Received: 22 February 2022 | Revised: 31 July 2022 | Accepted: 16 August 2022

DOI: 10.1111/1755-0998.13703

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Extracting abundance information from DNA-based data

Mingjie Luo[1,2] | Yinqiu Ji[1] | David Warton[3,4] | Douglas W. Yu[1,5,6]

[1]State Key Laboratory of Genetic Resources and Evolution and Yunnan Key Laboratory of Biodiversity and Ecological Security of Gaoligong Mountain, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, China

[2]Kunming College of Life Sciences, University of Chinese Academy of Sciences, Kunming, Yunnan, China

[3]School of Mathematics and Statistics, UNSW Sydney, Sydney, New South Wales, Australia

[4]Evolution and Ecology Research Centre, UNSW Sydney, Sydney, New South Wales, Australia

[5]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, Yunnan, China

[6]School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk, UK

Correspondence
Douglas W. Yu, State Key Laboratory of Genetic Resources and Evolution and Yunnan Key Laboratory of Biodiversity and Ecological Security of Gaoligong Mountain, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, China.
Email: dougwyu@mac.com

Funding information
Key Research Program of Frontier Sciences, CAS, Grant/Award Number: QYZDY-SSW-SMC024; University of Chinese Academy of Sciences; Kunming Institute of Zoology; State Key Laboratory of Genetic Resources and Evolution, Grant/Award Number: GREKF19-01, GREKF20-01 and GREKF21-01; University of East Anglia; Chinese Academy of Sciences, Grant/Award Number: SMC024; idiv.de, Grant/Award Number: sCom; Strategic Priority Research Program, Chinese Academy of Sciences, Grant/Award Number: XDA20050202

Handling Editor: Lucie Zinger

Abstract

The accurate extraction of species-abundance information from DNA-based data (metabarcoding, metagenomics) could contribute usefully to diet analysis and food-web reconstruction, the inference of species interactions, the modelling of population dynamics and species distributions, the biomonitoring of environmental state and change, and the inference of false positives and negatives. However, multiple sources of bias and noise in sampling and processing combine to inject error into DNA-based data sets. To understand how to extract abundance information, it is useful to distinguish two concepts. (i) Within-sample *across-species* quantification describes relative species abundances in one sample. (ii) Across-sample *within-species* quantification describes how the abundance of each individual species varies from sample to sample, such as over a time series, an environmental gradient or different experimental treatments. First, we review the literature on methods to recover *across-species* abundance information (by removing what we call "species pipeline biases") and *within-species* abundance information (by removing what we call "pipeline noise"). We argue that many ecological questions can be answered with just *within-species* quantification, and we therefore demonstrate how to use a "DNA spike-in" to correct for pipeline noise and recover *within-species* abundance information. We also introduce a model-based estimator that can be used on data sets without a physical spike-in to approximate and correct for pipeline noise.

KEYWORDS
Arthropoda, biomonitoring, community composition, DNA barcoding, environmental DNA, Insecta, internal standard, polymerase chain reaction, taxonomic bias

1 | INTRODUCTION

The accurate extraction of species-abundance information from DNA-based data could contribute usefully to the reconstruction of diets and quantitative food webs, the inference of species interactions, the modelling of population dynamics and species distributions, the biomonitoring of environmental state and change, and more prosaically the inference of false positives and negatives (Abrego et al., 2021; Carraro et al., 2020, 2021; Deagle et al., 2019; Peel et al., 2019; Rojahn et al., 2021; Thomas et al., 2016). Here we use the term "abundance" to mean any estimate of biomass or count of individuals.

However, there are four general obstacles to the extraction of abundance information from DNA-based data (Griffin et al., 2020 for more formal treatments; see Shelton et al., 2016), which we will call here: (i) *species capture biases*, (ii) *capture noise*, (iii) *species pipeline biases* and (iv) *pipeline noise*.

1. *Species capture biases*. Different species are more or less likely to be captured by a given sampling method or via non-random sampling designs. For instance, Malaise traps preferentially capture Diptera (deWaard et al., 2019), and different fish species, body sizes and physiological conditions vary in their environmental DNA (eDNA) shedding rates (Thalinger et al., 2021; Yates, Glaser, et al., 2021).
2. *Capture noise*. Steinke et al. (2021) have shown that Malaise traps separated by only 3 m fail to capture the same species compositions, from which we infer that abundances vary stochastically across traps. Levi et al. (2019) showed that counts of salmon could be estimated via qPCR of aquatic eDNA, but only after correcting for temporal fluctuations in streamflow. Other sources of capture noise include environmental variation in eDNA degradation rates, food availability, PCR inhibitors and transport rates (reviewed in Yates, Cristescu, & Derry, 2021).
3. *Species pipeline biases*. Species differ in body size (biomass bias), genome size, mitochondrial copy number, DNA extraction efficiency and PCR amplification efficiency (primer bias) (Amend et al., 2010; Bell et al., 2017; Elbrecht & Leese, 2015; Garrido-Sanz et al., 2021; Iwaszkiewicz-Eggebrecht et al., 2022; Krehenwinkel et al., 2017; McLaren et al., 2019; Pauvert et al., 2019; Piñol et al., 2015, 2019; Tang et al., 2015; Yang et al., 2021; Yu et al., 2012). Species can even differ in their propensity to survive a bioinformatic pipeline, such as when closely related species are clustered into one operational taxonomic unit (OTU) (Pauvert et al., 2019).
4. *Pipeline noise*. There is considerable noise in DNA-based pipelines, which breaks the relationship between starting sample biomasses and final numbers of reads per sample (Ji et al., 2020), caused in part by PCR stochasticity and the passing and pooling of small aliquots of liquid along wet-laboratory pipelines. In particular, it is common practice *to deliberately equalize the amount of data per sample* by "pooling samples in equimolar concentration" just before sequencing.

We do not consider species capture biases or capture noise further, referring the reader to the literature on eDNA occupancy correction (Doi et al., 2019; e.g. Dorazio & Erickson, 2018; Erickson, 2019; Griffin et al., 2020; Lyet et al., 2021; Stauffer et al., 2021) and the review by Yates, Cristescu, and Derry (2021). Instead, our purpose is to review methods for the extraction of abundance information from already-collected samples, because even if species capture biases and capture noise can be corrected, the combination of species pipeline biases and pipeline noise still *causes the number of DNA sequences assigned to a species in a sample to be an error-prone measure of the abundance of that species in that sample* (McLaren et al., 2019).

To start, we illustrate in a simplified way how pipeline noise and species pipeline biases (hereafter, species biases) combine to inject error into DNA-based data sets. We start with a notionally true sample × species table or OTU table (Figure 1), where OTU stands for operational taxonomic unit (i.e., a species hypothesis). Let each cell represent the true abundance (biomass or count) of that OTU in that sample.

Pipeline noise affects the *rows* (samples) of an OTU table. Thus, even though in the true table, OTU1 is six times as abundant in sample 4 vs. sample 1 (green cells in Figure 1a), in the observed table, OTU1 is only *two* times as abundant in sample 4 (green cells in Figure 1b). *Pipeline noise thus obscures how the abundance of each individual species varies across samples, where the samples could be a time series, an environmental gradient or different experimental treatments.*

Species bias affects the *columns* (OTUs) of an OTU table. Thus, even though in the true table, OTU2 and OTU1 are equally abundant in sample 3 (orange cells in Figure 1a), in the observed OTU table, OTU2 is two times as abundant as OTU1 in sample 3 (orange cells in Figure 1b). Species bias thus obscures *relative species abundances*, which is important for diet analysis (Deagle et al., 2019) and when relative abundance within a sample provides information on species contribution to ecosystem functioning or services (e.g., relative fish species biomasses).

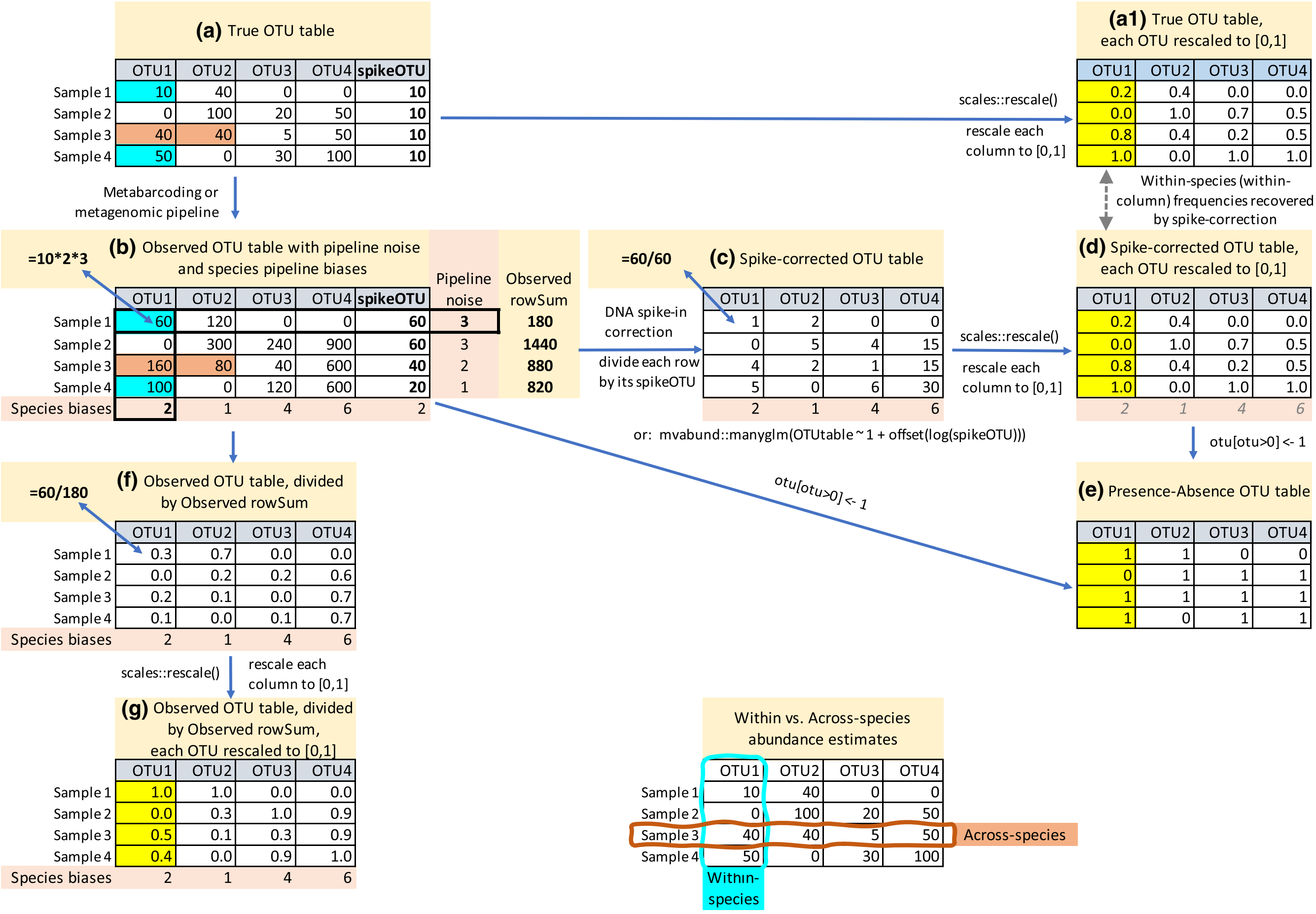
So how can we extract abundance information from DNA-based data? It is helpful to distinguish between two concepts from Ji et al. (2020; see also Garrido-Sanz et al., 2021; Ji et al., 2020):

1. *Within*-species quantification: for example, "Species A is more abundant in this sample than it is in that sample (e.g., two points on a time series)." This is achieved by removing pipeline noise (Figure 2a1,d).
2. *Across*-species quantification: for example, "Species A is more abundant than Species B *in this sample* (i.e., relative species abundance)." This is achieved by removing species biases.

We can state this mathematically as:

$$\log(\mu_{ij}) = a_i + a_j + \mathbf{x}_i'\mathbf{b} + \mathbf{x}_i'\mathbf{b}_j \quad (1)$$

where μ_{ij} is the abundance of species j in sample i, a_i is a measure of the overall abundance of a sample, a_j is a measure of how abundant species j is across samples, and we assume a vector of environmental variables $\mathbf{x}_i$ (whose transpose is $\mathbf{x}_i'$) haa an effect on total


scales::rescale()
rescale each column to [0,1]
Within-species (within-column) frequencies recovered by spike-correction
Metabarcoding or metagenomic pipeline
=10*2*3
=60/60
DNA spike-in correction
divide each row by its spikeOTU
scales::rescale()
rescale each column to [0,1]
or: mvabund::manyglm(OTUtable ~ 1 + offset(log(spikeOTU)))
otu[otu>0] <- 1
otu[otu>0] <- 1
=60/180
scales::rescale()
rescale each column to [0,1]
Within-species
Across-species

(a) True OTU table

	OTU1	OTU2	OTU3	OTU4	spikeOTU
Sample 1	10	40	0	0	10
Sample 2	0	100	20	50	10
Sample 3	40	40	5	50	10
Sample 4	50	0	30	100	10

(a1) True OTU table, each OTU rescaled to [0,1]

	OTU1	OTU2	OTU3	OTU4
	0.2	0.4	0.0	0.0
	0.0	1.0	0.7	0.5
	0.8	0.4	0.2	0.5
	1.0	0.0	1.0	1.0

(b) Observed OTU table with pipeline noise and species pipeline biases

	OTU1	OTU2	OTU3	OTU4	spikeOTU	Pipeline noise	Observed rowSum
Sample 1	60	120	0	0	60	3	180
Sample 2	0	300	240	900	60	3	1440
Sample 3	160	80	40	600	40	2	880
Sample 4	100	0	120	600	20	1	820
Species biases	2	1	4	6	2		

(c) Spike-corrected OTU table

OTU1	OTU2	OTU3	OTU4
1	2	0	0
0	5	4	15
4	2	1	15
5	0	6	30
2	1	4	6

(d) Spike-corrected OTU table, each OTU rescaled to [0,1]

OTU1	OTU2	OTU3	OTU4
0.2	0.4	0.0	0.0
0.0	1.0	0.7	0.5
0.8	0.4	0.2	0.5
1.0	0.0	1.0	1.0
2	*1*	*4*	*6*

(e) Presence-Absence OTU table

OTU1	OTU2	OTU3	OTU4
1	1	0	0
0	1	1	1
1	1	1	1
1	0	1	1

(f) Observed OTU table, divided by Observed rowSum

	OTU1	OTU2	OTU3	OTU4
Sample 1	0.3	0.7	0.0	0.0
Sample 2	0.0	0.2	0.2	0.6
Sample 3	0.2	0.1	0.0	0.7
Sample 4	0.1	0.0	0.1	0.7
Species biases	2	1	4	6

(g) Observed OTU table, divided by Observed rowSum, each OTU rescaled to [0,1]

	OTU1	OTU2	OTU3	OTU4
Sample 1	1.0	1.0	0.0	0.0
Sample 2	0.0	0.3	1.0	0.9
Sample 3	0.5	0.1	0.3	0.9
Sample 4	0.4	0.0	0.9	1.0
Species biases	2	1	4	6

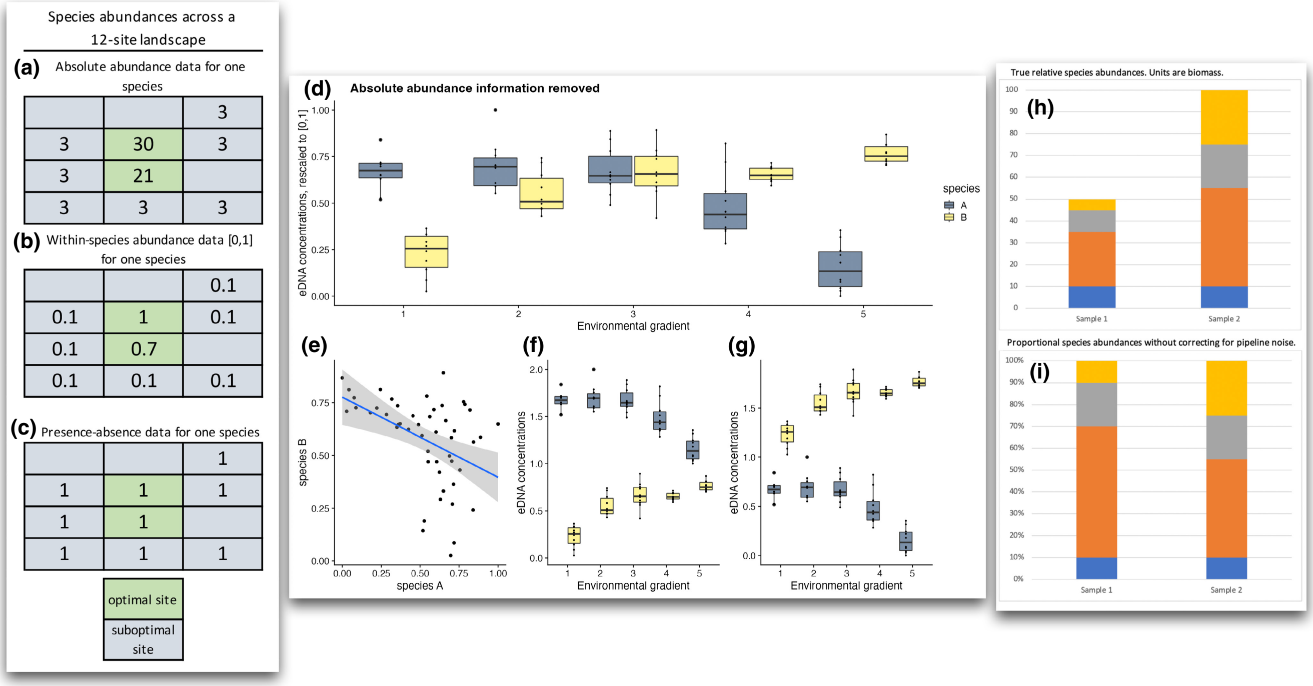
Within vs. Across-species abundance estimates

	OTU1	OTU2	OTU3	OTU4
Sample 1	10	40	0	0
Sample 2	0	100	20	50
Sample 3	40	40	5	50
Sample 4	50	0	30	100

FIGURE 1 Pipeline noise vs. species bias in OTU tables. (a) The true OTU table, with cell numbers representing the true abundance of DNA for each OTU (column) in each sample (row). The spikeOTU column shows that the same amount of DNA spike-in has been added to each sample. (a1) The true OTU table after rescaling each OTU column to the interval [0,1]. (b) The observed OTU table after amplicon sequencing, showing the combined effects of pipeline noise and species biases. Each cell in Table (a) is multiplied by the Pipeline noise and Species bias values in that cell's row and column. For instance, OTU1's true abundance in Sample 1 is 10 but appears as 60 (=10 × 2 × 3). Pipeline noise thus causes the original 10:50 ratio of OTU1 in Samples 1 and 4 (blue cells) to appear as 60:100, while species bias causes the original 40:40 ratio of OTU1 and OTU2 (orange cells) to appear as 160:80. (c) The observed OTU table after dividing each row by its observed spike-in reads, which removes pipeline noise. Note that species biases remain uncorrected. In statistical modelling, the observed spike-in values are an index of sampling (sequencing) effort and can be included as offset values. (d) Table (c) after rescaling each column to the interval [0,1], to allow direct comparison with the rescaled true-OTU Table (a1). Spike-in correction successfully recovers *within-species* abundance change from sample to sample. Species biases have not been removed but have now been ignored via rescaling. (e) If spike-in reads are not available, or if it is suspected that capture noise is uncorrectable and high, the observed OTU table can be transformed to presence/absence. However, this method loses ecological information (Figure 2c). (f) Pipeline noise cannot be reliably removed by using the total reads per sample as a proxy for sampling effort (Observed rowSum) because the observed rowSum is confounded by species composition. (g) Table (f) after rescaling each OTU column to the interval [0,1], to contrast with Tables (a1) and (d). Line graphs of the OTU tables are in the spreadsheet version of this table, in the Supporting Information. Code syntax from the R language (R Core Team, 2021), including the {mvabund} (Wang et al., 2012) and {scales} packages (https://scales.r-lib.org, accessed December 16, 2021)

abundance (via $\boldsymbol{b}$) as well as having a compositional effect, that is affecting different species in different ways (via $\boldsymbol{b}_j$). The responses to environmental variables ($\boldsymbol{b}$ and $\boldsymbol{b}_j$) are typically the main quantities of biological interest, being used to model and monitor species distributions. Pipeline noise biases our estimate of a_i, which would be zero for identical replicates in the absence of stochasticity, which in turn biases estimates of effects of environmental variables ($\boldsymbol{b}$ and $\boldsymbol{b}_j$). Species pipeline biases affect our estimate of a_j, affecting across-species quantification.

As we review and demonstrate below, some approaches remove pipeline noise, some remove species biases and some remove both. *Our take-home message is that removing only pipeline noise to achieve within-species quantification can be enough* to improve the inference of species interactions, the modelling of population dynamics and species distributions, the biomonitoring of environmental state and change, and the inference of false positives and negatives (Abrego et al., 2021; Carraro et al., 2020, 2021; Rojahn et al., 2021, and Figure 2).



(a)

		3
3	30	3
3	21	
3	3	3

(b)

		0.1
0.1	1	0.1
0.1	0.7	
0.1	0.1	0.1

(c)

		1
1	1	1
1	1	
1	1	1

FIGURE 2 The usefulness of within-species abundance information. (a) Imagine that a species is found in many sites but that only two sites are optimal (green cells) with high abundances, with the rest suboptimal (grey cells), with low abundances. (b) Even though species pipeline biases make it difficult to recover absolute abundances from DNA-based data, it is straightforward to use a spike-in to recover within-species abundance data (shown by rescaling to the interval [0,1]), still revealing that the green sites appear to be optimal habitat. (c) If the DNA-based data were converted to presence/absence, the distinction between green and grey habitat would be lost. (d) Along an environmental gradient from left to right, let species A decrease and species B increase in eDNA concentration (rescaled to [0,1]). (e) The two species are seen to be negatively correlated over the gradient, even though absolute abundance information is unavailable. Example adapted from Rojahn et al. (2021), who combined a similar result with additional information to infer the competitive exclusion of a native fish species by an invasive species. (f,g) Due to species pipeline biases, absolute abundances are not known, and either species A or B could be absolutely more abundant. (h) Two samples with different absolute and relative species abundances. (i) If only across-species quantification is achieved (e.g., via forward or reverse metagenomics), it is valid to compare species *within* each sample only, revealing that the dark orange species has the highest relative abundance in both samples 1 and 2. However, it would not be valid to conclude that the dark orange species has a greater absolute biomass in Sample 1 than in Sample 2, as can be seen by inspection of the true absolute abundances in (h). However, also achieving within-species quantification (via a spike-in) would make it possible to compare how each species' *absolute* abundances vary across samples

1.1 | Mini-review of methods to extract abundance information

1.1.1 | Multiplexed individual barcoding

The most straightforward approach is to DNA-barcode all the individual organisms and count them up, which achieves both within- and across-species quantification. This method only works on taxa that have body sizes suitable for separating individuals, such as bees (Gueuning et al., 2019). Once separated, individuals or portions thereof (e.g., a leg) are placed in separate wells of a 96-well plate and individually PCR'd. Each PCR requires a uniquely tagged pair of PCR primers, which allows all the PCR products to be pooled and then sequenced *en masse* on Illumina (Creedy et al., 2020; Meier et al., 2016; Ratnasingham, 2019), PacBio (Hebert et al., 2018) or MinION (Srivathsan et al., 2021) sequencers. This method now costs much less than $1 per individual. Wührl et al. (2022) further increase throughput with a robotic pipettor and camera that visually identifies small insects to higher taxonomic rank and sorts them into 96-well plates. However, this method is difficult to apply to very large numbers of individuals and cannot be applied to trace DNA or microbial taxa. Note that this approach could also be carried out via machine-learning-accelerated visual identifications of photos of arthropods (Schneider et al., 2022).

1.1.2 | Presence–absence in multiple subsamples

Presence–absence across multiple subsamples can be used as an index of within-species abundance. For instance, Abrego et al. (2021) summed all weekly detections (presences) per species in their mitogenomic arthropod data set to estimate an annual abundance measure for each species. However, pipeline noise can still be reflected in presence/absence data, albeit more weakly, especially when many subsamples are used. This method can achieve partial within-species quantification but probably not across-species quantification.

1.1.3 | Design less biased PCR primers

In some cases, the target taxon is nearly uniform in body size and DNA-extraction efficiency, and it can be possible to design PCR primers that bind similarly across species. For instance, Schenk et al. (2019) have reported that primers for the 28S D3–D5 and 18S V4 regions return nematode read frequencies that accurately recover relative species abundances, Verkuil et al. (2022) have reported that modified *COI* primers can recover the relative biomasses of insect orders from pied flycatcher faeces, and Ershova et al. (2021) have reported that increasing *COI* primer degeneracy (Leray-XT) can recover relative biomasses of marine zooplankton. This method achieves across-species quantification, albeit with error, but not within-species quantification.

1.1.4 | Quantitative/digital-droplet PCR

qPCR and ddPCR (quantitative and digital droplet PCR) can be used to estimate the sample DNA concentration of one species per assay. ddPCR is more sensitive than qPCR (Brys et al., 2021) and allows the detection of single copies of target DNA and absolute quantification through the partitioning of the PCR into 20,000 droplets and subsequent fluorescent detection of droplets that contain the target DNA (Hindson et al., 2011). The present paper does not review q/ddPCR except to note that single-species q/ddPCR applied to aquatic trace DNA can achieve within-species quantification, provided that one corrects for capture bias and noise in the form of variation in water discharge rates, surface-area to mass ratio, and/or eDNA transport and diffusion (Fukaya et al., 2021; Levi et al., 2019; Pochardt et al., 2020; Rourke et al., 2022; Shelton, Ramón-Laca, et al., 2022; Yates, Cristescu, & Derry, 2021; Yates, Glaser, et al., 2021). If applied to multiple species and if statistical models that relate DNA copy number to abundance can be fitted (Fukaya et al., 2021; Levi et al., 2019; Pochardt et al., 2020), then across-species quantification can also be achieved, albeit with nontrivial amounts of error. See also Rourke et al. (2022) for a recent, comprehensive review.

1.1.5 | Spike-in DNA

To achieve within-species quantification, researchers have advocated adding a fixed amount of an arbitrary DNA sequence to each sample, after tissue lysis and before DNA extraction. This "spike-in," also known as an internal standard (ISD, Harrison et al., 2021), must have a sequence that does not match any species that could be in the samples and be flanked by primer binding sequences that match the primers used to amplify the samples (Deagle et al., 2018; Harrison et al., 2021; Smets et al., 2016; Tkacz et al., 2018; Tsuji et al., 2022; Ushio et al., 2018). By design, each sample receives the same amount of spike-in, and all samples should therefore return the same number of spike-in reads after PCR and sequencing. However, due to pipeline noise, some samples return more spike-in reads because more of the sample's DNA made it through the metabarcoding pipeline; those samples have OTUs with "too many reads." Some samples return fewer spike-in reads because less of the sample's DNA made it through the metabarcoding pipeline; those samples have OTUs with "too few reads." The correction step is simple: divide each sample's OTU sizes by the number of spike-in reads in that sample (Abrego et al., 2021; Ji et al., 2020). OTUs in samples with large numbers of spike-in reads are reduced in size more than OTUs in samples with small numbers of spike-in reads. Alternatively, the number of spike-in reads per sample can be input as an offset term in a multivariate statistical model (Wang et al., 2012). This latter approach can be understood as estimating a_i in Equation 1 using $\hat{a}_i = \ln \sum_{j=1}^{q} z_{ij}$ where we have spike-in reads (z_{ij}) for q species (or synthetic sequences).

As an example, and following the pioneering work of Zhou et al. (2013), Ji et al. (2020) mapped whole-genome-sequenced (WGS, aka "shotgun sequencing") data sets of insects to mitochondrial genomes and barcodes and achieved nearly perfect within-species quantification (barcodes $R^2 = 93\%$, mitogenomes $R^2 = 95\%$) and almost direct proportionality between mapped reads and input DNA mass. The high accuracy was largely achieved by using a spike-in correction. However, the regression lines that related read number to input DNA for each species all had different intercepts, reflecting uncorrected species biases, and thus across-species quantification was not achieved. Harrison et al. (2021) provide an excellent, complementary review of the recent literature on spike-ins and also describe an alternative approach for modelling nonspike-corrected ("compositional") data sets. Figure 1 provides a worked example of spike-in correction, and the Excel spreadsheet used to produce Figure S1 is provided in the Supporting Information.

1.1.6 | Model-based pipeline-noise estimation

A related approach is to try to use the data itself to estimate the pipeline noise, rather than a physical spike-in. To do this we could fit the model stated in Equation 1 to data. However, fitting this full model with row effects can be computationally intensive, especially for large data sets, so a simple alternative is to approximate a_i using a one-step estimator (Warton, 2022):

$$\tilde{a}_i = \log \sum_{j=1}^{p} y_{ij} - \log \sum_{j=1}^{p} \hat{\mu}_{ij}^{(0)} \qquad (2)$$

where y_{ij} is the number of reads for OTU j in sample i, $\hat{\mu}_{ij}^{(0)}$ is its predicted value from a model that does not include a row effect and p is the total number of OTUs. We can then include $\tilde{a}_i$ as an offset in future models to (approximately) correct for pipeline bias.

The reason Equation 2 has two terms in it is that there are two reasons that a sample might end up generating many sequence reads: by chance (pipeline noise) and/or because some (or many) of the OTUs are abundant in the site where the sample was taken (ecology). Thus, if one has informative predictors $\boldsymbol{x}_i$ that can successfully predict which OTUs should be abundant in which samples, then it becomes possible to separate the two effects. log $\sum_{j=1}^{p} y_{ij}$ is a function of both effects, log $\sum_{j=1}^{p} \hat{\mu}_{ij}^{(0)}$ estimates the effect of the predictors on the OTUs (ecology) and their difference isolates the row effect (pipeline noise). This is related to the spike-in approach, the main difference being that the spike-in formula (for $\hat{a}_i$) has no second term involving μ_{ij} because, by design, the same amount of each spike-in species is included in every sample (the spike-in has no ecology). An important difference here however is that because the same data are being used to estimate both pipeline noise ($\boldsymbol{a}_i$) and ecological effects ($\boldsymbol{b}$), it will be difficult to tease these effects apart if the two are correlated. In fact, the common practice of adjusting samples to equimolar concentration before sequencing confounds these two effects. This problem does not however affect estimation of compositional effects ($\boldsymbol{b}_j$), often the main quantity of interest.

1.1.7 | Unique molecular identifiers (UMIs)

A UMI is a series of ~7–12 random bases ("NNNNNNN") added to the forward primer as an ultrahigh-diversity tag (Hoshino & Inagaki, 2017). Seven Ns produce $4^7 = 16{,}384$ uniquely identified forward primer molecules. Species contributing abundant DNA to a sample will capture many of these primer molecules and thus amplify many unique UMIs, while species contributing scarce DNA will amplify a low number. The relationship between UMI richness and DNA abundance is roughly linear but asymptotes for species with very high DNA abundance. After sequencing, the number of UMIs per OTU correlates with the starting number of template DNA molecules per species in that sample (Hoshino et al., 2021; Hoshino & Inagaki, 2017). This method thus mimics q/ddPCR in that if statistical relationships between DNA copy number and true abundance can be estimated, across-species quantification can be achieved. Within-species quantification can be achieved by also adding a spike-in.

1.1.8 | Estimate and eliminate PCR bias

Silverman et al. (2021) propose a straightforward way to estimate PCR bias, by pooling all samples to ensure that all species are present, and subjecting the pooled sample to different numbers of PCR cycles x_i , from low to high. For any given pair of species 1 and 2, the ratio of their reads $\frac{w_{i1}}{w_{i2}}$ after a given number of cycles is their starting DNA ratio $\frac{a_1}{a_2}$ multiplied by their relative amplification bias $\left(\frac{b_1}{b_2}\right)^{x_i}$, which increases with the number of cycles. This relationship can be linearized, and given the post-PCR relative read numbers at all cycle numbers, starting DNA ratios (and relative amplification biases) can be estimated.

$$\frac{w_{i1}}{w_{i2}} = \frac{a_1}{a_2}\left(\frac{b_1}{b_2}\right)^{x_i} \tag{3}$$

However, PCR is not the only source of species pipeline bias (e.g., Iwaszkiewicz-Eggebrecht et al., 2022), and McLaren et al. (2019) have pointed out that although it is not possible to estimate a priori the whole set of species biases in a given amplicon or metagenomic data set (because an unknown number of factors of unknown strengths combine to create the biases), it is reasonable to assume that *the ratio of the biases of every pair of species is fixed*. Given this, Williamson et al. (2021); see also Clausen & Willis, 2022) showed that if first one is able to estimate the absolute abundances of a *subset* of species in the samples (via multiple, species-specific q/ddPCR assays or flow cytometry), it is possible to infer the absolute abundances of all the species by inferring their ratios with the q/ddPCR-quantified species, allowing one to achieve across-species quantification. The authors dub this a "multiview data structure" because there are two views into the community of interest: q/ddPCR and sequencing. Note that because q/ddPCR is carried out after many of the wet-laboratory steps have been carried out, multiview modelling does not remove pipeline noise, and a spike-in is still needed to achieve within-species quantification. Shelton, Gold, et al., 2022) advocate a similar approach but via the construction of a "mock community" containing tissue of all species of interest and subjecting it to the same PCR protocol as the samples. From this mock community, species-specific PCR biases are calculated and used to extract across-species abundance information.

1.1.9 | Forward and reverse metagenomics

Another way to achieve across-species quantification is to avoid PCR by using a metagenomic approach. For marine phytoplankton, Pierella Karlusich et al. (2022) used shotgun-sequenced counts of the (mostly) single-copy *psbO* gene, which is part of the photosystem II complex, to estimate species relative abundances. For land plants, Lang et al. (2019) showed that WGS data sets from pollen samples mapped to the variable protein-coding regions in chloroplast genomes can achieve accurate across-species quantification, finding that read frequency correlated strongly and linearly with pollen-grain frequency in a nearly 1:1 relationship ($R^2 = 86.7\%$, linear regression). At the same time, Peel et al. (2019) showed that it is possible to skip the labour of assembling and annotating chloroplast genomes, by using long-read sequences produced by the MinION sequencers from Oxford Nanopore Technologies (ONT). In this protocol, unassembled genome skims of individual plant species, ideally sequenced at ≥1.0× depth, are used as reference databases. Mixed-species query samples of pollen are sequenced on MinIONs. The reads from each (reference) genome skim are mapped to each (query) long read, and each long read is assigned to the species whose skim maps at the highest percentage coverage. This "reverse metagenomic" (RevMet) protocol achieves across-species quantification, allowing biomass-dominant species to be identified in mixed-species pollen samples (and potentially, in root masses). Because RevMet uses the whole genome, it avoids species biases and pipeline noise created by ratios of chloroplasts to cells varying across species, condition, tissues and age, and it can potentially be applied to any taxon for which it is possible to generate large numbers of individual genome skims, potentially including soil fauna. However, metagenomics by itself does not remove pipeline noise and would have to be paired with a spike-in to achieve within-species quantification.

To sum up, multiple methods exist to extract abundance information from DNA-based data sets (Table 1). Some achieve within-species quantification by removing pipeline noise, some achieve across-species quantification by removing species biases, and some achieve both or can be combined to achieve both.

It is useful to understand that many ecological questions can be tackled with only within-species quantification (Figure 2). In the second half of this paper, we therefore provide a detailed protocol and experimental validation of spike-ins to achieve within-species quantification for metabarcoding data sets.

We carry out two tests. First, we start with a sample of known composition (a “mock soup” of 52 OTUs), and from this we create a dilution gradient of seven samples with a spike-in. We show the successful use of the spike-in correction to remove pipeline noise and recover the dilution gradient. We then repeat the experiment with seven Malaise trap samples, which have the advantage of being more realistic but the disadvantage of having unknown compositions. Again, we show the successful use of the spike-in to recover the seven dilution gradients made from the seven samples.

2 | METHODS

2.1 | Mock soup construction

In total, 286 arthropods were collected in Kunming, China (25°8′23″ N, 102°44′17″ E) (Luo et al., 2022). DNA was extracted from each individual using the DNeasy Blood & Tissue Kit (Qiagen). Genomic DNA concentration of each individual was quantified from three replicates using PicoGreen fluorescent dye. Then, 658-bp *COI* barcoding sequences were PCR'd with Folmer primers (LCO1490 and HCO2198) (Folmer et al., 1994) and Sanger-sequenced. After the 658-bp *COI* sequences were trimmed to 313 bp based on our metabarcoding primers (see 2.4 Primer design), 286 arthropods were clustered to 168 OTUs at 97% similarity. We selected 52 individuals with genomic DNA >20 ng μl^{-1}, representing 52 OTUs.

We created a mock-soup gradient of seven dilution levels. First, we created the highest concentration-level soup by pooling 61 ng of each of the 52 OTUs. The next soup was created by pooling 48.8 ng (=0.8 × 61) of each of the 52 OTUs, and so on to create a gradient of seven mock soups of differing absolute abundances, stepping down 0.8× each time. To make it possible to check for mundane experimental error (as opposed to failure of the spike-in to recover the gradient), we independently created this mock-soup gradient three times, for $n_{\text{tot}} = 21$ independent poolings (Figure 3a).

2.2 | Preparation of Malaise trap samples

In total, 244 Malaise trap samples from 96 sites, using 99.9% ethanol as the trapping liquid, were collected in and near a 384-km^2 forested landscape containing the H.J. Andrews Experimental Forest (44.2° N, 122.2° W), Oregon, USA, in July 2018 (Luo et al., 2022). Traps were left for seven nonrainy days. To equalize biomass across individuals, we only kept the heads of large individuals (body lengths >2 cm) and then transferred the samples to fresh 99.9% ethanol to store at room temperature until extraction. The samples were air dried individually on filter papers for less than 1 h and then transferred to 50-ml tubes or 5-ml tubes according to sample volume. The samples were then weighed. DNA was nondestructively extracted by soaking the samples in lysis buffer, using the protocols from Ji et al. (2020) and Nielsen et al. (2019). For this study, we selected seven samples spread over the study area, each of which is an independent test of our ability to recover the dilution gradient. After

TABLE 1 Summary of reviewed methods for extracting abundance information from DNA-based data. Each method is scored for whether it can achieve within-species or across-species quantification or both

Method	Description	Within-species abundance	Across-species abundance
Multiplexed individual barcoding	DNA-barcode every individual in every sample	✓	✓
Presence–absence in multiple subsamples	Take multiple subsamples and count presences	✓	?
Design less biased PCR primers	Self-explanatory		✓
Quantitative/digital-droplet PCR	Quantify a species' DNA concentration per sample	✓	✓ (with extra work)
Spike-in DNA	Add a fixed amount of external DNA to each sample to measure pipeline noise	✓	
Model-based pipeline-noise estimation	Estimate the effect of pipeline noise by removing the effect of environmental predictors	✓	
Unique molecular identifiers (UMIs)	Estimate the amount of starting DNA per sample and per species	✓	?
Estimate and eliminate PCR bias	Use calibration samples and/or PCR time series to estimate species-specific PCR biases		✓
Forward and reverse metagenomics	Map and count shotgun reads to reference sequences		✓

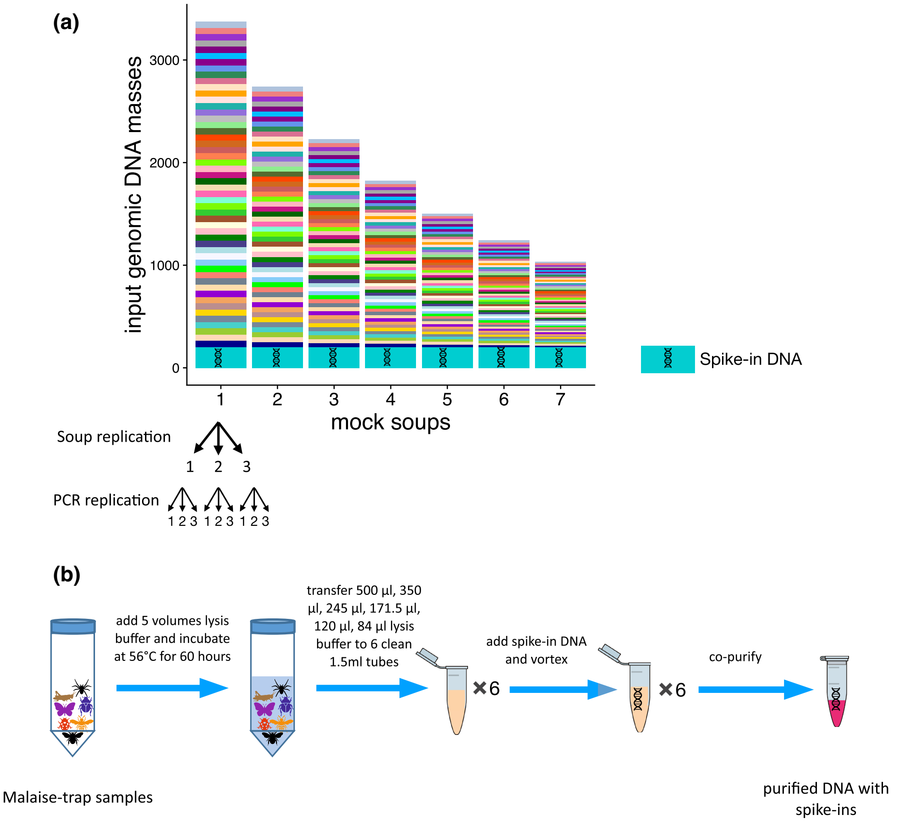

(a)
input genomic DNA masses
0
1000
2000
3000
1
2
3
4
5
6
7
mock soups
Spike-in DNA
Soup replication
1 2 3
PCR replication
1 2 3 1 2 3 1 2 3
(b)
add 5 volumes lysis buffer and incubate at 56°C for 60 hours
transfer 500 μl, 350 μl, 245 μl, 171.5 μl, 120 μl, 84 μl lysis buffer to 6 clean 1.5ml tubes
×6
add spike-in DNA and vortex
×6
co-purify
Malaise-trap samples
purified DNA with spike-ins

FIGURE 3 Preparation of mock and Malaise trap soups. (a) Mock soups. Each mock soup was constructed with equal masses of purified DNA from 52 OTUs. From soup "a" to soup "g," the input genomic masses of each of the 52 OTUs were 61, 48.8, 39, 31.2, 25, 20 and 16 ng. The same mass of spike-in DNA was then added to each soup (green DNA molecule). Each of the seven soups was made in triplicate, and all 21 soups were PCR'd in triplicate following the BEGUM pipeline (Yang et al., 2021) to detect and remove false reads. (b) Malaise trap sample protocol. Each bulk sample of arthropods was nondestructively DNA-extracted by soaking in 5× volume of lysis buffer. From each of the seven samples, 500, 350, 245, 171.5, 120 and 84 μl lysis buffer was used to create six dilution soups, a fixed amount of spike-in DNA was added, and the mixture was copurified

completion of lysis, we serially diluted the seven samples by using 0.7× lysis buffer volume (500, 350, 245, 171.5, 120 and 84 μl) to create six soups per sample (n_{tot} = 42). We used a QIAquick PCR purification kit (Qiagen) following the manufacturer instructions to purify lysis buffer on one spin column per soup (Figure 3b). We used a shallower gradient (0.7×) because our starting DNA amount was lower than with the mock soups.

2.3 | Adding spike-in DNA

2.3.1 | Spike-in DNA

For our spike-ins, we used three insect species from China (Lepidoptera: Bombycidae, Coleoptera: Elateridae, Coleoptera: Mordellidae), none of which is expected to appear in the Oregon Malaise trap samples. An alternative is to use one or more synthetic, random DNA sequences (Tkacz et al., 2018). Each of our three spike-ins is represented by a 658-bp *COI* fragment (Table S1) with primer binding sites that match the Folmer primers HCO2198 and LCO1490. For long-term storage, we inserted the *COI* fragments as plasmids into monoclonal bacteria. Plasmids were extracted using a TIANprep Mini Plasmid Kit following the manufacturer's instructions.

2.3.2 | Adding spike-in to the mock soups

Adding too much spike-in wastes sequencing data, while adding too little risks loss of abundance information in at least some samples when the number of spike-in reads is too low to use as a reliable correction factor. Thus, we quantified the *COI* copy numbers of the mock soups and the spike-in DNA by qPCR (Table S2, Figure S1) and chose a volume so that spike-in reads should make up 1% of the total number of *COI* copies in the lowest-concentration mock soups, balancing efficiency with reliability. We used all three spike-in species here and mixed them (Bombycidae/Elateridae/Mordellidae) in a ratio of 1:2:4, which was added directly to the mock soups' DNA since they were already purified.

2.3.3 | Adding spike-in to the Malaise trap samples

From the 244 Malaise trap samples, we first extracted 17 Malaise trap samples without adding spike-ins, and then we used qPCR to quantify the mean *COI* concentrations of these 17 samples in order to decide how much spike-in to add. Before adding the spike-ins, we discovered that the Bombycid DNA spike-in had degraded, and so we used only two spike-in species for the Malaise trap samples, at a ratio of 1:9 (Mordellidae/Elateridae). We then chose seven other samples for this study. In these samples, lysis buffer (500, 350, 245, 171.5, 120, 84 μl) from each sample was transferred into clean 1.5-ml tubes, and the spike-in DNA was added. We then purified the DNA with the Qiagen QIAquick PCR purification kit, following the manufacturer's instructions. DNA was eluted with 200 μl of elution buffer. In this way, the spike-in DNA was co-purified, co-amplified and co-sequenced along with the sample DNA (Figure 3b). We also recorded the total lysis buffer volume of each sample, for downstream correction.

2.4 | Primer design

For this study, we simultaneously tested two methods for extracting abundance information: spike-ins and UMIs. UMI tagging requires a two-step PCR procedure (Hoshino & Inagaki, 2017; Lundberg et al., 2013), first using tagging primers and then using amplification primers (Figure S2). The tagging primers include (i) the Leray-FolDegenRev primer pair to amplify the 313-bp *COI* amplicon of interest, (ii) a 1- or 2-nucleotide heterogeneity spacer on both the forward and reverse primers to increase sequence entropy for the Illumina sequencer, (iii) the same 6-nucleotide sequence on both the forward and reverse primers to "twin-tag" the samples for downstream demultiplexing, (iv) a 5 N random sequence on the forward primer and a 4 N random sequence on the reverse primer (9 N total) as the UMI tags, and (v) parts of the Illumina universal adapter sequences to anneal to the 3′ ends of the forward and reverse primers for the second PCR. By splitting the 9 N UMI into 5 N + 4 N over the forward and reverse primers, we avoid primer dimers. The amplification primers include (i) an index sequence on the forward primer pair for Illumina library demultiplexing, and (ii) the full length of the Illumina adapter sequences. For further explanation of the design of the tagging primers (except for the UMI sequences), see Yang et al. (2021).

2.5 | PCR and the BEGUM pipeline

The first PCR amplifies *COI* and concatenates sample tags and UMIs and runs for only two cycles using a KAPA 2G Robust HS PCR Kit (Roche KAPA Biosystems). We used the mlCOIintF–FolDegenRev primer pair (Leray et al., 2013; Yu et al., 2012, p. 2012), which amplifies a 313-bp fragment of the *COI* barcode; and we followed the BEGUM protocol (Yang et al., 2021; Zepeda-Mendoza et al., 2016), which is a wet-laboratory and bioinformatic pipeline that combines multiple independent PCR replicates per sample, twin-tagging and false positive controls to remove tag-jumping and reduce erroneous sequences. Twin-tagging means using the same tag sequence on both the forward and reverse primers in a PCR, and we use this design because during library index PCR for Illumina sequencing, occasional incomplete extensions can create new primers that already contain the tag of one amplicon, resulting in chimeric sequences with tags from two different amplicons (Schnell et al., 2015). Tag jumps thus almost always result in nonmatching tag sequences, and these are identified and removed in the BEGUM pipeline. We performed three PCR replicates per sample, which means we used three different twin-tags to distinguish the three independent PCR replicates. BEGUM removes erroneous sequences by filtering out the reads that appear in a low number of PCR replicates (e.g., only one PCR) at a low number of copies per PCR (e.g., only two copies), because true sequences are more likely to appear in multiple PCRs with higher copy numbers per PCR. The 20-μl reaction mix included 4 μl Enhancer, 4 μl Buffer A, 0.4 μl dNTP (10 mM), 0.8 μl per primer (10 mM), 0.08 μl KAPA 2G HotStart DNA polymerase (Roche KAPA Biosystems), 5 μl template DNA and 5 μl water. PCR conditions were initial denaturation at 95°C for 3 min, followed by two cycles of denaturation at 95°C for 1 min, annealing at 50°C for 90 s,and extension at 72°C for 2 min. Then the products were purified with 14 μl of KAPA pure beads (Roche KAPA Biosystems) to remove the primers and PCR reagents and were eluted into 16 μl of water.

The second PCR amplifies the tagged templates for building the libraries that can be sequenced directly on the Illumina platform. The 50-μl reaction mix included 5 μl TAKARA buffer, 4 μl dNTP (10 mM), 1.2 μl per primer (10 mM), 0.25 μl TAKARA Taq DNA polymerase, 15 μl DNA product from the first PCR and 23.35 μl water. PCR conditions were initial denaturation at 95°C for 3 min, five cycles of denaturation at 95°C for 30 s, annealing at 59°C for 30 s (−1°C per cycle), extension at 72°C for 30 s, followed by 25 cycles of denaturation at 95°C for 30 s, annealing at 55°C for 30 s, extension at 72°C for 30 s; a final extension at 72°C for 5 min, and cool down to 4°C.

From all second PCR products, 2 μl was roughly quantified on a 2% agarose gel with IMAGE LAB 2.0 (Bio-Rad). For each set of PCRs with the same index, amplicons were mixed at equimolar ratios to make a pooled library. One PCR-negative control was set for each library. We sent our samples to Novogene for PE250 sequencing on an Illumina NovaSeq 6000, requiring 0.8 GB raw data from each PCR.

2.6 | Bioinformatic processing

ADAPTERREMOVAL 2.1.7 was used to remove any remaining adapters from the raw data (Schubert et al., 2016). SICKLE 1.33 was used to trim away low-quality bases at the 3′ ends. BFC version 181 was used to denoise the reads (Li, 2015). Read merging was performed using PANDASEQ 2.11 (Masella et al., 2012). BEGUM was used to demultiplex the reads by sample tag and to filter out erroneous reads (https://github.com/shyamsg/Begum, accessed September, 2021). We allowed 2-bp primer mismatches to the twin-tags while demultiplexing, and we filtered at a stringency of accepting only reads that appeared in at least two PCRs at a minimum copy number of four reads per PCR, with minimum length of 300 bp. This stringency minimized the false positive reads in the negative PCR control.

For mock-soup data, we need to compare the UMI and read numbers in each PCR set. However, BEGUM cannot recognize UMIs. Also because of our complicated primer structure, there is no software available for our data to count the UMIs per OTU in each PCR set. Thus, we wrote a custom bash script to process the mock-soup data from the PANDASEQ output files, which include all the UMIs, tags and primers. First, we used BEGUM-filtered sequences as a reference to filter reads for each PCR set and put the UMI information on read headers. Then we carried out reference-based OTU clustering for each PCR set with QIIME 1.9.1 (pick_otus.py -m uclust_ref -s 0.99) (Caporaso et al., 2010; Edgar, 2010), using the OTU representative sequences from barcoding Sanger sequencing as the reference, counted UMIs and reads for each OTU in each PCR set, and generated two OTU tables, separately with UMI and read numbers.

For the Malaise trap data, we directly used the BEGUM pipeline. After BEGUM filtering, VSEARCH 2.14.1 (−-uchime_denovo) (Rognes et al., 2016) was used to remove chimeras. SUMACLUST 1.0.2 was used to cluster the sequences of Malaise trap samples into 97% similarity OTUs. The python script tabulateSumaclust.py from the DAME toolkit was used to generate the OTU table. Finally, we applied the R package {LULU} 0.1.0 with default parameters to merge oversplit OTUs (Frøslev et al., 2017). The OTU table and OTU representative sequences were used for downstream analysis.

2.7 | Statistical analyses

All statistical analyses were carried out in R 4.1.0 (R Core Team, 2021), and we used the {lme4} 1.1–27 package (Bates et al., 2015) to fit linear mixed-effects models, using OTU, soup replicate and PCR replicates as random factors, to isolate the variance explained by the sole (fixed-effect) predictor of interest: OTU size. Model syntax is given in the legend of Figure 4. We used the {MuMIn} 1.43.17 package (CRAN.R-project.org/package=MuMIn, accessed January 2, 2022) to calculate the variance explained by fixed effects only (marginal R^2). To carry out spike-in correction, we first calculated a weighted mean from the added spike-ins (e.g., mean[Bombycidae + Elateridae/2 + Mordellidae/4]), rescaled the new mean spike-in so that the smallest value is equal to 1, and divided each row's OTU size and UMI number by the weighted, scaled spike-in.

3 | RESULTS

3.1 | Bioinformatic processing of the Malaise trap samples and the mock soups

Five libraries yielded a total of 283,319,770 paired-end reads, of which 247,285,097 were merged successfully in PANDASEQ. After BEGUM sorting and demultiplexing, which removed a large number of tag-jumped reads and some reads <300 bp length, we retained

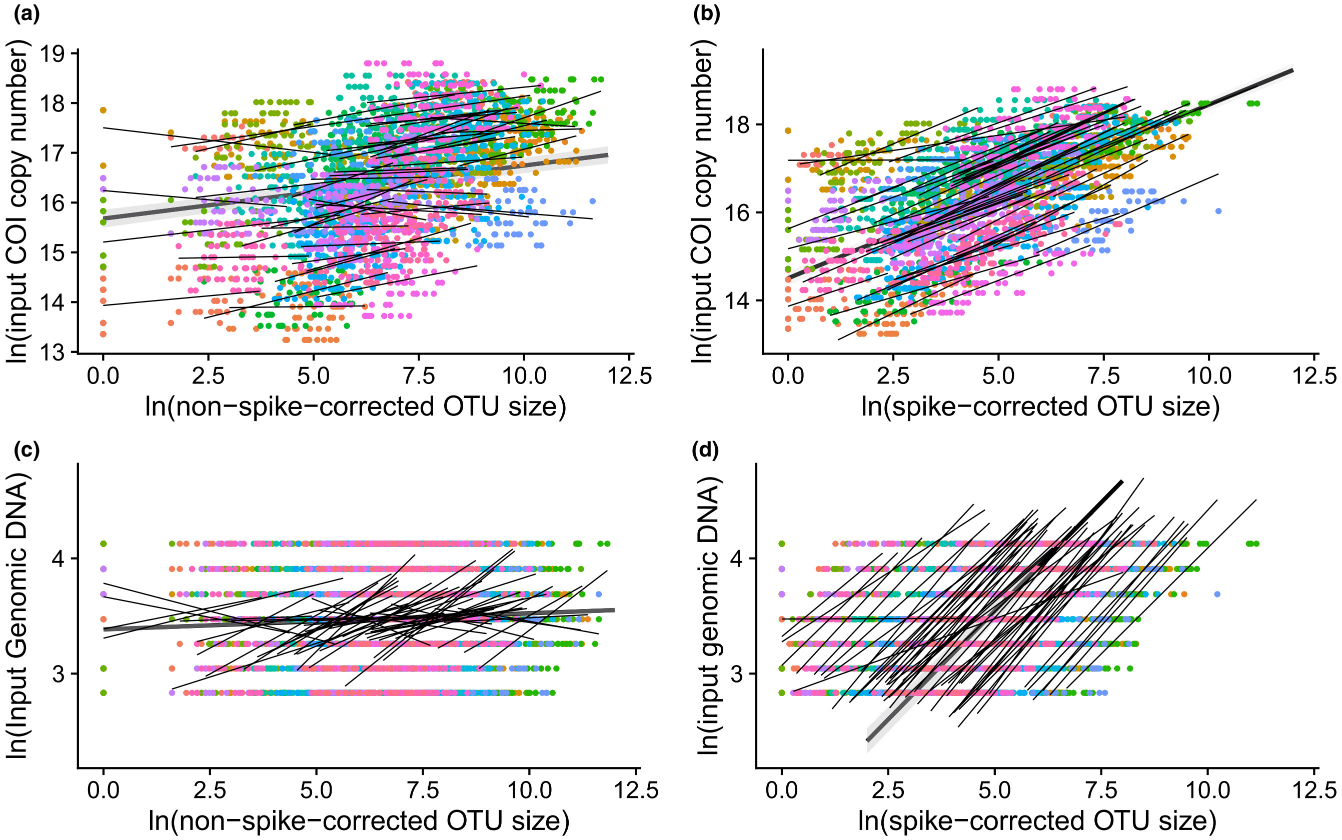


FIGURE 4 Recovery of within-species abundance change in *COI* copy number and in genomic DNA concentration in the mock-soup experiment. For visualization, all data points are shown (including all soup and PCR replicates), each thin line is fit to one of the OTUs across the seven serially diluted mock-soup samples, and the thick line represents the fitted model in which OTUs were treated as a random factor. (a) Nonspike-corrected OTU size (number of reads per OTU per soup) poorly predicts within-species variation in input *COI* copy number (linear mixed-effects model, marginal $R^2 = .04$, conditional $R^2 = .85$). (b) Spike-corrected OTU size successfully predicts within-species variation in input *COI* copy number (mixed-effects linear model, marginal $R^2 = .42$, conditional $R^2 = .96$), but species bias remains, as can be seen in the orders-of-magnitude variation in intercepts. (c) Nonspike-corrected read number poorly predicts within-species variation in input genomic DNA concentration (linear mixed-effects model, marginal $R^2 = .01$, conditional $R^2 = .01$). (d) Spike-corrected read number successfully predicts within-species variation in input genomic DNA concentration but more poorly for species represented by small OTUs (linear mixed-effects model, marginal $R^2 = .52$, conditional $R^2 = .95$) despite species bias (Figure 1). Model syntax: lme4::lmer(log.input_gDNA or log.inputCOI_copynumber ~ log.OTUsize + (log.OTUsize | OTUID) + (1 |soupRep/pcrRep)) (Bates et al., 2015). Marginal R^2 is variance explained by the fixed effect, and conditional R^2 is variance explained by the whole model

106,649,397 reads. After BEGUM's filtering of erroneous reads, we retained 76,289,802 reads, and after de novo chimera removal, we retained 73,818,971 reads. Sequences were clustered at 97% similarity into 1188 OTUs, and LULU combined the OTUs of the Malaise trap samples into 435 OTUs. After removing the spike-in OTUs, the seven Malaise trap samples contained a total of 432 OTUs. All 52 OTUs of the seven mock soups were recovered

3.2 | Mock soups, COI copy number

Without spike-in correction, OTU size (numbers of reads per OTU) predicts almost none of the within-species (dilution-gradient-caused) variation in *COI* copy number (R^2 = .04, all values marginal R^2), but with spike-in correction, OTU size predicts 42.0% of the variation (Figure 4a,b). As expected, UMI number by itself does not predict input *COI* copy number (R^2 = .05), but with spike-in correction, UMI number does predict *COI* copy number (R^2 = .42) (Figure S3a,b). Also as expected, spike-in correction does not achieve across-species quantification, as shown by the orders of magnitude variation in intercepts across the 52 OTUs. Note that this experiment pooled DNA extracts with equalized concentrations of genomic DNA mass per species, which suggests that PCR bias is the main source of species bias in this data set

3.3 | Mock soup within-species abundance in input genomic DNA mass

Of course, our goal is to estimate not *COI* copy number but specimen biomass. We thus tested how well OTU size and UMI numbers predicted genomic DNA concentration. Nonspike-corrected OTU size and UMI number both failed to predict input genomic DNA mass (R^2 <.02 for both, Figure 4c and Figure S3c), but spike-corrected OTU size and UMI number again both successfully predicted input genomic DNA mass (R^2 = .53 and .52, Figure 4d and Figure S3d).

3.4 | Malaise trap within-species abundance recovery

Recall that each of the seven selected Malaise trap samples was serially diluted by 0.7× to create six soups per sample. Nonspike-corrected OTU size did not predict within-species variation in input genomic-DNA mass (p = .33) (Figure 5a), but spike-corrected OTU size again did predict within-species variation in input genomic-DNA mass (R^2 = .53) (Figure 5b)

4 | DISCUSSION

We propose that there is a useful distinction to be made between *within*-species and *across*-species abundance information (Figures 1 and 2). Within-species abundance information can be enough to improve the inference of species interactions, the modelling of population dynamics and species distributions, the biomonitoring of environmental state and change, and the inference of false positives and negatives (Abrego et al., 2021; Carraro et al., 2020, 2021; Rojahn et al., 2021; Figure 2). We thus recommend that future quantitative eDNA studies should make clear which abundance measure is being estimated.

We experimentally show that spike-ins allow the recovery of within-species abundance change, by removing pipeline noise (Figures 4 and 5), even given the equimolar pooling step before library prep. In both experiments, we used a multispecies spike-in. The potential benefit of multiple species is the option to detect experimental error, which could be exposed by the spike-ins deviating strongly from their input ratios (Ji et al., 2020), but the cost is usage of sequence data on spike-in reads. Ushio et al. (2018) have also shown that spike-ins recover within-species abundance change, and they moreover showed that a spike-in can be used on trace fish eDNA in water samples. We note that Ushio et al.'s method is more complex than our method of counting the number of spike-in reads per sample, and so the optimal method for trace DNA remains an open research question.

In our first test, we serially diluted 52 OTUs into seven mock soups, and after spike-in correction (Figure 3), we were able to recover within-species abundance change in both input *COI* copy number and input genomic DNA (Figure 4), the latter of which should be more closely correlated with organism biomass. In our second test, we serially diluted each of the seven Malaise trap soups into six soups (Figure 3), and we were able to recover within-species abundance change in input genomic DNA (Figure 5).

Finally, our experimental protocol included UMIs, and we find that they can also recover within-species abundance change (Figure S3), but UMIs require a laborious two-step PCR protocol for no additional quantification benefit over the spike-in (Figure S3). On the other hand, UMIs have other advantages that could recommend them over a physical spike-in, such as not taking up sequencing data, which could make them more suitable for trace DNA sample types, contamination detection and error correction. Contaminant and erroneous sequences should be present at low abundances and thus capture few UMIs (Fields et al., 2021).

Additional alternatives to external spike-ins include a method introduced by Lundberg et al. (2021), who describe a two-step PCR method to use a single-copy host gene as a built-in spike-in. Also, in the Supporting Information code for Figure 4 (S4), we apply the model-based pipeline-noise estimator to the mock-soup data set and achieve an R^2 = 11.8% for prediction of *COI* copy number, which lies between the R^2 values achieved for the nonphysical-spike-corrected (R^2 = .04) and physical-spike-corrected values (R^2 = .42) (Figure 4b). We also achieve an R^2 = 21.3% for prediction of genomic DNA, again intermediate between the nonphysical-spike-corrected (R^2 = .0) and physical-spike-corrected values (R^2 = .53) (Figure 4d). In the Malaise trap data (Supporting Information S5), the model-based approach performed poorly at

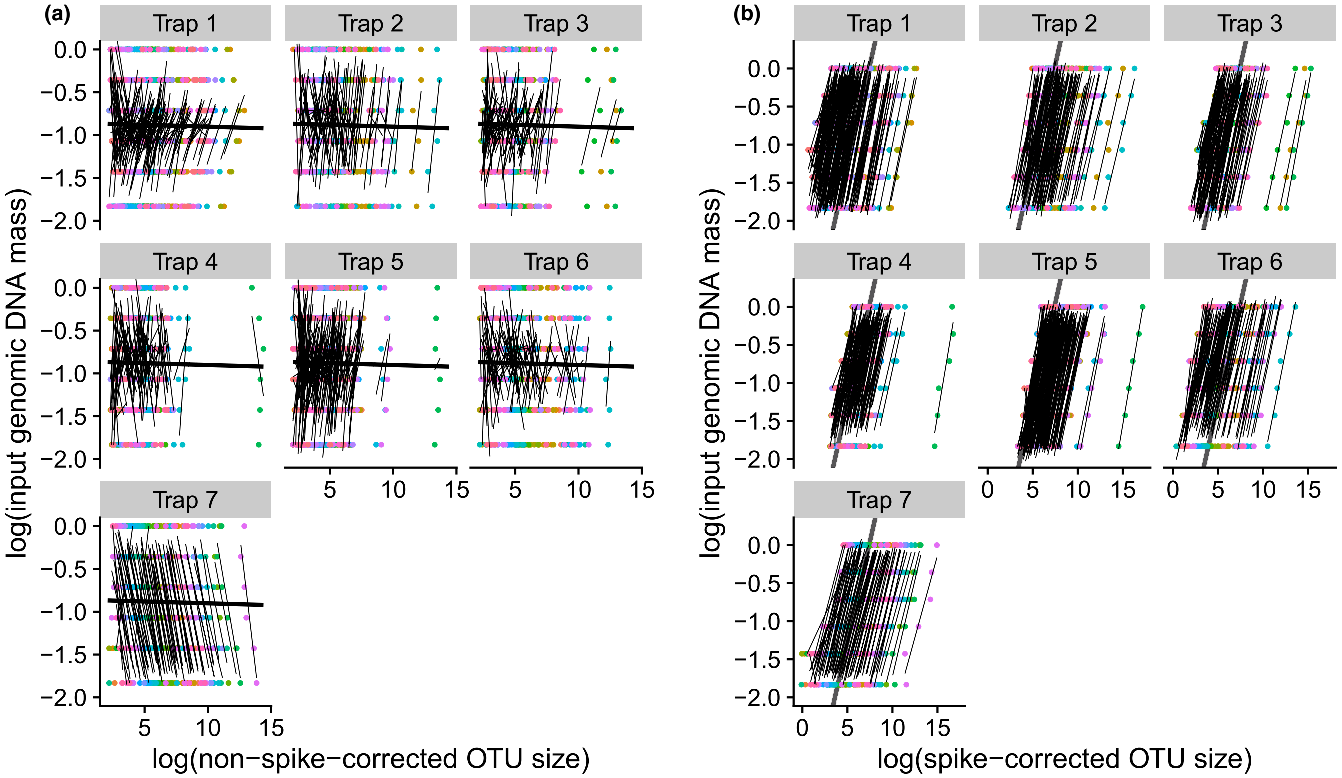


FIGURE 5 Prediction of within-species variation in genomic DNA concentration in the Malaise trap samples. For visualization, each thin line is fit to an OTU's serial dilution made from each of the seven Malaise trap samples, and the thick lines are the fitted model with sample and OTU as random factors. There are 176, 113, 111, 104, 196, 110 and 82 OTUs in samples 1–7, respectively. (a) Nonspike-corrected OTU size (read number per OTU and sample) does not predict within-species variation in genomic DNA concentration (marginal $R^2 = .0$, conditional $R^2 = .0$). (b) Spike-corrected OTU size successfully predicts within-species variation in genomic DNA concentration (marginal $R^2 = .53$, conditional $R^2 = .98$) despite species bias, represented by the different intercepts. A similar protocol was followed in Ji et al. (2020), where it was called “FSL correction.” Full model syntax: lme4::lmer(log.input_gDNA ~ log.OTUsize + [1 | sample/OTUID])

recovering genomic concentration. The issue was that samples had been pooled to equimolar concentration, which led to strong confounding of pipeline noise and differences in total abundance across samples. The model-based approach did however correctly infer that there were no compositional effects in this data set, consistent with a dilution gradient. This behaviour is as expected for the model-based method—it will recover relative not absolute DNA concentrations, and hence is a tool best used to study effects on compositional rather than total abundance.

Statistical analysis of DNA-based data sets will also need to exploit better within-species abundance information. The most straightforward method is to incorporate spike-in counts as an offset term in general linear models. For species distribution modelling, there is a need for software packages to utilize abundance data that ranges continuously over the interval [0,1], whereas to our knowledge, practitioners can effectively now only choose between presence/absence and absolute-abundance data.

We conclude with the acknowledgment that relative species abundance remains the more difficult abundance-estimation problem, given the many hidden sources of species bias along metabarcoding and metagenomic pipelines (McLaren et al., 2019), but promising solutions are now starting to be available for amplicon (Shelton, Gold, et al., 2022; Silverman et al., 2021; Williamson et al., 2021) and metagenomic data sets (Lang et al., 2019; Peel et al., 2019). Note that even if species biases can be corrected by using one of these techniques, it is still necessary to use a spike-in to correct for pipeline noise.

ACKNOWLEDGEMENTS

We thank Sarah Bourlat, Nathan Geraldi, Lucie Zinger and three anonymous reviewers for very helpful comments on the manuscript. The authors were supported by the Key Research Program of Frontier Sciences, Chinese Academy of Sciences (QYZDY-SSW-SMC024), the Strategic Priority Research Program, Chinese Academy of Sciences (XDA20050202), the State Key Laboratory of Genetic Resources and Evolution (GREKF19-01, GREKF20-01, GREKF21-01) at the Kunming Institute of Zoology, the sCom Working Group at iDiv.de, the University of East Anglia, and the University of Chinese Academy of Sciences.

CONFLICT OF INTEREST

DY is a cofounder of Nature Metrics, which provides commercial eDNA services.

DATA AVAILABILITY STATEMENT

Data and R scripts for Figures 4, 5 and S3, and the model-based estimator are available in Supplementary Information as RSTUDIO projects (S4 and S5). Other than the above, all sequence data (mock soup and Malaise trap), reference files, folder structure, output files and bioinformatic scripts (32.5 GB) are archived at https://doi.org/10.5061/dryad.2280gb5t8. The raw sequence data for the seven Malaise trap samples have also been submitted to NCBI's Short Read Archive with the rest of the Malaise trap data set (n_{tot} = 121) under BioProject number PRJNA869351 and will become available upon publication of the paper analysing that data set.

BENEFIT STATEMENT

There are no benefits to report. These samples were collected in Kunming, Yunnan, China (mock soup experiments) and HJ Andrews Experimental Forest, Oregon, USA (Malaise trap experiment).

ORCID

Douglas W. Yu https://orcid.org/0000-0001-8551-5609

REFERENCES

Abrego, N., Roslin, T., Huotari, T., Ji, Y., Schmidt, N. M., Wang, J., Yu, D. W., & Ovaskainen, O. (2021). Accounting for species interactions is necessary for predicting how arctic arthropod communities respond to climate change. *Ecography*, *44*(6), 885–896. https://doi.org/10.1111/ecog.05547

Amend, A. S., Seifert, K. A., & Bruns, T. D. (2010). Quantifying microbial communities with 454 pyrosequencing: Does read abundance count? *Molecular Ecology*, *19*(24), 5555–5565. https://doi.org/10.1111/j.1365-294X.2010.04898.x

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using **lme4**. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bell, K. L., Fowler, J., Burgess, K. S., Dobbs, E. K., Gruenewald, D., Lawley, B., Morozumi, C., & Brosi, B. J. (2017). Applying Pollen DNA Metabarcoding to the Study of Plant–Pollinator Interactions. *Applications in Plant Sciences*, *5*(6), 1600124. https://doi.org/10.3732/apps.1600124

Brys, R., Halfmaerten, D., Neyrinck, S., Mauvisseau, Q., Auwerx, J., Sweet, M., & Mergeay, J. (2021). Reliable eDNA detection and quantification of the European weather loach (*Misgurnus fossilis*). *Journal of Fish Biology*, *98*(2), 399–414. https://doi.org/10.1111/jfb.14315

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

Carraro, L., Mächler, E., Wüthrich, R., & Altermatt, F. (2020). Environmental DNA allows upscaling spatial patterns of biodiversity in freshwater ecosystems. *Nature Communications*, *11*(1), 3585. https://doi.org/10.1038/s41467-020-17337-8

Carraro, L., Stauffer, J. B., & Altermatt, F. (2021). How to design optimal eDNA sampling strategies for biomonitoring in river networks. *Environmental DNA*, *3*(1), 157–172. https://doi.org/10.1002/edn3.137

Clausen, D. S., & Willis, A. D. (2022). Modeling complex measurement error in microbiome experiments. *ArXiv:2204.12733 [Stat]*. http://arxiv.org/abs/2204.12733

Creedy, T. J., Norman, H., Tang, C. Q., Qing Chin, K., Andujar, C., Arribas, P., O'Connor, R. S., Carvell, C., Notton, D. G., & Vogler, A. P. (2020). A validated workflow for rapid taxonomic assignment and monitoring of a national fauna of bees (Apiformes) using high throughput DNA barcoding. *Molecular Ecology Resources*, *20*(1), 40–53. https://doi.org/10.1111/1755-0998.13056

Deagle, B. E., Clarke, L. J., Kitchener, J. A., Polanowski, A. M., & Davidson, A. T. (2018). Genetic monitoring of open ocean biodiversity: An evaluation of DNA metabarcoding for processing continuous plankton recorder samples. *Molecular Ecology Resources*, *18*(3), 391–406. https://doi.org/10.1111/1755-0998.12740

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

de Waard, J. R., Levesque-Beaudin, V., de Waard, S. L., Ivanova, N. V., McKeown, J. T. A., Miskie, R., Naik, S., Perez, K. H. J., Ratnasingham, S., Sobel, C. N., Sones, J. E., Steinke, C., Telfer, A. C., Young, A. D., Young, M. R., Zakharov, E. V., & Hebert, P. D. N. (2019). Expedited assessment of terrestrial arthropod diversity by coupling Malaise traps with DNA barcoding. *Genome*, *62*(3), 85–95. https://doi.org/10.1139/gen-2018-0093

Doi, H., Fukaya, K., Oka, S., Sato, K., Kondoh, M., & Miya, M. (2019). Evaluation of detection probabilities at the water-filtering and initial PCR steps in environmental DNA metabarcoding using a multispecies site occupancy model. *Scientific Reports*, *9*(1), 3581. https://doi.org/10.1038/s41598-019-40233-1

Dorazio, R. M., & Erickson, R. A. (2018). ednaoccupancy: An R package for multiscale occupancy modelling of environmental DNA data. *Molecular Ecology Resources*, *18*(2), 368–380. https://doi.org/10.1111/1755-0998.12735

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Elbrecht, V., & Leese, F. (2015). Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass—sequence relationships with an innovative metabarcoding protocol. *PLoS One*, *10*(7), e0130324. https://doi.org/10.1371/journal.pone.0130324

Erickson, R. A. (2019). Sampling designs for landscape-level eDNA monitoring programs. *Integrated Environmental Assessment and Management*, *12*, 760–771.

Ershova, E. A., Wangensteen, O. S., Descoteaux, R., Barth-Jensen, C., & Præbel, K. (2021). Metabarcoding as a quantitative tool for estimating biodiversity and relative biomass of marine zooplankton. *ICES Journal of Marine Science*, *78*(9), 3342–3355. https://doi.org/10.1093/icesjms/fsab171

Fields, B., Moeskjær, S., Friman, V., Andersen, S. U., & Young, J. P. W. (2021). MAUI-seq: Metabarcoding using amplicons with unique molecular identifiers to improve error correction. *Molecular Ecology Resources*, *21*(3), 703–720. https://doi.org/10.1111/1755-0998.13294

Folmer, O., Black, M., Hoeh, W., Lutz, R., & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, *3*(5), 294–299.

Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A. J. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications*, *8*(1), 1188. https://doi.org/10.1038/s41467-017-01312-x

Fukaya, K., Murakami, H., Yoon, S., Minami, K., Osada, Y., Yamamoto, S., Masuda, R., Kasai, A., Miyashita, K., Minamoto, T., & Kondoh, M. (2021). Estimating fish population abundance by integrating quantitative data on environmental DNA and hydrodynamic modelling.

Molecular Ecology, *30*(13), 3057–3067. https://doi.org/10.1111/mec.15530

Garrido-Sanz, L., Senar, M. À., & Piñol, J. (2021). Relative species abundance estimation in artificial mixtures of insects using mito-metagenomics and a correction factor for the mitochondrial DNA copy number. *Molecular Ecology Resources*, *22*(1), 153–167. https://doi.org/10.1111/1755-0998.13464

Griffin, J. E., Matechou, E., Buxton, A. S., Bormpoudakis, D., & Griffiths, R. A. (2020). Modelling environmental DNA data; Bayesian variable selection accounting for false positive and false negative errors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *69*(2), 377–392. https://doi.org/10.1111/rssc.12390

Gueuning, M., Ganser, D., Blaser, S., Albrecht, M., Knop, E., Praz, C., & Frey, J. E. (2019). Evaluating next-generation sequencing (NGS) methods for routine monitoring of wild bees: Metabarcoding, mitogenomics or NGS barcoding. *Molecular Ecology Resources*, *19*(4), 847–862. https://doi.org/10.1111/1755-0998.13013

Harrison, J. G., John Calder, W., Shuman, B., & Alex Buerkle, C. (2021). The quest for absolute abundance: The use of internal standards for DNA-based community ecology. *Molecular Ecology Resources*, *21*(1), 30–43. https://doi.org/10.1111/1755-0998.13247

Hebert, P. D. N., Braukmann, T. W. A., Prosser, S. W. J., Ratnasingham, S., de Waard, J. R., Ivanova, N. V., Janzen, D. H., Hallwachs, W., Naik, S., Sones, J. E., & Zakharov, E. V. (2018). A sequel to sanger: Amplicon sequencing that scales. *BMC Genomics*, *19*(1), 219. https://doi.org/10.1186/s12864-018-4611-3

Hindson, B. J., Ness, K. D., Masquelier, D. A., Belgrader, P., Heredia, N. J., Makarewicz, A. J., Bright, I. J., Lucero, M. Y., Hiddessen, A. L., Legler, T. C., Kitano, T. K., Hodel, M. R., Petersen, J. F., Wyatt, P. W., Steenblock, E. R., Shah, P. H., Bousse, L. J., Troup, C. B., Mellen, J. C., … Colston, B. W. (2011). High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Analytical Chemistry*, *83*(22), 8604–8610. https://doi.org/10.1021/ac202028g

Hoshino, T., & Inagaki, F. (2017). Application of stochastic labeling with random-sequence barcodes for simultaneous quantification and sequencing of environmental 16S rRNA genes. *PLoS One*, *12*(1), e0169431. https://doi.org/10.1371/journal.pone.0169431

Hoshino, T., Nakao, R., Doi, H., & Minamoto, T. (2021). Simultaneous absolute quantification and sequencing of fish environmental DNA in a mesocosm by quantitative sequencing technique. *Scientific Reports*, *11*(1), 4372. https://doi.org/10.1038/s41598-021-83318-6

Iwaszkiewicz-Eggebrecht, E., Granqvist, E., Buczek, M., Prus, M., Roslin, T., Tack, A. J. M., Andersson, A. F., Miraldo, A., Ronquist, F., & Łukasik, P. (2022). *Optimizing insect metabarcoding using replicated mock communities* [Preprint]. https://doi.org/10.1101/2022.06.20.496906

Ji, Y., Huotari, T., Roslin, T., Schmidt, N. M., Wang, J., Yu, D. W., & Ovaskainen, O. (2020). SPIKEPIPE: A metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes. *Molecular Ecology Resources*, *20*(1), 256–267. https://doi.org/10.1111/1755-0998.13057

Krehenwinkel, H., Wolf, M., Lim, J. Y., Rominger, A. J., Simison, W. B., & Gillespie, R. G. (2017). Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Scientific Reports*, *7*, 17668. https://doi.org/10.1038/s41598-017-17333-x

Lang, D., Tang, M., Hu, J., & Zhou, X. (2019). Genome-skimming provides accurate quantification for pollen mixtures. *Molecular Ecology Resources*, *19*(6), 1433–1446. https://doi.org/10.1111/1755-0998.13061

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., Boehm, J. T., & Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial *COI* region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, *10*(1), 34. https://doi.org/10.1186/1742-9994-10-34

Levi, T., Allen, J. M., Bell, D., Joyce, J., Russell, J. R., Tallmon, D. A., Vulstek, S. C., Yang, C., & Yu, D. W. (2019). Environmental DNA for the enumeration and management of Pacific salmon. *Molecular Ecology Resources*, *19*(3), 597–608. https://doi.org/10.1111/1755-0998.12987

Li, H. (2015). BFC: Correcting illumina sequencing errors. *Bioinformatics*, *31*(17), 2885–2887. https://doi.org/10.1093/bioinformatics/btv290

Lundberg, D. S., Ayutthaya, P. P. N., Strauß, A., Shirsekar, G., Lo, W.-S., Lahaye, T., & Weigel, D. (2021). Host-associated microbe PCR (hamPCR) enables convenient measurement of both microbial load and community composition. *eLife*, *10*, e66186. https://doi.org/10.7554/eLife.66186

Lundberg, D. S., Yourstone, S., Mieczkowski, P., Jones, C. D., & Dangl, J. L. (2013). Practical innovations for high-throughput amplicon sequencing. *Nature Methods*, *10*(10), 999–1002. https://doi.org/10.1038/nmeth.2634

Luo, M., Ji, Y., Warton, D., & Yu, D. W. (2022). *Dataset for "Extracting abundance information from DNA-based data."* DataDryad. https://doi.org/10.5061/dryad.2280gb5t8

Lyet, A., Pellissier, L., Valentini, A., Dejean, T., Hehmeyer, A., & Naidoo, R. (2021). EDNA sampled from stream networks correlates with camera trap detection rates of terrestrial mammals. *Scientific Reports*, *11*(1), 11362. https://doi.org/10.1038/s41598-021-90598-5

Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G., & Neufeld, J. D. (2012). PANDAseq: Paired-end assembler for illumina sequences. *BMC Bioinformatics*, *13*(1), 31. https://doi.org/10.1186/1471-2105-13-31

McLaren, M. R., Willis, A. D., & Callahan, B. J. (2019). Consistent and correctable bias in metagenomic sequencing experiments. *eLife*, *8*, e46923. https://doi.org/10.7554/eLife.46923

Meier, R., Wong, W., Srivathsan, A., & Foo, M. (2016). $1 DNA barcodes for reconstructing complex phenomes and finding rare species in specimen-rich samples. *Cladistics*, *32*(1), 100–110. https://doi.org/10.1111/cla.12115

Nielsen, M., Gilbert, M. T. P., Pape, T., & Bohmann, K. (2019). A simplified DNA extraction protocol for unsorted bulk arthropod samples that maintains exoskeletal integrity. *Environmental DNA*, *1*(2), 144–154. https://doi.org/10.1002/edn3.16

Pauvert, C., Buée, M., Laval, V., Edel-Hermann, V., Fauchery, L., Gautier, A., Lesur, I., Vallance, J., & Vacher, C. (2019). Bioinformatics matters: The accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline. *Fungal Ecology*, *41*, 23–33. https://doi.org/10.1016/j.funeco.2019.03.005

Peel, N., Dicks, L. V., Clark, M. D., Heavens, D., Percival-Alwyn, L., Cooper, C., Davies, R. G., Leggett, R. M., & Yu, D. W. (2019). Semi-quantitative characterisation of mixed pollen samples using MinION sequencing and Reverse Metagenomics (RevMet). *Methods in Ecology and Evolution*, *10*(10), 1690–1701. https://doi.org/10.1111/2041-210X.13265

Pierella Karlusich, J. J., Pelletier, E., Zinger, L., Lombard, F., Zingone, A., Colin, S., Gasol, J. M., Dorrell, R. G., Henry, N., Scalco, E., Acinas, S. G., Wincker, P., Vargas, C., & Bowler, C. (2022). A robust approach to estimate relative phytoplankton cell abundances from metagenomes. *Molecular Ecology Resources*. https://doi.org/10.1111/1755-0998.13592

Piñol, J., Mir, G., Gomez-Polo, P., & Agustí, N. (2015). Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular Ecology Resources*, *15*(4), 819–830. https://doi.org/10.1111/1755-0998.12355

Piñol, J., Senar, M. A., & Symondson, W. O. C. (2019). The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. *Molecular Ecology*, *28*(2), 407–419. https://doi.org/10.1111/mec.14776

Pochardt, M., Allen, J. M., Hart, T., Miller, S. D. L., Yu, D. W., & Levi, T. (2020). Environmental DNA facilitates accurate, inexpensive, and multiyear population estimates of millions of anadromous fish. *Molecular Ecology Resources*, *20*(2), 457–467. https://doi.org/10.1111/1755-0998.13123

R Core Team. (2021). *R: A language and environment for statistical computing* (4.0.4). R Foundation for Statistical Computing. https://www.R-project.org

Ratnasingham, S. (2019). mBRAVE: the multiplex barcode research and visualization environment. *Biodiversity Information Science and Standards*, *3*, e37986. https://doi.org/10.3897/biss.3.37986

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Rojahn, J., Pearce, L., Gleeson, D. M., Duncan, R. P., Gilligan, D. M., & Bylemans, J. (2021). The value of quantitative environmental DNA analyses for the management of invasive and endangered native fish. *Freshwater Biology*, *66*(8), 1619–1629. https://doi.org/10.1111/fwb.13779

Rourke, M. L., Fowler, A. M., Hughes, J. M., Broadhurst, M. K., DiBattista, J. D., Fielder, S., Wilkes Walburn, J., & Furlan, E. M. (2022). Environmental DNA (eDNA) as a tool for assessing fish biomass: A review of approaches and future considerations for resource surveys. *Environmental DNA*, *4*(1), 9–33. https://doi.org/10.1002/edn3.185

Schenk, J., Geisen, S., Kleinboelting, N., & Traunspurger, W. (2019). Metabarcoding data allow for reliable biomass estimates in the most abundant animals on earth. *Metabarcoding and Metagenomics*, *3*, e46704. https://doi.org/10.3897/mbmg.3.46704

Schneider, S., Taylor, G. W., Kremer, S. C., Burgess, P., McGroarty, J., Mitsui, K., Zhuang, A., de Waard, J. R., & Fryxell, J. M. (2022). Bulk arthropod abundance, biomass and diversity estimation using deep learning for computer vision. *Methods in Ecology and Evolution*, *13*(2), 346–357. https://doi.org/10.1111/2041-210X.13769

Schnell, I. B., Bohmann, K., & Gilbert, M. T. P. (2015). Tag jumps illuminated—Reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*, *15*(6), 1289–1303. https://doi.org/10.1111/1755-0998.12402

Schubert, M., Lindgreen, S., & Orlando, L. (2016). AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Research Notes*, *9*(1), 88. https://doi.org/10.1186/s13104-016-1900-2

Shelton, A. O., Gold, Z. J., Jensen, A. J., D'Agnese, E., Andruszkiewicz, E., & Kelly, P. (2022). Toward Quantitative Metabarcoding. *BioRXiv*. https://doi.org/10.1101/2022.04.26.489602

Shelton, A. O., O'Donnell, J. L., Samhouri, J. F., Lowell, N., Williams, G. D., & Kelly, R. P. (2016). A framework for inferring biological communities from environmental DNA. *Ecological Applications*, *26*(6), 1645–1659. https://doi.org/10.1890/15-1733.1

Shelton, A. O., Ramón-Laca, A., Wells, A., Clemons, J., Chu, D., Feist, B. E., Kelly, R. P., Parker-Stetter, S. L., Thomas, R., Nichols, K. M., & Park, L. (2022). Environmental DNA provides quantitative estimates of Pacific hake abundance and distribution in the open ocean. *Proceedings of the Royal Society B: Biological Sciences*, *289*(1971), 20212613. https://doi.org/10.1098/rspb.2021.2613

Silverman, J. D., Bloom, R. J., Jiang, S., Durand, H. K., Dallow, E., Mukherjee, S., & David, L. A. (2021). Measuring and mitigating PCR bias in microbiota data sets. *PLoS Computational Biology*, *17*(7), e1009113. https://doi.org/10.1371/journal.pcbi.1009113

Smets, W., Leff, J. W., Bradford, M. A., McCulley, R. L., Lebeer, S., & Fierer, N. (2016). A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing. *Soil Biology and Biochemistry*, *96*, 145–151. https://doi.org/10.1016/j.soilbio.2016.02.003

Srivathsan, A., Lee, L., Katoh, K., Hartop, E., Kutty, S. N., Wong, J., Yeo, D., & Meier, R. (2021). MinION barcodes: Biodiversity discovery and identification by everyone, for everyone. *BioRXiv*. doi:10.1101/2021.03.09.434692

Stauffer, S., Jucker, M., Keggin, T., Marques, V., Andrello, M., Bessudo, S., Cheutin, M.-C., Borrero-Pérez, G. H., Richards, E., Dejean, T., Hocdé, R., Juhel, J.-B., Ladino, F., Letessier, T. B., Loiseau, N., Maire, E., Mouillot, D., Mutis Martinezguerra, M., Manel, S., … Waldock, C. (2021). How many replicates to accurately estimate fish biodiversity using environmental DNA on coral reefs? *Ecology and Evolution*, *11*(21), 14630–14643. doi:10.1101/2021.05.26.445742

Steinke, D., Braukmann, T. W., Manerus, L., Woodhouse, A., & Elbrecht, V. (2021). Effects of Malaise trap spacing on species richness and composition of terrestrial arthropod bulk samples. *Metabarcoding and Metagenomics*, *5*, e59201. https://doi.org/10.3897/mbmg.5.59201

Tang, M., Hardman, C. J., Ji, Y., Meng, G., Liu, S., Tan, M., Yang, S., Moss, E. D., Wang, J., Yang, C., Bruce, C., Nevard, T., Potts, S. G., Zhou, X., & Yu, D. W. (2015). High-throughput monitoring of wild bee diversity and abundance via mitogenomics. *Methods in Ecology and Evolution*, *6*(9), 1034–1043. https://doi.org/10.1111/2041-210X.12416

Thalinger, B., Rieder, A., Teuffenbach, A., Pütz, Y., Schwerte, T., Wanzenböck, J., & Traugott, M. (2021). The effect of activity, energy use, and species identity on environmental DNA shedding of freshwater fish. *Frontiers in Ecology and Evolution*, *9*, 623718. https://doi.org/10.3389/fevo.2021.623718

Thomas, A. C., Deagle, B. E., Eveson, J. P., Harsch, C. H., & Trites, A. W. (2016). Quantitative DNA metabarcoding: Improved estimates of species proportional biomass using correction factors derived from control material. *Molecular Ecology Resources*, *16*(3), 714–726. https://doi.org/10.1111/1755-0998.12490

Tkacz, A., Hortala, M., & Poole, P. S. (2018). Absolute quantitation of microbiota abundance in environmental samples. *Microbiome*, *6*(1), 110. https://doi.org/10.1186/s40168-018-0491-7

Tsuji, S., Inui, R., Nakao, R., Miyazono, S., Saito, M., Kono, T., & Akamatsu, Y. (2022). Quantitative environmental DNA metabarcoding reflects quantitative capture data of fish community obtained by electrical shocker. *In BioRXiv*. https://doi.org/10.1101/2022.04.27.489619

Ushio, M., Murakami, H., Masuda, R., Sado, T., Miya, M., Sakurai, S., Yamanaka, H., Minamoto, T., & Kondoh, M. (2018). Quantitative monitoring of multispecies fish environmental DNA using high-throughput sequencing. *Metabarcoding and Metagenomics*, *2*, e23297. https://doi.org/10.3897/mbmg.2.23297

Verkuil, Y. I., Nicolaus, M., Ubels, R., Dietz, M. W., Samplonius, J. M., Galema, A., Kiekebos, K., de Knijff, P., & Both, C. (2022). DNA metabarcoding quantifies the relative biomass of arthropod taxa in songbird diets: Validation with camera-recorded diets. *Ecology and Evolution*, *12*(5), e8881. doi:10.1101/2020.11.26.399535

Wang, Y., Naumann, U., Wright, S. T., & Warton, D. I. (2012). Mvabund—An R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*, *3*(3), 471–474. https://doi.org/10.1111/j.2041-210X.2012.00190.x

Warton, D. (2022). *Eco-Stats—Data Analysis in Ecology* (1st ed.). Springer International Publishing. https://link.springer.com/book/9783030884420

Williamson, B. D., Hughes, J. P., & Willis, A. D. (2021). A multiview model for relative and absolute microbial abundances. *Biometrics*. doi:10.1111/biom.13503

Wührl, L., Pylatiuk, C., Giersch, M., Lapp, F., von Rintelen, T., Balke, M., Schmidt, S., Cerretti, P., & Meier, R. (2022). DiversityScanner: Robotic handling of small invertebrates with machine learning methods. *Molecular Ecology Resources*, *22*(4), 1626–1638. doi:10.1101/2021.05.17.444523

Yang, C., Bohmann, K., Wang, X., Cai, W., Wales, N., Ding, Z., Gopalakrishnan, S., & Yu, D. W. (2021). Biodiversity Soup II: A bulk-sample metabarcoding pipeline emphasizing error reduction. *Methods in Ecology and Evolution*, *12*(7), 1252–1264. https://doi.org/10.1111/2041-210X.13602

Yates, M. C., Cristescu, M. E., & Derry, A. M. (2021). Integrating physiology and environmental dynamics to operationalize environmental

DNA (eDNA) as a means to monitor freshwater macro-organism abundance. *Molecular Ecology*, *30*(24), 6531–6550. https://doi.org/10.1111/mec.16202

Yates, M. C., Glaser, D. M., Post, J. R., Cristescu, M. E., Fraser, D. J., & Derry, A. M. (2021). The relationship between eDNA particle concentration and organism abundance in nature is strengthened by allometric scaling. *Molecular Ecology*, *30*(13), 3068–3082. https://doi.org/10.1111/mec.15543

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., & Ding, Z. (2012). Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, *3*(4), 613–623. https://doi.org/10.1111/j.2041-210X.2012.00198.x

Zepeda-Mendoza, M. L., Bohmann, K., Carmona Baez, A., & Gilbert, M. T. P. (2016). DAMe: A toolkit for the initial processing of datasets with PCR replicates of double-tagged amplicons for DNA metabarcoding analyses. *BMC Research Notes*, *9*(1), 255. https://doi.org/10.1186/s13104-016-2064-9

Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X., Zhou, L., Tang, M., Fu, R., Li, J., & Huang, Q. (2013). Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *GigaScience*, *2*(1), 4. https://doi.org/10.1186/2047-217X-2-4

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: Luo, M., Ji, Y., Warton, D., & Yu, D. W. (2022). Extracting abundance information from DNA-based data. *Molecular Ecology Resources*, *00*, 1–16. https://doi.org/10.1111/1755-0998.13703