





**ARTICLE**

# Toward quantitative metabarcoding

Andrew Olaf Shelton<sup>1</sup>  | Zachary J. Gold<sup>1,2</sup>  | Alexander J. Jensen<sup>2,3</sup>  |  
 Erin D'Agnese<sup>3</sup> | Elizabeth Andruszkiewicz Allan<sup>3</sup> | Amy Van Cise<sup>4</sup> |  
 Ramón Gallego<sup>3,5</sup> | Ana Ramón-Laca<sup>2,3</sup> | Maya Garber-Yonts<sup>3</sup> |  
 Kim Parsons<sup>1</sup> | Ryan P. Kelly<sup>3</sup> 

<sup>1</sup>Conservation Biology Division,  
 Northwest Fisheries Science Center,  
 National Marine Fisheries Service,  
 National Oceanic and Atmospheric  
 Administration, Seattle, Washington, USA

<sup>2</sup>CICOES, University of Washington and  
 Northwest Fisheries Science Center,  
 National Marine Fisheries Service, Seattle,  
 Washington, USA

<sup>3</sup>School of Marine and Environmental  
 Affairs, University of Washington, Seattle,  
 Washington, USA

<sup>4</sup>North Gulf Oceanic Society, Visiting  
 Scientist at Northwest Fisheries Science  
 Center, National Oceanic and  
 Atmospheric Administration, Seattle,  
 Washington, USA

<sup>5</sup>Departamento de Biología, Universidad  
 Autónoma de Madrid, Unidad de  
 Genética, Madrid, Spain

**Correspondence**

Andrew Olaf Shelton  
 Email: [ole.shelton@noaa.gov](mailto:ole.shelton@noaa.gov)

**Handling Editor:** Hideyuki Doi

**Abstract**

Amplicon-sequence data from environmental DNA (eDNA) and microbiome studies provide important information for ecology, conservation, management, and health. At present, amplicon-sequencing studies—known also as metabarcoding studies, in which the primary data consist of targeted, amplified fragments of DNA sequenced from many taxa in a mixture—struggle to link genetic observations to the underlying biology in a quantitative way, but many applications require quantitative information about the taxa or systems under scrutiny. As metabarcoding studies proliferate in ecology, it becomes more important to develop ways to make them quantitative to ensure that their conclusions are adequately supported. Here we link previously disparate sets of techniques for making such data quantitative, showing that the underlying polymerase chain reaction mechanism explains the observed patterns of amplicon data in a general way. By modeling the process through which amplicon-sequence data arise, rather than transforming the data post hoc, we show how to estimate the starting DNA proportions from a mixture of many taxa. We illustrate how to calibrate the model using mock communities and apply the approach to simulated data and a series of empirical examples. Our approach opens the door to improve the use of metabarcoding data in a wide range of applications in ecology, public health, and related fields.

**KEYWORDS**

amplicon sequencing, bias adjustment, community structure, compositional analysis, diet analysis, environmental DNA

**INTRODUCTION**

During the past decade, rapid technological advances in the collection and analysis of trace genetic material from sampled environmental media water ([Ficetola et al., 2008; Thomsen et al., 2012], soil [Andersen et al., 2012], feces [Pompanon et al., 2012], or even air [Lynggaard et al., 2022]; hereafter environmental

DNA [eDNA]) have opened new frontiers for environmental surveillance. Studies using eDNA have focused on diverse topics including monitoring biodiversity (Creer et al., 2016), managing invasive species (Jerde et al., 2013), characterizing diet (Deagle et al., 2013), and supporting fisheries management (Fukaya et al., 2021; Shelton et al., 2022), from tropical forests (Lopes et al., 2017) to the deep sea (Everett & Park, 2018).

These ecological studies and many others use techniques essentially identical to those in microbial ecology, microbiome, and public-health applications, and all share a set of analytical challenges. In most amplicon-based studies (hereafter: “metabarcoding”), a single oligonucleotide primer set targets a region of DNA shared among a taxonomic group of interest (see Taberlet et al., 2012) to be amplified via polymerase chain reaction (PCR) and subsequently sequenced, with the primer design determining which taxa are likely to be amplified and thus detected. The result is a mixture of DNA sequences from many taxa; the challenge is to determine whether and how the abundance of those sequence reads corresponds to the starting composition of DNA prior to amplification.

There is agreement that metabarcoding data contain information about the taxa present in a sample, and therefore inform estimates of taxonomic richness (reviewed in Taberlet et al., 2018). However, using metabarcoding data to estimate the composition (i.e., taxon-specific proportions) of DNA within a sample is more controversial. Metabarcoding data consist of counts of unique DNA sequences detected by a DNA sequencing platform (e.g., for a given sample we might observe three copies of sequence A, 1001 copies of sequence B, etc.). Important bioinformatic decisions bear on how multiple sequences are combined to represent species or genera or higher taxonomic groups (Macé et al., 2022), but the resulting data themselves are counts of reads associated with particular taxa for each sample. Many uses of these data in an ecological setting share an (often implicit) assumption that the reads emerging from DNA sequences are an accurate depiction of the sample composition prior to amplification (e.g., Laporte et al., 2021). Connecting DNA composition in a sample to the community of taxa present the environment involves the additional understanding of and assumptions about the “ecology of environmental DNA” (see e.g., Barnes & Turner, 2016); we do not discuss these additional aspects here.

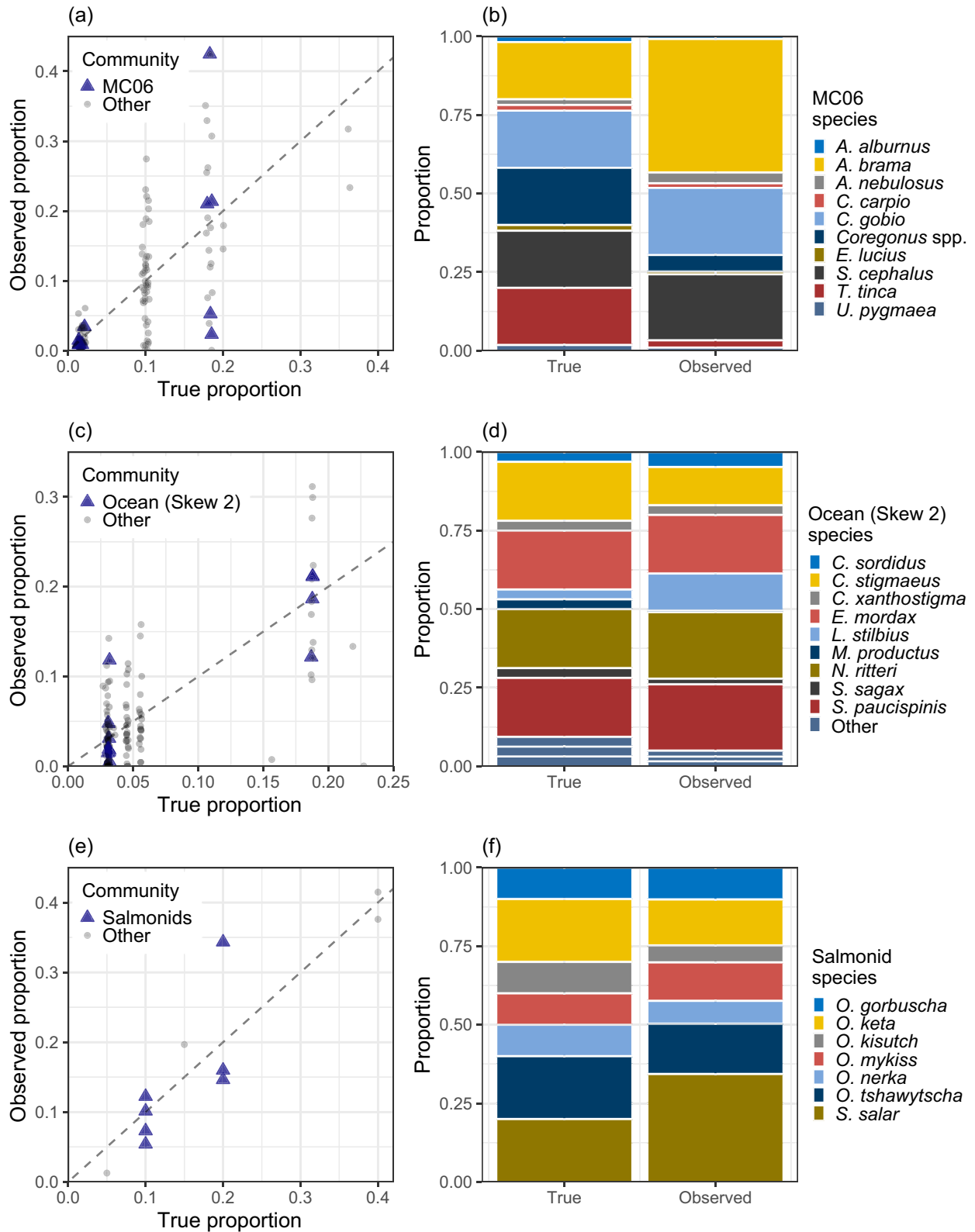
There is abundant evidence that the relationship between the true composition of DNA contained in a sample and the reads emerging from the sequencer is far from simple. This fact has been documented in the microbiome and microbial literature (Gloor et al., 2017; McLaren et al., 2019; Silverman et al., 2021) but less so in the ecological literature (but see e.g., Thomas et al., 2016). The most compelling evidence for this phenomenon comes from the analysis of mock communities in which researchers create a known mixture of DNA from a suite of taxa of interest and compare the relative abundance of the reads from each taxa against the known community. Read counts following amplification and sequencing invariably fail to match—or often, even approximate—the mock-community DNA starting proportions. For example, relative read counts deviate strongly from the mock communities created

for freshwater mussels (Coghlan et al., 2021), freshwater invertebrates (Fernández et al., 2018), arthropods (Kreihenwinkel et al., 2017; Piñol et al., 2015), freshwater fish (Hänfling et al., 2016; Rivera et al., 2021), marine vertebrates (Andruszkiewicz et al., 2017; Port et al., 2016), fungi (Adams et al., 2013; De Filippis et al., 2017; Palmer et al., 2018), diet studies from a range of organisms (Ando et al., 2020; Ford et al., 2016; Thomas et al., 2016; Tournayre et al., 2020), and microbiome studies (McLaren et al., 2019; Silverman et al., 2021). Beyond their obvious taxonomic diversity, these analyses span a wide range of methodological implementations (i.e., primers and protocols), and are reproducible. Because the differences between expected and observed read proportions often appear idiosyncratic and species specific, it is difficult to know how to interpret sequencing reads for quantitative use.

## The problem

We illustrate the problem using empirical data from three fish-community datasets: one from British lake communities (Hänfling et al., 2016; Figure 1a,b), one from Pacific marine communities (this paper; Figure 1c,d), and one from fecal diet samples from fish-eating killer whales (*Orcinus orca*, this paper; Figure 1e,f). For each study, we first plot the true proportion of DNA from a mock community of known composition against the estimated proportion of reads detected from each taxon (Figure 1a,c,e).

If the estimated proportions based on sequencing data accurately reflected the original known proportions of the mock community, all points would lie on or very near the 1:1 reference line. While it is reassuring that there is some suggestion of a relationship—larger true proportions are associated with larger observed proportions—points are scattered well above and below the reference line with some points more than double or less than half their true proportions. It is tempting to examine (Figure 1a,c,e; note that observations are scattered approximately equally above and below the reference line), and conclude that the true proportions are well estimated on average (e.g., Lamb et al., 2019), but to do so would be a mistake. Because these values are proportions of reads associated with each taxon, metabarcoding reads are compositional: the sum of proportions across taxa must equal one, and therefore the points on this graph are not independent. Indeed, if one taxon is above the reference line, one or more other taxa must, by definition, fall below the reference line. Consequently, a positive relationship is very likely for metabarcoding datasets, and, with enough randomly assembled mock communities, the estimated slope will be very near 1. For metabarcoding data nonindependence among the observations renders most standard statistical



**FIGURE 1** Comparisons between the composition of a known mock community (“true”) and the estimated proportions derived from raw read counts following sequencing (“observed”). Left panels show the relationship for species within mock communities constructed by Hänfling et al. (2016) ((a); 10 mock communities of freshwater fishes; cytochrome b mtDNA), and for this study ((c); two mock communities of Pacific Ocean fishes; 12S rRNA; and (e) two mock communities of southern resident killer whale fecal samples; 16 S rRNA). Each point (dot or triangle) represents a single species in a mock community. For panel (a), each point is a single technical replicate; whereas in panels (c) and (e), each point is the average across three technical replicates. Communities highlighted (triangle symbols) are shown in the stacked bar charts in the right panels (b, d, f).

analyses (e.g., ordinary regression or correlation analyses) inappropriate (see Erb et al., 2020; Gloor et al., 2017).

Examining the true and estimated compositions for a single community (triangles in Figure 1a,c,e, are shown in stacked bar charts in Figure 1b,d,f), it becomes obvious that the true and estimated communities differ substantially. A few taxa match their true abundance closely (see e.g., *Oncorhynchus gorbuscha* in Figure 1f, *Engraulis mordax* in Figure 1d) but some taxa are strongly over-represented (*Abramis brama* in Figure 1b; *Leuroglossus stilbuis* in Figure 1d) while others are under-represented (*Sardinops sagax* in Figure 1d; *Oncorhynchus kisutch* in Figure 1f).

Two important questions stem from these observations: First, why do these biases in taxonomic composition arise? Second, how do we correct for such biases? We address each in the following sections.

## The process

While there are several aspects of sequencing data that differ from other types of surveys, PCR amplification and the biases it can introduce into metabarcoding data have repeatedly been identified as the major factor limiting its usefulness (Beng & Corlett, 2020; Deagle et al., 2010, 2013; Gloor et al., 2017; Kelly et al., 2019; Krehenwinkel et al., 2017; McLaren et al., 2019; Shelton et al., 2016; Silverman et al., 2021). Thus correcting for PCR-driven biases remains the major challenge for making metabarcoding results reflective of DNA concentrations, although other processes besides amplification can hinder the reconstruction of the relationship between metabarcoding data and taxonomic abundance (e.g., variability in DNA deposition and persistence, copy number variation, biases in taxonomic assignment, PCR inhibition; Beng & Corlett, 2020; Gohl et al., 2016; McLaren et al., 2019). Understanding the mechanisms by which metabarcoding data are produced is therefore essential for reliably understanding PCR-based data.

We use a structure for understanding PCR-driven biases that has been used previously (Kelly et al., 2019; McLaren et al., 2019; Silverman et al., 2021) and allows PCR-driven variation to be a function of variation among taxa in PCR amplification efficiency. Specifically, for any taxon  $i$ , the number of sequence reads produced during a PCR reaction are governed by an efficiency parameter  $a_i$ , which is characteristic of the interaction between the particular primer set and each taxon (or unique sequence variant) being amplified. Thus, for any given taxon, we expect the number of reads to be directly related to the efficiency of amplification and the starting concentration of DNA template. Let  $A_i$  be the expected number of

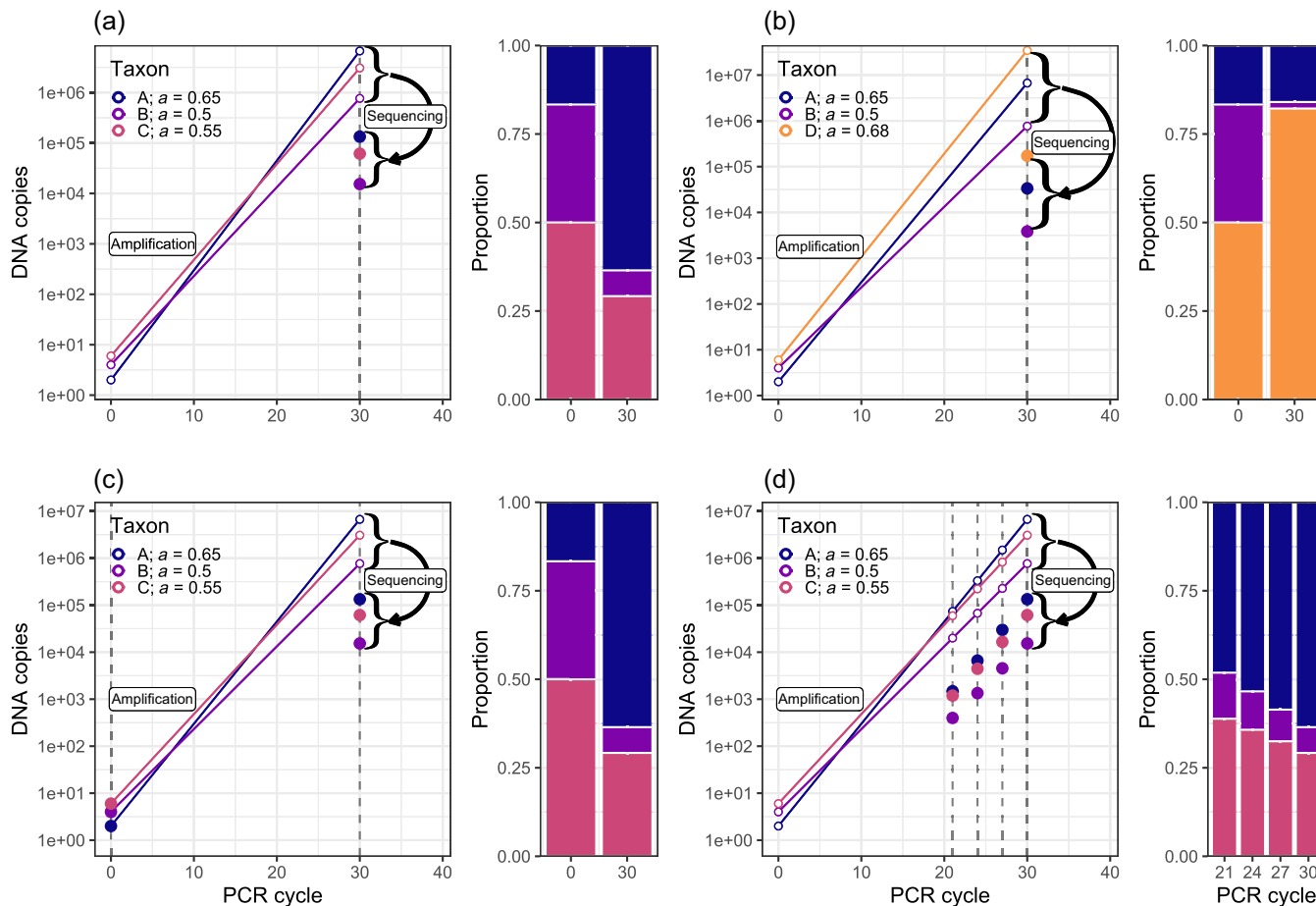
sequence reads after PCR,  $c_i$  be the true number of DNA copies in the reaction attributable to taxon  $i$ ,  $a_i$  be the amplification efficiency (bounded on  $(0, 1)$ ), and  $N_{\text{PCR}}$  be the number of PCR cycles used in the reaction:

$$A_i = c_i(1 + a_i)^{N_{\text{PCR}}}. \quad (1)$$

Note that this equation only applies during the exponential phase of PCR, before reagents have been exhausted and the amplification process has stopped, a valid assumption for most eDNA applications which start with very low DNA concentrations. If we could perfectly observe the DNA molecules, the above equation alone would be sufficient to understand the value of interest,  $c_i$ . Unfortunately PCR and sequencing technology does not allow for such direct observations. For any useful primer set,  $a_i$  is typically not close to 0 (a value of 0 would indicate no amplification during PCR) and  $N_{\text{PCR}}$  is large ( $> 30$ ), so the number of sequence reads expected for any taxon with  $c_i > 0$  is very large (e.g., with  $c_i = 2$ ,  $a_i = 0.75$ , and  $N_{\text{PCR}} = 36$ ,  $A_i = 1.12 \times 10^9$ ). Given the simultaneous amplification of many taxa,  $10^{10}$  or more DNA copies are typically produced.

DNA sequencing instruments report only a small fraction of the total amplicons following amplification (often on the order of  $10^6$  to  $10^7$  reads per sample, depending upon the sequencing instrument, indexing decisions, and so on). Thus only a small fraction of the total generated amplicons are observed. Assuming that the sequencing mechanism reports an unbiased subsample of amplicons, we can think of the observed reads for each taxon ( $Y_i$ ) as proportional to the true amplicon abundance:  $Y_i \propto A_i$ . This sampling changes what in Equation (1) appears to be a single-taxon process—each taxon being amplified independently—into a multitaxon, compositional process; the number of amplicons observed for taxon  $i$  will depend both upon the amplicons produced for taxon  $i=1$  and the amplicons from taxa  $i=2, 3, \dots, I$  in the same reaction.

The consequences of (1) among-taxon variation in amplification rate, and (2) the compositional nature of the resulting data for inferences about ecological communities are profound. We provide two graphical examples of the potential effects of amplification variability on inferences for communities of three hypothetical taxa (Figure 2a,b). Ecologists are interested in quantifying the DNA present before any PCR amplification (PCR cycle 0 in Figure 2), but if taxa differ in amplification efficiencies ( $a_i$ ), the relative abundance of amplicons can shift dramatically over the course of PCR amplification such that when the reads are observed following sequencing (filled points), the composition of the observed sequences differ substantially from the starting DNA molecules.



**FIGURE 2** Two examples of simulated changes in communities due to polymerase chain reaction (PCR) bias (a, b) and graphical illustrations of potential solutions (c, d). In all panels, filled circles indicate known values or observations and empty circles indicate unknown states. (a, b) Communities of three taxa with variation in amplification efficiency. Left panels illustrate the PCR process in which slopes of lines are determined by the taxon-specific amplification efficiencies ( $a_i$ ). Right panels compare the true starting-community DNA composition at PCR cycle 0 and the observed community composition at PCR cycle 30. Taxa A and B are shared across communities, with identical starting proportions in each, but radically different proportions following PCR. (c, d) Two schematic illustrations of techniques to enable estimation of amplification bias among taxa. The mock-community approach (c) provides known initial taxon composition (PCR cycle 0) that can be compared with the composition of sequences observed at PCR cycle 30 to estimate relative amplification rates. The variable PCR cycle calibration (d) uses a sample with unknown community composition at PCR cycle 0, but observes the community of amplicons multiple points in the PCR process, sampling at different numbers of amplification cycles. (d) Change in relative community composition across PCR cycles provides information about relative amplification rates. The right panels of both (c) and (d) show the observed composition of the samples after a given number of PCR cycles.

Note that the relative abundances and even the rank order of abundances can change between the initial and final PCR cycle (0 and 30, respectively; Figure 2). Furthermore, it is vital to understand that this is a multivariate process; the observed relative abundance of any one taxon is dependent upon the other taxa co-occurring with it in the sample. Compare the fate of taxon “A” in the two different communities depicted in Figure 2a,b. In both panels, taxon “A” initially comprises one-sixth ( $\approx 17\%$ ) of the community, but in the first community it makes up more than 60% of amplicons after PCR amplification (Figure 2a); however, in the second community it declines slightly to  $\approx 16\%$  of amplicons after PCR

amplification (Figure 2b). This is not a result of anything changing about taxon “A” but rather the fact that a taxon with a higher amplification efficiency (taxon “D”) comprises half the community in Figure 2b, while it is replaced by a taxon with lower efficiency (taxon “C”) in the community in Figure 2a.

While Figure 2a,b present a stylized example, it makes four important points. First, varying amplification efficiencies among taxa have the potential to dramatically affect the amplicon counts observed after sequencing. Second, the patterns of bias produced by allowing for amplification variation qualitatively match the patterns observed in the empirical analysis of mock communities



(Figure 1). Third, in the presence of varying amplification efficiencies, it is impossible to determine the initial composition of a sample simply by observing the relative abundance of amplicons associated with each taxon after sequencing; many possible combinations of parameter values for starting proportion and amplification efficiency would yield the same observed proportions post-PCR. Finally, interpreting amplicon counts cannot be done on a taxon-by-taxon basis, but must be done in a multivariate context. We can therefore quickly identify and disregard approaches that will clearly not resolve the problem. Specifically, simple data transformations—including logarithms, roots, or any other monotonic transformation—will not succeed in correcting for biases introduced by amplification variation among taxa (e.g., Kelly et al., 2019).

## METHODS

Various techniques have been proposed to reconcile the differences between the amplicon data we observe and the abundances of the underlying organism-specific DNA concentrations in the environment. These range from process-based statistical approaches solidly grounded in theory (McLaren et al., 2019; Silverman et al., 2021), to laboratory methods such as tagging molecules individually prior to amplification to distinguish replicate amplicons from unique template molecules after sequencing (e.g., qSeq, Hoshino & Inagaki, 2017; Hoshino et al., 2021, and other molecular ID tags [MIDs]) to various post hoc transformations and corrections (Kelly et al., 2019; Krehenwinkel et al., 2017; Thomas et al., 2016).

Here we treat observed sequence reads as arising from the mechanics of the PCR reaction and subsequent DNA sequencing, developing a statistical model grounded in the PCR process itself. We set out a quantitative model for amplicon data to account for the effects of amplification bias, estimating the proportion of each taxon's DNA in the original PCR template (i.e., community composition of our samples) prior to PCR and sequencing. The key to doing so is estimating taxon-specific amplification-efficiency parameters. We use multinomial logistic regression to account for PCR bias, and make clear the general applicability of these models to all kinds of amplicon-based studies. We emphasize that the models we use are not unique—other researchers have provided closely related approaches in the medical and microbiome literature (e.g., McLaren et al., 2019; Silverman et al., 2021)—but these techniques are underused and underappreciated in the ecological literature. We also note that there are some similarities between existing correction procedures (Krehenwinkel et al., 2017; Thomas et al., 2016) and parts of our approach.

We summarize the primary statistical model in the main text and present extensions in Appendix S1. After the model description, we highlight the calibration requirements for these models and emphasize how adding a few steps to molecular data-collection protocols can greatly improve their value for ecological inference. Given the potential for uncorrected data to lead to potentially large errors in ecological inference, we aim to make a complex statistical model approachable to most practitioners.

## Compositional models for amplicon data

The sampling process associated with DNA sequencing severs the link between the absolute abundance of the initial DNA copy count or concentration ( $c_i$ ; see Equations 1 and 2) and the count of DNA sequences observed (the  $Y_i$ ). Gloor et al. (2017) present illuminating examples illustrating the challenges of compositional data. Specifically, we can write the ratio of observed sequences for taxa  $i$  and  $j$ ,  $\left(\frac{Y_i}{Y_j}\right)$  as a function of the initial ratio of taxa  $i$  and  $j$ ,  $\frac{c_i}{c_j}$ , modified by the product of the ratio of amplification efficiencies for the two taxa,  $\frac{1+a_i}{1+a_j}$ , and the number of PCR cycles:

$$\log\left(\frac{Y_i}{Y_j}\right) \propto \log\left(\frac{c_i}{c_j}\right) + N_{\text{PCR}} \log\left(\frac{1+a_i}{1+a_j}\right). \quad (2)$$

Note that in Equation (2) there are many possible values that yield the same ratios (e.g., the pair  $\{c_i = 2, c_j = 1\}$  produce the same ratio as the pair  $\{c_i = 4, c_j = 2\}$ ;  $\frac{c_i}{c_j} = 2$ ), emphasizing the loss of information about absolute scale when dealing with compositional data and ratios.

To solve this scaling problem, we can arbitrarily define one taxon to be a reference taxon ( $R$ ) and define a new set of parameters relative to this reference taxon. Let  $\beta_R = 0$  be the log-abundance of the reference taxon in the initial sample and  $\beta_i$  be the abundance of taxon  $i$  relative to the reference ( $\beta_i > 0$  indicates taxon  $i$  is more abundant than the reference,  $\beta_i < 0$  the opposite). Similarly, let  $\alpha_i$  be the log-efficiency relative to the reference taxon ( $\alpha_R = 0$ ), then  $\nu_i$  is the log-abundance of taxon  $i$  relative to the reference taxon after sequencing:

$$\nu_i = \beta_i + N_{\text{PCR}} \alpha_i. \quad (3)$$

By definition,  $\nu_R = 0$ , and the choice of reference taxon is arbitrary, but not unimportant. In this formulation,  $\nu_i$  is a linear function having slope  $\alpha_i$  and intercept  $\beta_i$ ; the intercept determines the proportion of DNA for taxon  $i$  present in the sample before PCR, and the equation defines the proportional abundance for any number of focal taxa.

We acknowledge the stochastic processes that contribute to the observed read counts beyond the deterministic skeleton presented in Equation (3), and we can model this stochasticity using a multinomial likelihood (Egozcue et al., 2020; Silverman et al., 2021). A full model for observed counts for all  $I$  taxa is:

$$\begin{aligned}
 \mathbf{Y} &\sim \text{Multinomial}(\boldsymbol{\mu}, N), \\
 \mu_i &= \frac{e^{\nu_i}}{\sum_{i=1}^I e^{\nu_i}}, \\
 \nu_i &= \beta_i + N_{\text{PCR}}\alpha_i + \epsilon_i, \\
 \epsilon_i &\sim N(0, \tau_i), \tag{4}
 \end{aligned}$$

where bold text indicates vectors and  $N$  is the observed total number of sequences in the sample. The second line of Equation (4) is the softmax transformation which produces proportions for each taxon ( $\mu_i$ ) from the ratios of abundance. The parameter  $\epsilon$  allows for overdispersion in the counts beyond the variability provided by the multinomial distribution, capturing the substantial variance among technical replicates often observed in metabarcoding data. As will be seen below,  $\epsilon$  can be important because read counts may vary substantially across replicate PCR reactions, but this may not be estimable for some metabarcoding datasets, which commonly lack technical replication. The model described in Equation (4) is known as a multinomial logistic regression model, and defining log-ratios relative to a reference taxon is known as the additive log-ratio transform (ALR; Aitchison, 1986).

Above we have written a simple scalar-valued form to improve readability, but the regression component quickly generalizes and can take on more complicated structures from the diverse world of generalized linear models (e.g., Silverman et al., 2021). Ecologists are rarely interested in a model for a single sample as written in Equation (4) but in a model for many samples taken across space (e.g., latitude, habitats) and/or time (seasons, months, years) and the model generalizes to accommodate such questions. For example, in the context of diet data it may be valuable to add terms describing measured covariates (e.g., temperature) or factors such as season, and the stochastic term can be modified to allow for additional covariance structure among samples. We present a general model in Appendix S1 to make such modifications explicit. Similarly, it may be desirable to use transformations other than the ALR depending on the application (e.g., centered log-ratio transform [CLR] or isometric log-ratio transform; Pawlowsky-Glahn & Egozcue, 2016; Silverman et al., 2021).

Most metabarcoding datasets include a single sequencing replicate for each sample collected after PCR amplification (corresponding to a single  $N_{\text{PCR}}$ ; note that PCR amplification can include either one or two-step PCR protocols; see “Methods”; Appendices S2 and S3). Thus many datasets have only  $I$  observations (one sequence count for each taxon). The model above requires at minimum  $I - 1$   $\beta$  parameters and  $I - 1$   $\alpha$  parameters. Thus with standard metabarcoding data, there are more parameters than data points and this model cannot be estimated. In the next sections we discuss how to integrate other data sources to calibrate the model, make the parameters identifiable, and allow researchers to generally correct for amplification bias.

### Calibration methods

There are at least four general strategies that can provide the information necessary to estimate amplification biases: (1) create mock communities of known DNA template composition and conduct PCR amplification and sequence this known community; (2) use samples of unknown composition, but vary the number of PCR cycles among technical replicates and subsequently sequence all replicates; (3) model amplicon results alongside another independent set of observations of the same community; (4) attach unique molecular identifiers to source molecules prior to amplification and analyze unique identifiers post-PCR. We focus on the first approach here because we believe it to be broadly applicable and we have had success implementing it in practice. Silverman et al. (2021) provide an investigation using variable PCR cycles applied to microbiome data. Gold et al. (2022) provides a substantial treatment of the third approach, there with amplicons and visual counts of larval fish communities in ethanol-preserved jars. Hoshino and colleagues review the fourth approach (qSeq; Hoshino & Inagaki, 2017; Hoshino et al., 2021).

To provide intuition about how these methods allow us to estimate amplification rates, we present two simple graphical illustrations for approaches using mock communities and variable PCR (Figure 2c,d). For both approaches, the amplification efficiency for each taxon is directly related to the slope of the line in Figure 2c,d (see also Equation 3).

### Mock communities

To calibrate using a mock community, we create the starting community of DNA, generally from vouchered tissue extractions. We therefore know the proportions of each taxon before PCR (cycle 0) and we observe the

amplicons following sequencing (Figure 2c). Thus we have an estimate of the initial and final community composition (filled points in Figure 2c), allowing us to estimate the relative amplification efficiency ( $\alpha$ , relative slope parameters) for each taxon. We can then apply these estimates of  $\alpha$  to samples of unknown community composition, yielding an estimate of the parameters of ecological interest,  $\beta$  (intercept parameters in Equation 4; Figure 2c). In practice it makes sense to use joint models that simultaneously incorporate observations from mock communities alongside samples of unknown composition. We use this technique in our simulations and empirical applications below.

## Variable PCR cycles

While mock communities offer a reasonable approach to estimating amplification biases, they have several drawbacks. Most obviously, they require constructing an appropriate mock community for a given application. When there are large numbers of taxa of interest or source DNA from important taxa are unavailable, creating a mock community of known composition may be difficult or impossible, as is often the case in microbial community studies, for example. In such cases, it is possible to modify the PCR protocol for technical replicates to bracket a range of PCR cycles and observe the amplicons at each end point (Figure 2d). Importantly this variable PCR cycle calibration can be applied to samples of unknown composition. In regression terms, the intercept parameters ( $\beta$ ) are unknown in the variable PCR cycle approach, but the change in relative composition across a range of PCR cycles enables an estimation of the relative amplification of different taxa (Figure 2d),  $\alpha$ , and in combination with observed proportions of sequences after PCR, yields estimates of starting proportions  $\beta$ . While this approach has been used to good effect with significant sampling effort (Silverman et al., 2021), we have not had success using it in practice.

## Applications

The quantitative model described above is only valuable if it can be useful in practice. We developed code to implement the model outlined in Equation (4) in *Stan* via the **R** language (*Rstan* v.2.21.2; *R* v.4.1.2) and then tested it against several datasets, both simulated and empirical. We provide simulations to illustrate the model's ability to recover known parameters and illustrate model limitations (Appendix S1). We then apply the mock-community calibration method to several

empirical examples to estimate community composition. Each empirical example is drawn from a different ecological context; these include British lake fishes (unreplicated PCR reactions from mock communities, cytochrome b *mtDNA*; data from Hänfling et al., 2016), Pacific Ocean fishes (replicated PCR reactions from mock communities, 12S rRNA; original data), and diet data derived from the gut contents of southern resident killer whales (SRKW; replicated PCR reactions, 16S rRNA; original data). We present details of model estimation and prior distributions for the model in Appendix S1 and a worked example using a small part of the Pacific fish dataset in Appendix S4. Because we focus here on eukaryotic communities, we include the above examples in the main text, but for completeness we also show an analysis of human bacterial microbiomes (replicated PCR reactions, 16S rRNA; data from Gohl et al., 2016) in Appendix S1; 16S bacterial data were also shown in Silverman et al. (2021) and McLaren et al. (2019). All data and code used in the applications are provided in the online supplement.

## British lakes

Hänfling et al. (2016) sampled eDNA from a series of freshwater lakes in northern England, and alongside these environmental samples, reported data from 10 cytochrome b mock communities having known starting compositions and consisting of partially overlapping mixes of 21 species drawn from the target lake-fish communities. These samples were amplified in triplicate but the three reactions were pooled and each mock community was sequenced once on an Illumina MiSeq system. See Hänfling et al. (2016) for analytical and bioinformatic details of the original dataset; the authors provided read counts, compositions for mock communities, and analytical code as part of the publication's supplementary material.

We took species-specific starting DNA concentrations and resulting read-numbers from the authors' Table S5 and, consistent with our method of subsetting data to the relevant target group, omitted any reads from species not included in the mock-community preparations (i.e., for which starting concentrations were zero; Appendix S3: Table S1). We note that this approach is not equivalent to simply ignoring contamination by selectively omitting data (see "Discussion"). Instead, it focuses the analysis on the subgroup of interest, and results in estimated proportions for species in that subgroup, rather than for the sequencing run as a whole. For purposes of model-fitting, we treated all PCR reactions as having 30 amplification cycles (appendix 6 in Hänfling et al., 2016).



For model-fitting and cross-validation we created a version of the metabarcoding model with no overdispersion term (i.e., fix  $\epsilon = 0$  in Equation 4) because, in the absence of technical replicates, it is impossible to estimate overdispersion and so the read counts are assumed to follow a multinomial distribution. We fitted two models, the model described above and a model that assumes equivalent amplification among all species (i.e., fix  $\alpha = 0$  and  $\epsilon = 0$  in Equation 4). For each model we then used the odd-numbered mock communities (1, 3, 5, 7, 9) as samples with known starting concentrations, and even communities (2, 4, 6, 8, 10) as unknowns to be estimated. Because we knew the starting composition of each of those communities, we used this second set for external cross-validation. We then did the reciprocal cross-validation, using even-numbered communities as known and odd-numbered communities as unknown, giving us a complete set of 10 cross-validated, out-of-sample estimates. We compare posterior estimates of the community compositions relative to the known starting-community composition of the mock community using Aitchison distance (Aitchison, 1982).

## Pacific Ocean fish

To analyze a larger suite of species and to estimate overdispersion among technical replicates, we extracted DNA from voucher tissue samples from the Scripps Institute of Oceanography Marine Vertebrate Collection, generating template communities of temperate fish communities from the Northeast Pacific (see detailed methodologies in Appendix S3). We generated two species pools: “North,” comprised of 22 fish species found in coastal Alaskan waters, and “Ocean,” containing 18 species representative of fish found in the California current system (Appendix S3: Table S2). Six species were present in both pools. For each species pool we then constructed three mock communities with varying DNA composition. In one community, all species’ were equally abundant (“even” community). In the other two, we selected 12 species, four species each comprised 18.75% of the community and the remaining eight species comprised 3.125% (“skew 1” and “skew 2” communities; Appendix S3: Table S2). We amplified each mock community in separate triplicate reactions using the MiFish Universal Teleost 12S primer set (Miya et al., 2015) using a two-step PCR protocol. Appendix S3 provides further information about assembly of the mock communities and bioinformatic processing.

As with the British lakes data, we divided our data into two groups, a set used for estimating amplification parameters and a set held out for out-of-sample

cross-validation. We used two communities (Ocean even, North even) as mock communities with known composition and the remaining four (Ocean skew 1, Ocean skew 2, North skew 1, North skew 2) to calculate out-of-sample predictive accuracy. Thus, we used a one-third of our mock communities to predict the remaining two-thirds. Each community had three technical replicates and therefore we used the full model as described in Equation (4). We also fitted a second model assuming equivalent among-species variation in amplification (fix  $\alpha = 0$  in Equation 4) and used Aitchison distance to compare the predicted versus true community composition for each model.

## Killer whale diet

To understand the impact of amplification bias on diet estimation, we examined data generated from a small subset of fecal samples collected from SRKW. SRKW are fish-eating whales that live primarily along the coast of Washington State, USA and British Columbia, Canada, and are listed as endangered under the Endangered Species Act. Understanding SRKW diet is important for understanding prey availability and supporting the recovery of their population (Hanson et al., 2021). We combine mock communities for common SRKW prey items with a small number of field collected fecal samples to understand how amplification bias may change estimates of diet composition.

We generated two mock communities representative of SRKW prey species from genomic DNA extracted from individual, vouchered fish-fin or muscle samples (Ford et al., 2016). Whole genomic DNA was normalized to a concentration of 0.5 ng/ $\mu$ l using a qPCR SYBR assay of a fragment of the 16S SSU rRNA gene, before being combined in known proportions into two mock communities comprising 10 species (seven salmonid species and three other species; Appendix S3: Table S4).

We included eight fecal samples of unknown species composition (“field samples”) to analyze alongside the mock communities. Samples were collected during the month of September in 2017 ( $n = 6$ ), 2018 ( $n = 1$ ), and 2021 ( $n = 1$ ). We amplified DNA extracts with primers targeting the 16S rDNA region (Ford et al., 2016) for both mock communities and field samples, using a two-step, 32-cycle PCR with two technical replicates for each field sample and four technical replicates for each mock community. Full laboratory and bioinformatic protocols are described in Appendix S3, as is information for incidental take permits under the United States Endangered Species Act and the United States Marine Mammal Protection Act.

Unlike in the British lakes and Pacific fish examples, we do not have additional, known SRKW diet communities with which to make out-of-sample estimates of model accuracy. Instead, we estimate two models [one with amplification variability (Equation 4) and one without amplification variability (Equation 4 but with  $\alpha = 0$ )] to illustrate how including amplification biases modifies estimated diet composition field samples. For analysis, we include only the 10 prey species included in the mock community and excluded other rare species arising in the sequenced samples; in total other species never comprised more than 0.65% of reads in any sample. We present predicted compositions for individual samples as well as used the posterior estimates for each individual fecal sample to derive an average diet composition from these eight samples.

## RESULTS

### Simulations

Our simulations suggest that calibration with mock communities effectively corrects for amplification bias (Appendix S1: Figure S1). With limited variability among technical replicates, calibration recovers taxon starting proportions accurately and precisely when amplification efficiencies have low-to-moderate variability among taxa (Appendix S1: Figure S1). The mean posterior estimates are unbiased and strongly correlated with the true proportions, and credible intervals are small; model estimates uniformly better approximate true proportions than do observed proportions.

### British lakes

We estimated substantial variation in the relative amplification efficiencies ( $\alpha_i$ ) of lake-fish species, which varied over roughly 0.3 units (Figure 3d; we used *A. brama* [common bream] as the reference species;  $\alpha_R = 0$ ). Estimated amplification efficiencies derived from using the odd-numbered versus even-numbered communities for calibration were quite similar, suggesting amplification efficiency derived from different mock communities are consistent. In general, lower efficiency species have greater uncertainty because they are observed rarely or are absent entirely in the observed sequences (Figure 3d).

Model estimates of DNA community composition more closely approximated the true starting concentrations (relative to a null model assuming a constant  $\alpha$  across species) in all 10 communities in cross-validated out-of-sample predictions (Figure 3). We illustrate

this improvement for two example communities (Figure 3a,b), as well as in the summarized Aitchison distance for all communities (Figure 3c; note that community MC07 had much larger Aitchison distance; see Appendix S1: Figures S3 and S4).

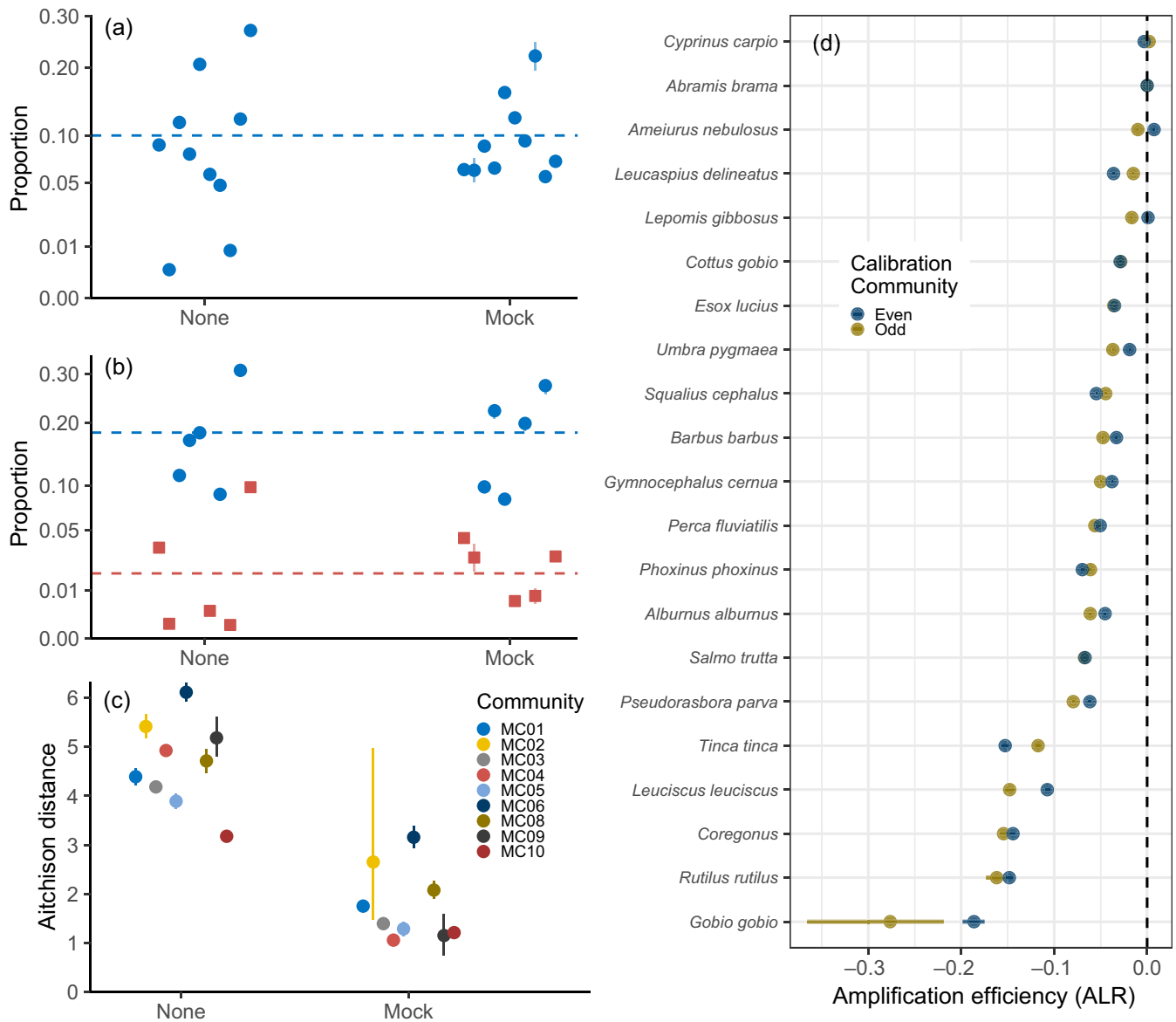
Despite the improved fit relative to the true composition, the credible intervals for all parameters were unreasonably small and only rarely did the credible interval for estimated proportions include the true composition. For example, no species has credible intervals that include the true proportion in Figure 3a,b, indicating overconfidence in estimates of composition as well as amplification efficiency (Figure 3d). This is largely a result of having a single data point for each species in each community, and the necessary assumption of multinomial sampling variability. As we show directly below, technical replication can resolve this issue.

### Pacific Ocean fish

For mock communities of 34 Pacific Ocean fish, we estimated substantial variation among species in the amplification efficiency with the 12S MiFish primer (Figure 4e). Note that rather than presenting direct estimates of  $\alpha_i$ , as in Figure 3, we transformed our amplification using the CLR (0 in Figure 4, which uses the geometric mean among species as 0 rather than the reference species as 0). This allows for consideration of taxon-specific amplification efficiencies relative to the average efficiency in the community.

In line with simulation and the British lake data, models that account for amplification variability produced estimates of species composition more similar to the true, underlying composition than models that did not account for amplification variability for all communities examined (Figure 4a–d). However, in contrast with the British lakes data, credible intervals for each species included the true species composition for 11 of 12 species in the Ocean skew 1 community (Figure 4a) and 10 of 12 species in the North skew 2 community (Figure 4b). Without calibration, the credible interval included the true proportion in 8 of 12 species in both Ocean skew 1 and North skew 2 (Figure 4a,b). Larger credible intervals better reflect the variability common in metabarcoding datasets, and here are a direct result of sequencing multiple technical replicates and allowing the model to estimate overdispersion ( $\tau_i$  in Equation 4). Estimates of  $\tau_i$  range from 0.25 to 3.9 (posterior mean) among species indicating increased variability relative to the multinomial distribution.

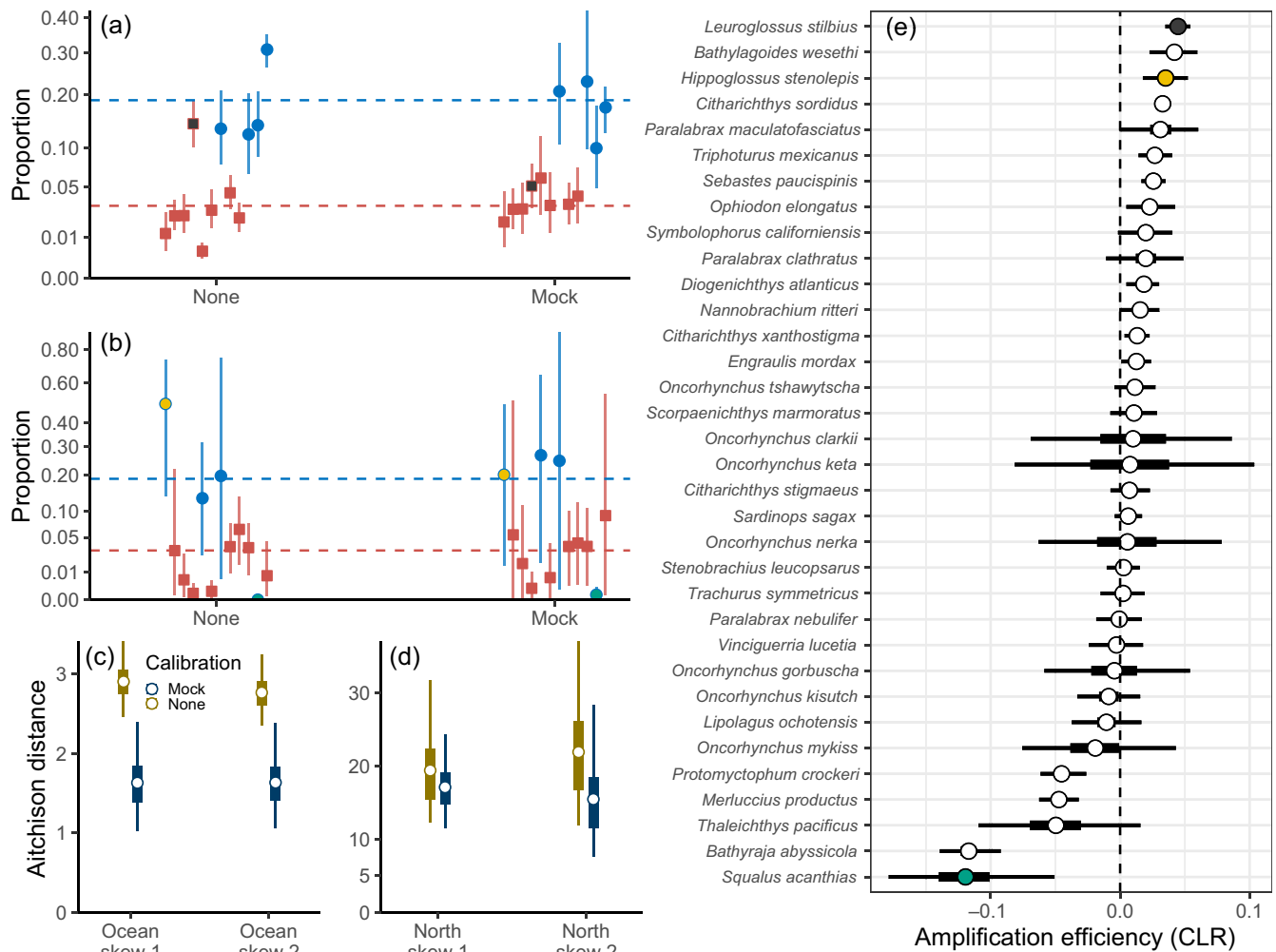
We highlight the behavior of a few key species with colors in Figure 4 to emphasize the effect of amplification



**FIGURE 3** Comparison of calibration methods for British lakes fish communities. (a) Posterior mean estimates (95% CI) of species composition for the 10 species in the mock-community MC03 without estimated amplification variability (“none”) or with amplification variability estimated using a mock community (“mock”). Dashed blue line shows the true composition for each species. (b) Posterior mean estimates (95% CI) of species composition for the 10 species in the mock community, MC08. Dashed blue line shows the true composition for the species identified with a circle. Dashed red line shows the true composition for the species identified with a square. Otherwise as in (a). (c) The similarity between the true composition and the estimated composition as measured by Aitchison distance. Posterior mean and 95% CI shown. Smaller values indicate greater similarity. (d) Posterior mean estimates (95% CI) of relative amplification efficiency derived from two calibration sets (odd-numbered communities used as the mock community or even-numbered communities used as the mock community). *Abramis brama* is the reference species ( $\alpha = 0$ ) for both communities. ALR, additive log-ratio.

efficiency on model predictions. For species with higher-than-average amplification efficiency (large  $\alpha_i$ ), the posterior estimates after calibration closely approximate the true species proportions even when the uncalibrated observations of those species are particularly misleading (see *L. stilbius* in Figure 4a,e and *Hippoglossus stenolepis* in Figure 4b,e). Conversely, species that amplify poorly offer little information on which to base

model fits, and so the posterior estimates will have little relationship to their true values (*Squalus acanthias*; Figure 4b,e). Despite comprising 20% of the true composition in North skew 2, *Squalus acanthias* had zero observed sequence reads in all three technical replicates. Because these data are compositional, inaccurate estimates of one species can substantially degrade the accuracy of estimates for all species in the community (Figure 4b).



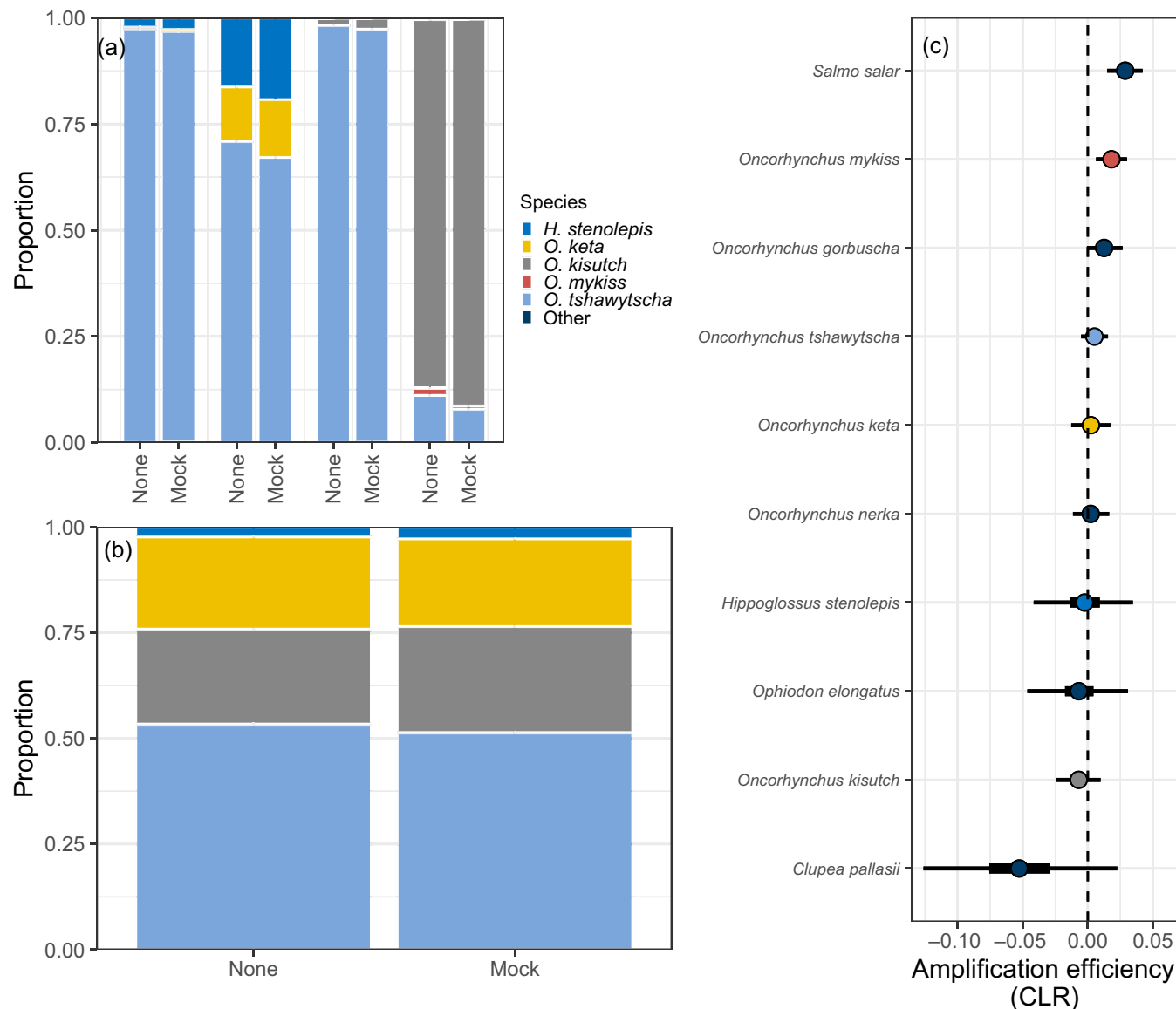
**FIGURE 4** Comparison of calibration methods for Pacific Ocean fish communities. (a) Posterior mean estimates (95% CI) of species composition for the 12 species in the mock-community Ocean skew 1 without estimated amplification variability (“none”) or with amplification variability estimated using a mock community (“mock”). Dashed blue line shows the true composition for the species identified with a circle. Dashed red line shows the true composition for the species identified with a square. Colors correspond to species noted in (e). (b) Posterior mean estimates (95% CI) of species composition for the 12 species in the mock community, North skew 2. Otherwise as in (a). (c, d) The similarity between the true composition and the estimated composition as measured by Aitchison distance for the four out-of-sample predicted communities. Posterior mean, interquartile range, and 95% CI shown. Smaller values indicate greater similarity. (e) Posterior mean estimates (95% CI) of relative amplification efficiency after centered log-ratio transformation (CLR). Dashed line indicates the geometric mean amplification efficiency among species. Yellow, black, and blue colors highlight species of interest in (a) and (b). *Citharichthys sordidus* was the reference species.

A consequence of getting a single species badly wrong can be seen in the large Aitchison distances indicating low similarity between the true and predicted community (compare y-axes of Figure 4c,d). We present an analysis removing *Squalus acanthias* from the community in Appendix S1.

## Killer whale diet

We found modest variation in amplification efficiency among the 10 species in the mock community; notably herring (*Clupea pallasii*) were underamplified and Atlantic

salmon (*Salmo salar*) were overamplified, but the remaining species were quite similar to one another (Figure 5). Furthermore, only five of the 10 focal species had more than 1% of the reads in any individual sample (Figure 5a) and herring and Atlantic salmon were not among those five species. As a result, the model had a small (but nonzero) effect on estimated diet composition from both individual samples (Figure 5a) and the among-sample average diet (Figure 5b). Species that were estimated to be above-average amplifiers decreased slightly in the across sample average (e.g., Chinook salmon decreased from 0.531 [0.484, 0.563], mean [95% CI], to 0.512 [0.468, 0.551] after adjusting for



**FIGURE 5** Comparison of calibration methods for southern resident killer whale diets. (a) Posterior mean estimates of diet for four individual samples without estimated amplification variability (“none”) or with amplification variability estimated using a mock community (“mock”). (b) Posterior mean estimates of diet composition across eight field samples collected during September between 2017 and 2021. (c) Posterior mean estimates (interquartile and 95% CI) of relative amplification efficiency after centered log-ratio transformation (CLR). Dashed line indicates the geometric mean amplification efficiency among species. Colors correspond to the colors in (a) and (b).

amplification variability) while below-average amplifiers increased slightly (e.g., coho salmon increased from 0.225 [0.190, 0.26] to 0.25 [0.198, 0.30]). SRKW diet estimates from metabarcoding underscore that when there is little amplification bias among focal taxa, the model output should look very similar to estimates from summarizing raw reads.

## DISCUSSION

By modeling the processes of PCR and sequencing, we provide a general method for correcting for the data

distortions generated by amplification bias. The challenges for amplicon data outlined above are general and apply to any kind of multitaxon PCR-based data. By linking this interpretable, mechanistic model to the existing statistical literature on compositional data analysis, we hope to popularize this technique, such that ecological inferences from metabarcoding may rest on a solid foundation. We find that our approach works well under a wide range of simulated scenarios, and for a diverse set of empirical examples spanning a range of ecological communities and subdisciplines.

Calibrating metabarcoding datasets is tractable and yields estimates of community composition under many



real-world conditions. Importantly, in all of the cases that we have examined, accounting for amplification bias improves the estimation of ecological communities relative to approaches that treat raw sequence counts as reflecting underlying communities. However, the approach is not without limitations or uncertainty; where an assay amplifies target species poorly, or where the variance among replicated samples is high—that is, cases in which the signal:noise ratio is very low—any inference about the DNA from ecological communities will be difficult and uncertain.

Thus metabarcoding datasets are very similar to “traditional” (nonmolecular) ecological information. Rather than a free-for-all of information having an unknown relationship to the living elements of sampled ecosystems, metabarcoding instead joins visual surveys, nets, traps, culture, and other observation methods in requiring contextual information to become interpretable. Just as cryptic species may elude visual surveys—and so, go unobserved—species that amplify poorly will be rarely observed in the metabarcoding data. Just as it is difficult to predict the abundance of patchily distributed species in a net or visual survey, it is difficult to predict the abundance of a species’ amplicons when the variation among technical replicates is high. Just as researchers should understand the sampling characteristics of their traditional ecological sampling tools in order to best understand the resulting data, so too should researchers understand the behavior of a given primer set in the context of molecular ecological data.

## Calibration, model performance, and extensions

It is now clear that simple tabulations of proportions of amplicon-sequence reads are likely to provide misleading inferences due to amplification bias. The model we present corrects for those biases to yield estimates of proportional contributions of each of the taxa (or amplicon sequence variants, etc.) to the original biological sample prior to PCR. Such calibration requires adding information beyond the raw observations of sequence reads and is not a trivial exercise. However, by dedicating a small portion of a sequencing run to calibration samples, researchers can derive robust estimates of their samples’ underlying DNA compositions. We note that there are several additional paths for calibrating metabarcoding data that are described elsewhere, and these either complement or may be combined with the approaches we discuss (Gold et al., 2022; Hoshino et al., 2021; Silverman et al., 2021).

Our model yields good estimates of community composition, particularly when amplification efficiencies do not vary excessively among taxa and among-replicate variability is low. While simulations suggest that the approach works well for an arbitrarily large number of taxa, in practice, the number of taxa will likely be constrained by the feasibility and patience required to construct mock communities. For valid inference, all taxa of interest must be included and observed in at least one mock community. However, the approach does allow researchers to subset data in arbitrary ways, focusing on only the taxa or samples of interest; model output will reflect this subsetting by estimating the composition of the selected ingroup. Such subsetting already occurs implicitly in most metabarcoding datasets and is evidenced by the species reported to be reliably detected by a given primer and molecular protocol. Therefore measuring and documenting amplification bias is a way of making explicit which components of the DNA from an ecological community can be measured by given primer and protocol and which cannot. A further point is that even species that are very rare in the environment (e.g., an endangered species) and therefore likely to have low eDNA concentration in the environment, should be able to be reliably detected as long as that taxon has a relatively high amplification rate.

Intuitively, the model fails in situations in which amplicons provide little information about the original community composition, either because of poor amplification or high variance. Less intuitively, the model can also fail when a few taxa amplify poorly and the rest amplify well; poor estimates for some taxa will affect the estimates for all other taxa because the data are compositional. Iteratively fitting the model can solve this problem, by dropping taxa with low information value and focusing on the remainder. Again, we view this iterative procedure positively because it forces researchers to be explicit about which taxa can be validly assessed by a given metabarcoding approach.

From a statistical perspective, we have presented a relatively simple model and applied it to relatively small datasets. However, the form of the model is easy to extend. The broad suite of statistical tools developed for regression can be easily incorporated; this includes making compositions a linear or nonlinear function of covariates, adding random effects, and incorporating spatiotemporal statistical models (see also Appendix S1). Furthermore, there are clear paths to generalize this framework to accommodate more than one genetic locus or dataset as well, offering a way of synthesizing information across genetic markers by treating data from each locus as an observation of a common ecological community composition. While such models can be

computationally difficult, there are few conceptual blocks to such advancement.

## Practical problems in metabarcoding studies

Our approach links the rapidly expanding field of metabarcoding in ecological applications with statistical methods developed in related fields and offers solutions to several practical problems of amplicon-sequencing techniques. We provide an extensive discussion of practical considerations in Appendix S2 and focus on a few points of general interest here including decontamination/denoising and the application of traditional ecological statistics to these data.

## Decontamination and denoising

Whether due to low-level laboratory contamination (Leonard et al., 2007), index-hopping and related technical problems (Carøe & Bohmann, 2020; Costello et al., 2018; Schnell et al., 2015), true detections of unintended taxonomic targets, or other mechanisms, metabarcoding datasets frequently contain observations of nontarget taxa. The question then arises whether, or how, these nontarget taxa might be responsibly identified and excluded from downstream analysis. For example, in diet-data analysis, might sequences from the host species be safely ignored? When detections of pigs, chickens, or others arise from PCR reagents themselves, how might we exclude these reads as contaminants?

The technique we present here can be applied to subsets of data by simply excluding nontarget taxa, thus changing the denominator for the overall read depth and shrinking the universe of taxa with DNA proportions to be estimated. Because the model is explicitly compositional, the analysis of any subset of the data remains valid; the resulting estimated proportions will sum to one, reflecting the proportions of the taxa in the subset analyzed, rather than in the entire raw dataset. If, for example, we wish to focus only on the five mammal species present in an environmental dataset, we may analyze only the reads of those five taxa in a set of (say) water samples. The resulting proportions will sum to one, reflecting the contributions of each of the five species to the analyzed subset, saying nothing about the proportions of those five species in the dataset as a whole. We emphasize that this subsetting procedure merely makes explicit what is inherent in any amplicon study: the amplified molecules observed reflect some, but not

all, of the molecules in the environment, namely those templates susceptible to the assay being used.

## Diversity indices and existing ecological statistics

Ecological studies frequently use Shannon, Simpson's, or other summary indices of diversity. However, sequence reads—whether in raw form or monotonically transformed—do not lend themselves to such indices, given that the indices rely upon proportional estimates of the underlying species present. For example, Shannon Entropy ( $\sum_{i=1}^N (p_i \log(p_i))$ ), for proportions  $p$  of species  $i = 1, \dots, N$ , when applied to raw metabarcoding reads, is meaningless and divorced from its connection to the underlying biology. In general, we care not about  $p$  as the probability of observing a sequence read from a given taxon, but rather  $p$  the probability of observing evidence of the underlying DNA collected, prior to distortion by PCR. Our model explicitly estimates the proportion of species' DNA in the sample, and so its output is appropriate for these common diversity indices and other standard downstream ecological analysis appropriate to proportion data. We note, however, the proportion of species DNA present does not necessarily reflect the proportion of species counts or biomass present in the environment due to generating, decay, and fate-and-transport phenomena (Barnes & Turner, 2016). Our work is agnostic about the connection between the abundance (or biomass) of taxa in natural communities and the DNA concentrations for taxa in the environment. However, our model does provide a necessary connection between observed amplicons after sequencing and collected DNA from field samples and our contribution is a necessary but not sufficient component for fully characterizing communities from eDNA.

## The need for replication

Technical replication supplies the data necessary to evaluate the signal:noise ratio in any study. When replicates are available, our model treats overdispersion—additional variance relative to that expected under a multinomial sampling model—as a random variable ( $\epsilon_i$ ) at the level of biological samples, drawn from a common distribution having a standard deviation  $\tau$ . Replication is expensive in metabarcoding studies due to sequencing costs and labor. Thus it is often desirable in practice to minimize the amount of technical replication in a study, in order to maximize effort elsewhere. In the absence of technical replicates, strong assumptions must be made about

the amount of variability in observed sequences. The example of British lakes (Figure 3) presents a potential consequence of no replication, overly precise estimates of community composition arise from assuming a multinomial likelihood. However, one might minimize replication after developing enough familiarity with a system to confidently assert a parameter value for overdispersion. Another approach might be to replicate some—but not all—samples in a study, yielding estimates of overdispersion that can be used throughout the dataset. Understanding the origin of overdispersion in sequence counts and best modeling approaches to account for such effects is a significant area deserving of further research in the metabarcoding community, perhaps most especially as to handling zero counts arising from a variety of mechanisms (Egozcue et al., 2020; Silverman et al., 2020).

## Conclusions

Across a broad swath of ecological, microbiological, and biomedical studies, it has become clear that simple read proportions or monotonic transformations calculated from metabarcoding studies have the potential to be deeply misleading. We outline approaches to correct for biases introduced by metabarcoding processes, but acknowledge that the laboratory and statistical effort to adjust for these biases are nontrivial and will hinder their rapid adoption. However, in such a rapidly advancing field, we trust that our work will lead to improved laboratory methods and statistical software to make implementation of these approaches routine.

## ACKNOWLEDGMENTS

We thank M. Fisher, J. Samhour, K. Vennemann, O. Wangenstein, and two anonymous reviewers for constructive comments on previous versions of the manuscript. We thank B. Hänfling and coauthors for providing excellent data and supplementary materials in their 2016 paper.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## DATA AVAILABILITY STATEMENT

Data and code (Shelton et al., 2022) are available in Zenodo at <https://doi.org/10.5281/zenodo.7158929>.

## ORCID

Andrew Olaf Shelton  <https://orcid.org/0000-0002-8045-6141>

Zachary J. Gold  <https://orcid.org/0000-0003-0490-7630>

Alexander J. Jensen  <https://orcid.org/0000-0002-2911-8884>

Ryan P. Kelly  <https://orcid.org/0000-0001-5037-2441>

## REFERENCES

- Adams, R. I., M. Miletto, J. W. Taylor, and T. D. Bruns. 2013. "Dispersal in Microbes: Fungi in Indoor Air Are Dominated by Outdoor Air and Show Dispersal Limitation at Short Distances." *The ISME Journal* 7: 1262–73.
- Aitchison, J. 1982. "The Statistical Analysis of Compositional Data." *Journal of the Royal Statistical Society: Series B: Methodological* 44: 139–60.
- Aitchison, J. 1986. *The Statistical Analysis of Compositional Data*. London: Chapman & Hall Ltd.
- Andersen, K., K. L. Bird, M. Rasmussen, J. Haile, H. Breuning-Madsen, K. H. Kjaer, L. Orlando, M. T. P. Gilbert, and E. Willerslev. 2012. "Meta-Barcoding of 'dirt' DNA from Soil Reflects Vertebrate Biodiversity." *Molecular Ecology* 21: 1966–79.
- Ando, H., H. Mukai, T. Komura, T. Dewi, M. Ando, and Y. Isagi. 2020. "Methodological Trends and Perspectives of Animal Dietary Studies by Noninvasive Fecal DNA Metabarcoding." *Environmental DNA* 2: 391–406.
- Andruszkiewicz, E. A., H. A. Starks, F. P. Chavez, L. M. Sassoubre, B. A. Block, and A. B. Boehm. 2017. "Biomonitoring of Marine Vertebrates in Monterey Bay Using eDNA Metabarcoding." *PLoS One* 12: e0176343.
- Barnes, M. A., and C. R. Turner. 2016. "The Ecology of Environmental DNA and Implications for Conservation Genetics." *Conservation Genetics* 17: 1–17.
- Beng, K. C., and R. T. Corlett. 2020. "Applications of Environmental DNA (eDNA) in Ecology and Conservation: Opportunities, Challenges and Prospects." *Biodiversity and Conservation* 29: 2089–121.
- Carøe, C., and K. Bohmann. 2020. "Tagsteady: A Metabarcoding Library Preparation Protocol to Avoid False Assignment of Sequences to Samples." *Molecular Ecology Resources* 20: 1620–31.
- Coghlan, S. A., C. A. Currier, J. Freeland, T. J. Morris, and C. C. Wilson. 2021. "Community eDNA Metabarcoding as a Detection Tool for Documenting Freshwater Mussel (Unionidae) Species Assemblages." *Environmental DNA* 3: 1172–91.
- Costello, M., M. Fleharty, J. Abreu, Y. Farjoun, S. Ferreira, L. Holmes, B. Granger, et al. 2018. "Characterization and Remediation of Sample Index Swaps by Non-redundant Dual Indexing on Massively Parallel Sequencing Platforms." *BMC Genomics* 19: 332.
- Creer, S., K. Deiner, S. Frey, D. Porazinska, P. Taberlet, W. K. Thomas, C. Potter, and H. M. Bik. 2016. "The Ecologist's Field Guide to Sequence-Based Identification of Biodiversity." *Methods in Ecology and Evolution* 7: 1008–18.
- De Filippis, F., M. Laiola, G. Blaiotta, and D. Ercolini. 2017. "Different Amplicon Targets for Sequencing-Based Studies of Fungal Diversity." *Applied and Environmental Microbiology* 83: e00905-17.
- Deagle, B. E., A. Chiaradia, J. McInnes, and S. N. Jarman. 2010. "Pyrosequencing Faecal DNA to Determine Diet of Little Penguins: Is What Goes in What Comes Out?" *Conservation Genetics* 11: 2039–48.

- Deagle, B. E., A. C. Thomas, A. K. Shaffer, A. W. Trites, and S. N. Jarman. 2013. "Quantifying Sequence Proportions in a DNA-Based Diet Study Using Ion Torrent Amplicon Sequencing: Which Counts Count?" *Molecular Ecology Resources* 13: 620–33.
- Egozcue, J. J., J. Graffelman, M. I. Ortego, and V. Pawlowsky-Glahn. 2020. "Some Thoughts on Counts in Sequencing Studies." *NAR Genomics and Bioinformatics* 2: lqaa094.
- Erb, I., G. B. Gloor, and T. P. Quinn. 2020. "Compositional Data Analysis and Related Methods Applied to Genomics—A First Special Issue from NAR Genomics and Bioinformatics." *NAR Genomics and Bioinformatics* 2: lqaa103.
- Everett, M. V., and L. K. Park. 2018. "Exploring Deep-Water Coral Communities Using Environmental DNA." *Deep Sea Research Part II: Topical Studies in Oceanography* 150: 229–41.
- Fernández, S., S. Rodríguez, J. L. Martínez, Y. J. Borrell, A. Ardura, and E. García-Vázquez. 2018. "Evaluating Freshwater Macroinvertebrates from eDNA Metabarcoding: A River Nalón Case Study." *PLoS One* 13: e0201741.
- Ficetola, G. F., C. Miaud, F. Pompanon, and P. Taberlet. 2008. "Species Detection Using Environmental DNA from Water Samples." *Biology Letters* 4: 423–5.
- Ford, M. J., J. Hempelmann, M. B. Hanson, K. L. Ayres, R. W. Baird, C. K. Emmons, J. I. Lundin, G. S. Schorr, S. K. Wasser, and L. K. Park. 2016. "Estimation of a Killer Whale (*Orcinus orca*) Population's Diet Using Sequencing Analysis of DNA from Feces." *PLoS One* 11: e0144956.
- Fukaya, K., H. Murakami, S. Yoon, K. Minami, Y. Osada, S. Yamamoto, R. Masuda, et al. 2021. "Estimating Fish Population Abundance by Integrating Quantitative Data on Environmental DNA and Hydrodynamic Modelling." *Molecular Ecology* 30: 3057–67.
- Gloor, G. B., J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. 2017. "Microbiome Datasets Are Compositional: And this Is Not Optional." *Frontiers in Microbiology* 8: 2224.
- Gohl, D. M., P. Vangay, J. Garbe, A. MacLean, A. Hauge, A. Becker, T. J. Gould, et al. 2016. "Systematic Improvement of Amplicon Marker Gene Methods for Increased Accuracy in Microbiome Studies." *Nature Biotechnology* 34: 942–9.
- Gold, Z., R. P. Kelly, A. O. Shelton, A. Thompson, K. D. Goodwin, R. Gallego, K. Parsons, L. R. Thompson, D. Kacev, and P. H. Barber. 2022. "Message in a Bottle: Archived DNA Reveals Impacts of a Marine Heatwave on Fish Assemblages over Multiple Decades." *Science Advances*. bioRxiv 2022.07.27. 501788. <https://doi.org/10.1101/2022.07.27.501788>.
- Hänfling, B., L. Lawson Handley, D. S. Read, C. Hahn, J. Li, P. Nichols, R. C. Blackman, A. Oliver, and I. J. Winfield. 2016. "Environmental DNA Metabarcoding of Lake Fish Communities Reflects Long-Term Data from Established Survey Methods." *Molecular Ecology* 25: 3101–19.
- Hanson, M. B., C. K. Emmons, M. J. Ford, M. Everett, K. Parsons, L. K. Park, J. Hempelmann, et al. 2021. "Endangered Predators and Endangered Prey: Seasonal Diet of Southern Resident Killer Whales." *PLoS One* 16: e0247031.
- Hoshino, T., and F. Inagaki. 2017. "Application of Stochastic Labeling with Random-Sequence Barcodes for Simultaneous Quantification and Sequencing of Environmental 16 S rRNA Genes." *PLoS One* 12: e0169431.
- Hoshino, T., R. Nakao, H. Doi, and T. Minamoto. 2021. "Simultaneous Absolute Quantification and Sequencing of Fish Environmental DNA in a Mesocosm by Quantitative Sequencing Technique." *Scientific Reports* 11: 1–9.
- Jerde, C. L., W. L. Chadderton, A. R. Mahon, M. A. Renshaw, J. Corush, M. L. Budny, S. Mysorekar, and D. M. Lodge. 2013. "Detection of Asian Carp DNA as Part of a Great Lakes Basin-Wide Surveillance Program." *Canadian Journal of Fisheries and Aquatic Sciences* 70: 522–6.
- Kelly, R. P., A. O. Shelton, and R. Gallego. 2019. "Understanding PCR Processes to Draw Meaningful Conclusions from Environmental DNA Studies." *Scientific Reports* 9: 1–14.
- Krehenwinkel, H., M. Wolf, J. Y. Lim, A. J. Rominger, W. B. Simison, and R. G. Gillespie. 2017. "Estimating and Mitigating Amplification Bias in Qualitative and Quantitative Arthropod Metabarcoding." *Scientific Reports* 7: 1–12.
- Lamb, P. D., E. Hunter, J. K. Pinnegar, S. Creer, R. G. Davies, and M. I. Taylor. 2019. "How Quantitative Is Metabarcoding: A Meta-Analytical Approach." *Molecular Ecology* 28: 420–30.
- Laporte, M., E. Reny-Nolin, V. Chouinard, C. Hernandez, E. Normandeau, B. Bougas, C. Côté, S. Behmel, and L. Bernatchez. 2021. "Proper Environmental DNA Metabarcoding Data Transformation Reveals Temporal Stability of Fish Communities in a Dendritic River System." *Environmental DNA* 3: 1007–22.
- Leonard, J. A., O. Shanks, M. Hofreiter, E. Kreuz, L. Hodges, W. Ream, R. K. Wayne, and R. C. Fleischer. 2007. "Animal DNA in PCR Reagents Plagues Ancient DNA Research." *Journal of Archaeological Science* 34: 1361–6.
- Lopes, C. M., T. Sasso, A. Valentini, T. Dejean, M. Martins, K. R. Zamudio, and C. F. Haddad. 2017. "eDNA Metabarcoding: A Promising Method for Anuran Surveys in Highly Diverse Tropical Forests." *Molecular Ecology Resources* 17: 904–14.
- Lynggaard, C., M. F. Bertelsen, C. V. Jensen, M. S. Johnson, T. G. Frøslev, M. T. Olsen, and K. Bohmann. 2022. "Airborne Environmental DNA for Terrestrial Vertebrate Community Monitoring." *Current Biology* 32: 701–7, e5.
- Macé, B., R. Hocdé, V. Marques, P.-E. Guerin, A. Valentini, V. Arnal, L. Pellissier, and S. Manel. 2022. "Evaluating Bioinformatics Pipelines for Population-Level Inference Using Environmental DNA." *Environmental DNA* 4: 674–86.
- McLaren, M. R., A. D. Willis, and B. J. Callahan. 2019. "Consistent and Correctable Bias in Metagenomic Sequencing Experiments." *eLife* 8: e46923.
- Miya, M., Y. Sato, T. Fukunaga, T. Sado, J. Poulsen, K. Sato, T. Minamoto, et al. 2015. "MiFish, a Set of Universal PCR Primers for Metabarcoding Environmental DNA from Fishes: Detection of More than 230 Subtropical Marine Species." *Royal Society Open Science* 2: 150088.
- Palmer, J. M., M. A. Jusino, M. T. Banik, and D. L. Lindner. 2018. "Non-biological Synthetic Spike-in Controls and the AMPtk Software Pipeline Improve Mycobiome Data." *PeerJ* 6: e4925.
- Pawlowsky-Glahn, V., and J. J. Egozcue. 2016. "Spatial Analysis of Compositional Data: A Historical Review." *Journal of Geochemical Exploration* 164: 28–32.
- Piñol, J., G. Mir, P. Gomez-Polo, and N. Agustí. 2015. "Universal and Blocking Primer Mismatches Limit the Use of High-Throughput DNA Sequencing for the Quantitative



- Metabarcoding of Arthropods.” *Molecular Ecology Resources* 15: 819–30.
- Pompanon, F., B. E. Deagle, W. O. Symondson, D. S. Brown, S. N. Jarman, and P. Taberlet. 2012. “Who Is Eating What: Diet Assessment Using Next Generation Sequencing.” *Molecular Ecology* 21: 1931–50.
- Port, J. A., J. L. O’Donnell, O. C. Romero-Maraccini, P. R. Leary, S. Y. Litvin, K. J. Nickols, K. M. Yamahara, and R. P. Kelly. 2016. “Assessing Vertebrate Biodiversity in a Kelp Forest Ecosystem Using Environmental DNA.” *Molecular Ecology* 25: 527–41.
- Rivera, S. F., F. Rimet, V. Vasselon, M. Vautier, I. Domaizon, and A. Bouchez. 2021. “Fish eDNA Metabarcoding from Aquatic Biofilm Samples: Methodological Aspects.” *Molecular Ecology Resources* 22: 1440–53.
- Schnell, I. B., K. Bohmann, and M. T. P. Gilbert. 2015. “Tag Jumps Illuminated—Reducing Sequence-to-Sample Misidentifications in Metabarcoding Studies.” *Molecular Ecology Resources* 15: 1289–303.
- Shelton, A. O., Z. J. Gold, A. J. Jensen, E. D’Agnese, E. A. Allan, A. Van Cise, R. Gallego, et al. 2022. “Code and Data for ‘Toward Quantitative Metabarcoding’.” Zenodo, Dataset. <https://doi.org/10.5281/zenodo.7158929>.
- Shelton, A. O., J. L. O’Donnell, J. F. Samhuri, N. Lowell, G. D. Williams, and R. P. Kelly. 2016. “A Framework for Inferring Biological Communities from Environmental DNA.” *Ecological Applications* 26: 1645–59.
- Shelton, A. O., A. Ramón-Laca, A. Wells, J. Clemons, D. Chu, B. E. Feist, R. P. Kelly, et al. 2022. “Environmental DNA Provides Quantitative Estimates of Pacific Hake Abundance and Distribution in the Open Ocean.” *Proceedings of the Royal Society B* 289: 20212613.
- Silverman, J. D., R. J. Bloom, S. Jiang, H. K. Durand, E. Dallow, S. Mukherjee, and L. A. David. 2021. “Measuring and Mitigating PCR Bias in Microbiota Datasets.” *PLoS Computational Biology* 17: e1009113.
- Silverman, J. D., K. Roche, S. Mukherjee, and L. A. David. 2020. “Naught all Zeros in Sequence Count Data Are the Same.” *Computational and Structural Biotechnology Journal* 18: 2789–98.
- Taberlet, P., A. Bonin, L. Zinger, and E. Coissac. 2018. *Environmental DNA: For Biodiversity Research and Monitoring*. New York: Oxford University Press.
- Taberlet, P., E. Coissac, F. Pompanon, C. Brochmann, and E. Willerslev. 2012. “Towards Next-Generation Biodiversity Assessment Using DNA Metabarcoding.” *Molecular Ecology* 21: 2045–50.
- Thomas, A. C., B. E. Deagle, J. P. Eveson, C. H. Harsch, and A. W. Trites. 2016. “Quantitative DNA Metabarcoding: Improved Estimates of Species Proportional Biomass Using Correction Factors Derived from Control Material.” *Molecular Ecology Resources* 16: 714–26.
- Thomsen, P. F., J. Kielgast, L. L. Iversen, P. R. Møller, M. Rasmussen, and E. Willerslev. 2012. “Detection of a Diverse Marine Fish Fauna Using Environmental DNA from Seawater Samples.” *PLoS One* 7: e41732.
- Tournayre, O., M. Leuchtman, O. Filippi-Codaccioni, M. Trillat, S. Piry, D. Pontier, N. Charbonnel, and M. Galan. 2020. “In Silico and Empirical Evaluation of Twelve Metabarcoding Primer Sets for Insectivorous Diet Analyses.” *Ecology and Evolution* 10: 6310–32.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Shelton, Andrew Olaf, Zachary J. Gold, Alexander J. Jensen, Erin D’Agnese, Elizabeth Andruszkiewicz Allan, Amy Van Cise, Ramón Gallego, et al. 2023. “Toward Quantitative Metabarcoding.” *Ecology* 104(2): e3906. <https://doi.org/10.1002/ecy.3906>