# 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 650223
[2] Kunming College of Life Sciences, University of Chinese Academy of Sciences, Kunming, China
[3] School of Mathematics and Statistics, UNSW Sydney, NSW 2052, Australia
[4] Evolution and Ecology Research Centre, UNSW Sydney, NSW 2052, Australia
[5] Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming Yunnan, China 650223
[5] School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk, UK NR47TJ

## 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, capture bias, capture noise, species pipeline biases, and pipeline noise all combine to inject error into DNA-based datasets. We focus on methods for correcting species pipeline biases and pipeline noise, as the first two errors are addressed extensively in the ecological literature. To extract abundance information, it is useful to distinguish two concepts. (1) *Across*-species quantification describes relative species abundances within one sample. (2) *Within*-species quantification describes how the abundance of each individual species varies from sample to sample, as in a time series, an environmental gradient, or experimental treatments. Firstly, we review methods to remove species pipeline biases and pipeline noise. Secondly, we demonstrate experimentally (with detailed protocol) how to use a 'DNA spike-in' to remove pipeline noise and recover within-species abundance information. We also introduce a model-based estimator that can be employed on datasets without a physical spike-in to approximately estimate and correct for pipeline noise.

## 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 (Thomas et al. 2016; Deagle et al. 2019; Peel et al. 2019; Carraro et al. 2020, 2021; Abrego et al. 2021; Rojahn et al. 2021). *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 (see Shelton et al. 2016; Griffin et al. 2020 for more formal treatments), which we will call here: (1) *species capture biases*, (2) *capture noise*, (3) *species pipeline biases*, and (4) *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 eDNA shedding rates (Thalinger et al. 2021; Yates et al. 2021*b*).
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 quantitative PCR of aquatic environmental DNA, 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 et al. 2021*a*).
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; Yu et al. 2012; Elbrecht and Leese 2015; Piñol et al. 2015, 2019; Tang et al. 2015; Bell et al. 2017; Krehenwinkel et al. 2017; McLaren et al. 2019; Pauvert et al. 2019; Garrido-Sanz et al. 2021; Yang et al. 2021). 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 (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-lab pipelines. In particular, it is common practice *to deliberately equalise 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 (e.g. Dorazio and Erickson 2018; Doi et al.

2019; Erickson 2019; Griffin et al. 2020; Lyet et al. 2021; Stauffer et al. 2021) and the review by Yates et al. (2021*a*). 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 datasets. We start with a notionally true sample X 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 versus sample 1 (green cells in Figure 1 A), in the observed table, OTU1 is only *two* times as abundant in sample 4 (green cells in Figure 1 B). *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 1 A), in the observed OTU table, OTU2 is two times as abundant as OTU1 in sample 3 (orange cells in Figure 1 B). 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):

1. *Within*-species quantification: E.g. “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 2 A1, D).
2. *Across*-species quantification: E.g. “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:

$$\log(\mu_{ij}) = a_i + a_j + \boldsymbol{x}_i'\boldsymbol{b} + \boldsymbol{x}_i'\boldsymbol{b}_j \qquad \text{Eq. 1}$$

where $\mu_{ij}$ 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 environmental variables $\boldsymbol{x}_i$ have an effect on total abundance (via $\boldsymbol{b}$) as well as having a compositional effect, *i.e.* 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, (Carraro et al. 2020, 2021; Abrego et al. 2021; Rojahn et al. 2021), and the inference of false positives and negatives (Figure 3).

*Mini-review of methods to extract abundance information*

*Multiplexed individual barcoding*. – The most straightforward approach is to DNA-barcode individual organisms and count them up, which achieves both within- and across-species quantification. This approach assumes bulk samples, with a target taxon (e.g. Insecta) that has body sizes suitable for mass individual sorting. DNA extraction, uniquely tagged PCR, and subsequent pooling and sequencing on an Illumina (Meier et al. 2016; Ratnasingham 2019; Creedy et al. 2020), PacBio (Hebert et al. 2018), or MinION (Srivathsan et al. 2021) now costs much less than $1 per individual. Wührl et al. (2021) further increase throughput with a robotic pipettor and camera that visually IDs small insects to higher taxonomic rank and sorts them into 96-well plates for individual DNA barcoding. This method achieves both within- and across-species quantification but is difficult to apply to very large numbers of samples and cannot be applied to trace DNA or microbial taxa.

*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 dataset 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.

*Design better 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. Similarly, Verkuil et al. (2020) have reported that modified COI primers can recover the relative biomasses of insect orders from Pied Flycatcher faeces. This method achieves across-species quantification but not within-species quantification.

*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 is qPCR (Brys et al. 2021) and allows the detection of single copies of target DNA and absolute quantification through the partitioning of the PCR reaction into 20,000 droplets and subsequent fluorescent detection of droplets that contain the target DNA (Hindson et al. 2011). This paper does not review q/ddPCR except to note that single-species q/ddPCR applied to aquatic trace DNA can estimate absolute abundances, provided that one corrects for water flux and/or surface-area to mass ratio (Levi et al. 2019; Pochardt et al. 2020; Fukaya et al. 2021; Yates et al. 2021*b*, 2021*a*). If applied to multiple species and if statistical models that relate DNA copy number to abundance can be fitted (Levi et al. 2019; Pochardt et al. 2020; Fukaya et al. 2021), then across-species quantification can also be achieved, albeit with non-trivial amounts of error.

*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, also known as an internal standard (ISD, Harrison et al. 2021). The 'spike-in' sequence of course should not match any species that could be in the samples and is flanked by primer binding sequences that match the primers used (Smets et al. 2016; Deagle et al. 2018; Tkacz et al. 2018; Harrison et al. 2021). Because each sample receives the same amount of spike-in, all samples should in theory return the same number of spike-in reads, but pipeline noise ensures that this never happens. Some samples return more spike-in reads, meaning that their OTUs are too large (too many reads), and some samples return fewer spike-in reads, meaning that their OTUs are too small. The correction step is simple: divide each sample's OTU sizes by the number of spike-in reads in that sample (Ji et al. 2020; Abrego et al. 2021) or include the spike-in reads as an offset term in a multivariate statistical model (Wang et al. 2012). This latter approach can be understood as estimating $a_i$ in Eq. 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).

Samples with a high number of spike-in reads must have had a low amount of template DNA, and vice versa, so dividing shrinks OTU sizes accordingly. For instance, following the pioneering work of Zhou et al. (2013), Ji et al. (2020) mapped WGS datasets 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 achieved by employing a spike-in, replicates to correct for sequence-run noise (an additional source of pipeline noise), and a percent-coverage filter to remove false positives. However, the regression lines relating 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 non-spike-corrected ('compositional') datasets.

*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 Eq. 1 to data. However, fitting this full model with row effects can be computationally intensive, especially for large datasets, so a simple alternative is to approximate $a_i$ using a one-step estimator (Warton 2022):

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

where $\hat{\mu}_{ij}^{(0)}$ is the 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.

As an intuitive explanation, note that a sample might end up generating many sequence reads by chance (pipeline noise) and/or because predictors $\boldsymbol{x}_i$ suggest that many of the OTUs are expected to be abundant in that sample (ecology). $\log \sum_{j=1}^{p} y_{ij}$ 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 $\mu_{ij}$ since, by design, the same amount of each spike-in species is included in every sample. 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.

*Unique Molecular Identifiers (UMIs).* – A UMI is a series of ~7-12 random bases ('NNNNNNN') added to the forward primer as an ultra-high diversity tag (Hoshino and 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 and Inagaki 2017; Hoshino et al. 2021). 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.

*Multiview modelling*. – Williamson et al. (2021) have proposed a method for achieving across-species quantification. Originally, McLaren et al. (2019) pointed out that although it is not possible to estimate *a priori* the species biases in a given amplicon or metagenomic dataset (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*. Thus, if one first estimates the absolute abundances of a subset of species in the dataset (via multiple, species-specific q/ddPCR assays), it is possible to infer the absolute abundances of the rest of the species (by inferring from data 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-lab steps have been carried out, multiview modelling does not remove pipeline noise. Thus, a spike-in is still needed to achieve within-species quantification.

*Forward and reverse metagenomics*. – Lang et al. (2019) showed that WGS (whole-genome sequencing) datasets 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 ≥1.0X 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 percent 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 datasets. 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. However, many ecological questions can be answered by achieving only within-species quantification (Figure 3), and in the second half of this paper, we provide a detailed protocol and experimental validation for spike-ins to achieve within-species quantification.

A. True OTU table

| | OTU1 | OTU2 | OTU3 | OTU4 | spikeOTU | True rowSum |
|---|---|---|---|---|---|---|
| Sample 1 | 10 | 20 | 0 | 0 | 20 | **30** |
| Sample 2 | 0 | 100 | 20 | 50 | 20 | **170** |
| Sample 3 | 40 | 40 | 5 | 50 | 20 | **135** |
| Sample 4 | 60 | 0 | 30 | 100 | 20 | **190** |

scales::rescale() each OTU (column) to interval [0,1]

A1. True OTU table, each OTU rescaled to [0,1]

| | OTU1 | OTU2 | OTU3 | OTU4 |
|---|---|---|---|---|
| Sample 1 | 0.2 | 0.2 | 0.0 | 0.0 |
| Sample 2 | 0.0 | 1.0 | 0.7 | 0.5 |
| Sample 3 | 0.7 | 0.4 | 0.2 | 0.5 |
| Sample 4 | 1.0 | 0.0 | 1.0 | 1.0 |

Metabarcoding or metagenomic pipeline

Within-OTU (within-column) frequencies recovered by spike-correction

B. Observed OTU table with pipeline noise and species pipeline biases

| | OTU1 | OTU2 | OTU3 | OTU4 | spikeOTU | Pipeline noise | Observed rowSum |
|---|---|---|---|---|---|---|---|
| Sample 1 | 60 | 60 | 0 | 0 | 120 | 3 | **120** |
| Sample 2 | 0 | 300 | 240 | 900 | 120 | 3 | **1440** |
| Sample 3 | 160 | 80 | 40 | 600 | 80 | 2 | **880** |
| Sample 4 | 120 | 0 | 120 | 600 | 40 | 1 | **840** |
| Species biases | 2 | 1 | 4 | 6 | 2 | | |

DNA spike-in correction

divide each row by its spikeOTU

C. Spike-corrected OTU table

| | OTU1 | OTU2 | OTU3 | OTU4 |
|---|---|---|---|---|
| Sample 1 | 0.5 | 0.5 | 0.0 | 0.0 |
| Sample 2 | 0.0 | 2.5 | 2.0 | 7.5 |
| Sample 3 | 2.0 | 1.0 | 0.5 | 7.5 |
| Sample 4 | 3.0 | 0.0 | 3.0 | 15.0 |
| Species biases | 2 | 1 | 4 | 6 |

scales::rescale()

D. Spike-corrected OTU table, each OTU rescaled to [0,1]

| | OTU1 | OTU2 | OTU3 | OTU4 |
|---|---|---|---|---|
| Sample 1 | 0.2 | 0.2 | 0.0 | 0.0 |
| Sample 2 | 0.0 | 1.0 | 0.7 | 0.5 |
| Sample 3 | 0.7 | 0.4 | 0.2 | 0.5 |
| Sample 4 | 1.0 | 0.0 | 1.0 | 1.0 |
| Species biases | 2 | 1 | 4 | 6 |

mvabund::manyglm(OTUtable ~ 1 + offset(log(spikeOTU)))

otu[otu>0] <- 1

otu[otu>0] <- 1

E. Presence-Absence OTU table

| | OTU1 | OTU2 | OTU3 | OTU4 |
|---|---|---|---|---|
| Sample 1 | 1 | 1 | 0 | 0 |
| Sample 2 | 0 | 1 | 1 | 1 |
| Sample 3 | 1 | 1 | 1 | 1 |
| Sample 4 | 1 | 0 | 1 | 1 |

F. Observed OTU table, divided by Observed rowSum

| | OTU1 | OTU2 | OTU3 | OTU4 |
|---|---|---|---|---|
| Sample 1 | 0.5 | 0.5 | 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 | 0.4 | 1.0 | 0.9 |
| Sample 3 | 0.4 | 0.2 | 0.3 | 1.0 |
| Sample 4 | 0.3 | 0.0 | 0.9 | 1.0 |
| Species biases | 2 | 1 | 4 | 6 |

2

**Figure 1**. **Pipeline noise versus species bias in OTU tables**. **A**. The true OTU table, with cell numbers representing the true amounts of DNA or biomass for each OTU (column) in each sample (row). The true rowSum column thus shows the total DNA or biomass per sample, and the spikeOTU column shows that the same amount of standard DNA spike-in has been added to each sample. **A1**. The true OTU table after rescaling each OTU column to the interval [0,1] ('quasiprobabilities'). **B**. The observed OTU table after amplicon or shotgun 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. Pipeline noise causes the original 10:60 ratio of Sample 1 and Sample 4 in OTU1 (green cells) to appear as 30:60, and species bias causes the original 40:40 ratio of OTU1 and OTU2 (orange cells) to appear as 80:160. **C**. The observed OTU table after dividing each row by the observed spike-in reads, to remove pipeline noise. Note that species biases remain uncorrected. **D**. Table C after rescaling each OTU 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. In statistical modelling, observed spike-in values are an index of sampling (sequencing) effort and can be included as offset values. **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 potentially loses ecological information (**Figure 3**). **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 4 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 **Figure 2**. 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 16 Dec 2021). (**358 words**)

**Figure 2**. **Line graphs of Figure 1, transposed so that each line is a column from the corresponding OTU table**. With spike-in correction, *within*-species abundance information is recovered (follow each OTU's line from sample to sample), but *across*-species abundance information is not (compare species abundances within each sample). Blue borders indicate that within-species abundance information has been preserved by using spike-in correction. Red borders indicate the loss of within-species abundance information. (**69 words**)

A. Absolute abundance data for one species

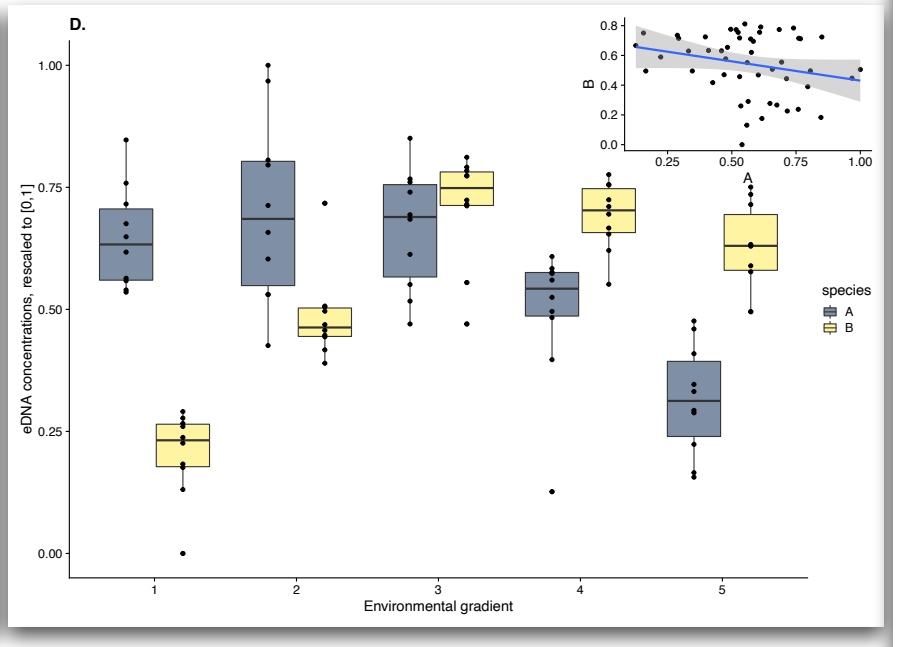
| | | |
|---|---|---|
| | | 3 |
| 3 | 30 | 3 |
| 3 | 21 | |
| 3 | 3 | 3 |

B. Within-species abundance data, [0,1]

| | | |
|---|---|---|
| | | 0.1 |
| 0.1 | 1 | 0.1 |
| 0.1 | 0.7 | |
| 0.1 | 0.1 | 0.1 |

C. Presence-absence data for one species

| | | |
|---|---|---|
| | | 1 |
| 1 | 1 | 1 |
| 1 | 1 | |
| 1 | 1 | 1 |



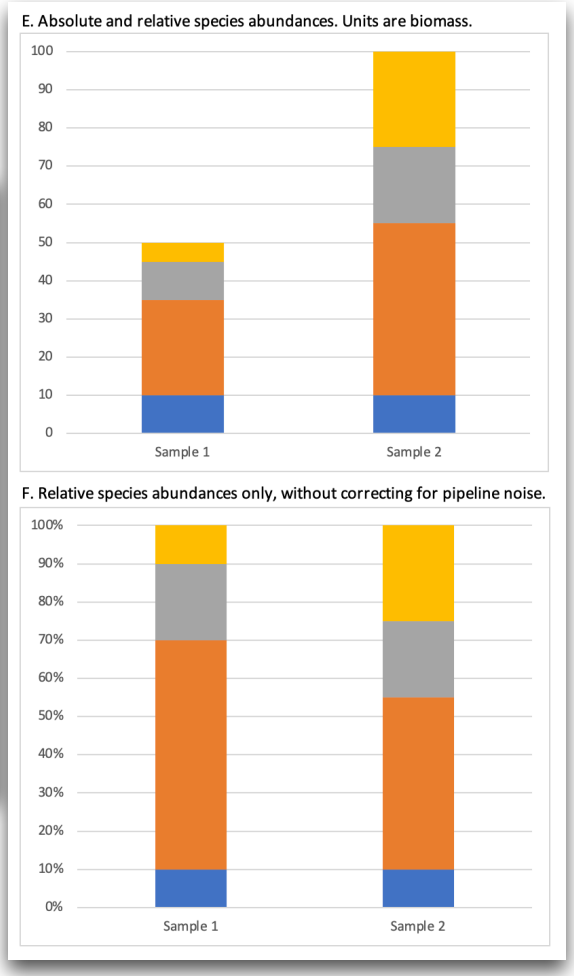




**Figure 3**. **The usefulness of achieving only within-species or only across-species quantification**. **A.** Imagine that a species is found in many sites but that only two sites are optimal (green) with high abundances, while most sites are suboptimal (grey) with low abundances maintained by immigration. **B**. Let the existence of unquantified species biases make it impossible to recover absolute abundances from DNA-based data, but as long as pipeline noise is removed with a spike-in, it is still possible to track how the species' abundance varies across sites (rescaled to [0,1]), which generates the hypothesis that the green sites are 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, species A decreases, and species B increases in eDNA concentration. Even without correcting for species biases, removal of pipeline noise (and capture noise) allows the detection of a negative correlation in abundances (inset). Example adapted from Rojahn et al. (2021) who combined a similar result with external information to infer the competitive exclusion of a native fish species by an invasive species. Note that due to species biases, it is not possible to infer from these within-species abundance gradients which species is more abundant than the other (nor is it correct to infer that the two species are of similar abundance). **E**. Two samples with different absolute and relative species abundances. **F**. If only across-species quantification is achieved (e.g. via forward or reverse metagenomics), it is only valid to compare across species *within* a sample. Thus, it is valid to conclude that the dark orange species has the highest relative abundance in both samples 1 and 2, but it is not valid to conclude that the dark orange species has a greater biomass in Sample 1 than in Sample 2, as can be seen by inspection of the true absolute abundances in **E**. However, adding a spike-in and also achieving within-species quantification would recover total biomasses and make it possible to compare how each species' abundances vary across samples (E). (**347 words**)

## Methods

*2.1 Mock soup construction.* – 286 arthropods were collected in Kunming, China (25°8'23" N, 102°44'17" E). DNA was extracted from each individual using the DNeasy Blood & Tissue Kit (Qiagen GmbH, Germany). Genomic DNA concentration of each individual was quantified from three replicates using PicoGreen fluorescent dye. 658-bp COI barcoding sequences were PCR'd with Folmer primers (LCO1490 and HCO2198) and Sanger-sequenced. After the 658-bp COI sequences were trimmed to 313 bp based on our metabarcoding primers (see below), 286 arthropods were clustered to 168 OTUs at 97% similarity. We selected 52 individuals with genomic DNA > 20 ng/µl, representing 52 OTUs. Each mock soup was constructed with equal masses of purified genomic DNA from the 52 OTUs. Input genomic DNA masses of each OTU from soup 1 to soup 7 decreased by 0.8X to create seven mock soups of differing absolute abundances. Each of the seven soups was made in triplicate ($n_{tot}$ = 21) (Figure 4 A).

*2.2 Preparation of Malaise-trap samples.* – 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, United States in July 2018. Traps were left for 7 non-rainy 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 an hour and then transferred to 50 ml tubes or 5 ml tubes according to sample volume. The samples were then weighed. DNA was non-destructively extracted by soaking the samples in lysis buffer, using the protocol from Ji et al. (2020) and Nielsen et al. (2019). For this study, we selected seven samples spread over the study area. After completion of lysis, we serially diluted the 7 samples by using 0.7X lysis buffer volume (500 µl, 350 µl, 245 µl, 171.5 µl, 120 µl and 84 µl) to create six soups per sample ($n_{tot}$ = 42). We used QIAquick PCR purification kit (Qiagen GmbH, Germany) following the manufacturer instructions to purify lysis buffer on one spin column per soup (Figure 4 B).

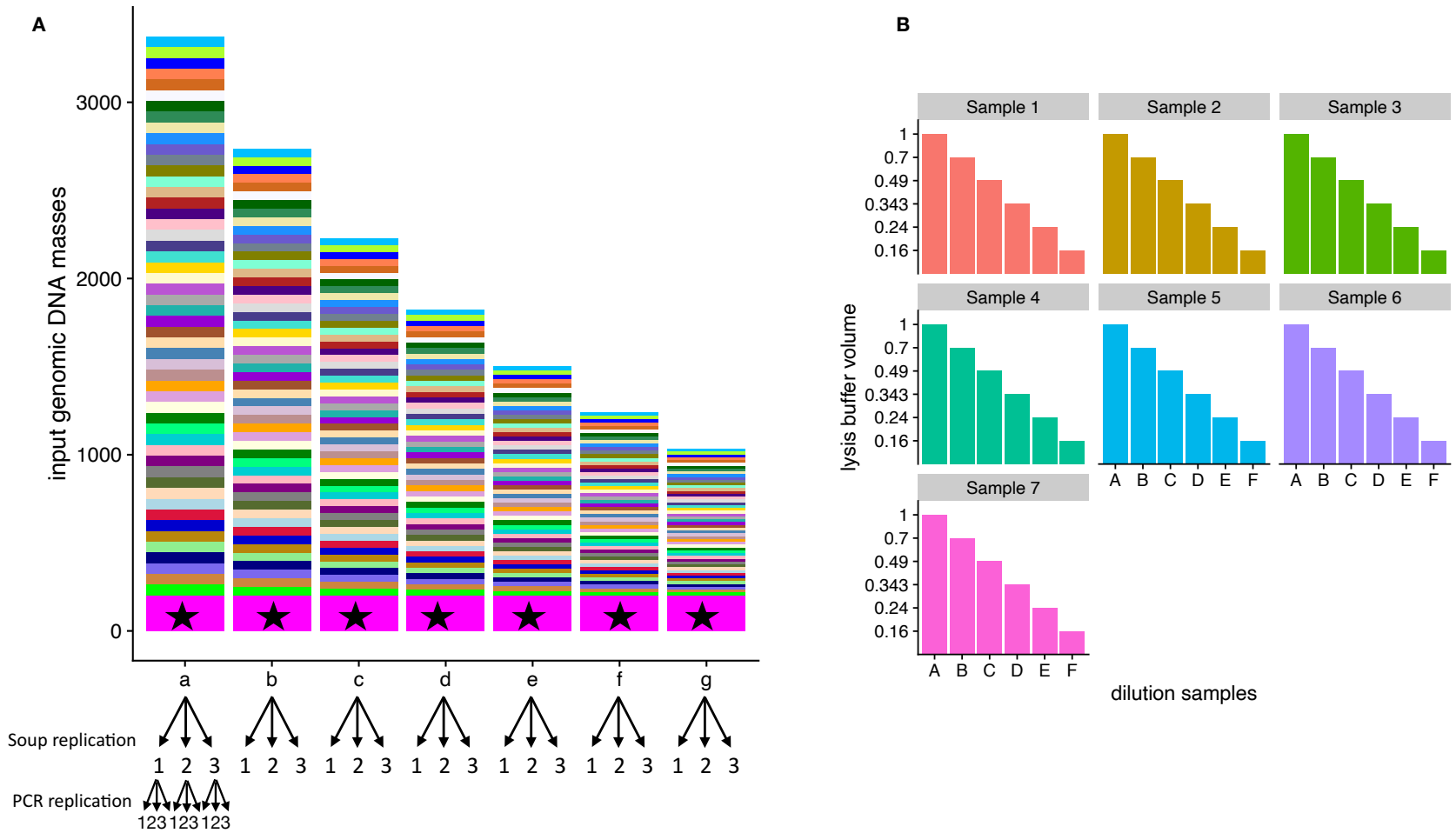



**Figure 4**. **Preparation of mock and Malaise-trap soups**. A. Mock soups. Each mock soup was constructed with equal masses of purified DNA from 52 OTUs. From soup "a" to soup "g", the input genomic masses of each of the 52 OTUs were 61, 48.8, 39, 31.2, 25, 20 and 16 ng. The same mass of spike-in DNA was then added to each soup (star). 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 soups. 500 µl, 350 µl, 245 µl, 171.5 µl, 120 µl and 84 µl of lysis buffer were used to create 6 soups (A-F) per sample. The spike-in DNA was added before extraction. Each soup ($n_{tot}$ = 42) was PCR'd in triplicate. (**139 words**)

334 *2.3 Adding spike-in DNA*

335 *2.3.1 Spike-in DNA.* – For our spike-ins, we used three insect species from China
336 (Lepidoptera:Bombycidae, Coleoptera:Elateridae, Coleoptera:Mordellidae), none of
337 which is expected to appear in the Oregon Malaise-trap samples. An alternative is to
338 use one or more synthetic, random DNA sequences (Tkacz et al. 2018). Each of our
339 three spike-ins is represented by a 658-bp COI fragment (Table 1) with primer binding
340 sites that match the Folmer primers HCO2198 and LCO1490. For long-term storage,
341 we inserted the COI fragments as plasmids into monoclonal bacteria. Plasmids were
342 extracted using TIANprep Mini Plasmid Kit (Beijing, China) following manufacturer's
343 instructions.

344 *2.3.2 Adding spike-in to the mock soups.* – Adding too much spike-in wastes
345 sequencing data, while adding too little risks loss of abundance information in at least
346 some samples when the number of spike-in reads is too low to use as a reliable
347 correction factor. Thus, we quantified the COI copy numbers of the mock soups and
348 the spike-in DNA by qPCR and chose a volume so that spike-in reads should make up
349 1% of the total number of COI copies in the lowest-concentration mock soups,
350 balancing efficiency with reliability. We used all three spike-in species here and mixed
351 them (Bombycidae:Elateridae:Mordellidae) in a ratio of 1:2:4, which was added directly
352 to the mock soups' DNA since they were already purified.

353 2.3.3 *Adding spike-in to the Malaise-trap samples.* – From the 244 Malaise-trap
354 samples, we first extracted 17 Malaise-trap samples without adding spike-ins, and then
355 we used qPCR to quantify the mean COI concentrations of these 17 samples in order
356 to decide how much spike-in to add. Before adding the spike-ins, we discovered that
357 the Bombycid DNA spike-in had degraded, and so we used only two spike-in species
358 for the Malaise trap samples, at a ratio of 1:9 (Mordellidae:Elateridae). For the 7
359 samples used in this study, lysis buffer (500 µl, 350 µl, 245 µl, 171.5 µl, 120 µl, 84 µl)
360 from each sample was transferred into clean 1.5 ml tubes, and the spike-in DNA was
361 added. We then purified the DNA with the Qiagen QIAquick PCR purification kit,
362 following the manufacturer instructions. DNA was eluted with 200 µl of elution buffer. In
363 this way, the spike-in DNA was co-purified, co-amplified, and co-sequenced along with
364 the sample DNA (Figure 5). We also recorded the total lysis buffer volume of each
365 sample, for downstream correction.

366 *2.4 Primer design.* – For this study, we simultaneously tested two methods for
367 extracting abundance information: spike-ins and UMIs (Unique Molecular Identifiers).
368 UMI tagging requires a two-step PCR procedure (Lundberg et al. 2013; Hoshino and
369 Inagaki 2017), first using tagging primers and then using amplification primers (Figure

6). The tagging primers include (1) the Leray-FolDegenRev primer pair to amplify the 313-bp COI amplicon of interest, (2) a 1- or 2-nucleotide heterogeneity spacer on both the forward and reverse primers to increase sequence entropy for the Illumina sequencer, (3) the same 6-nucleotide sequence on both the forward and reverse primers to ‘twin-tag’ the samples for downstream demultiplexing, (4) a 5N random sequence on the forward primer and a 4N random sequence on the reverse primer (9N total) as the UMI tags, (5) and 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 9N UMI into 5N + 4N over the forward and reverse primers, we avoid primer dimers. The amplification primers include (1) an index sequence on the forward primer pair for Illumina library demultiplexing, and (2) 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).

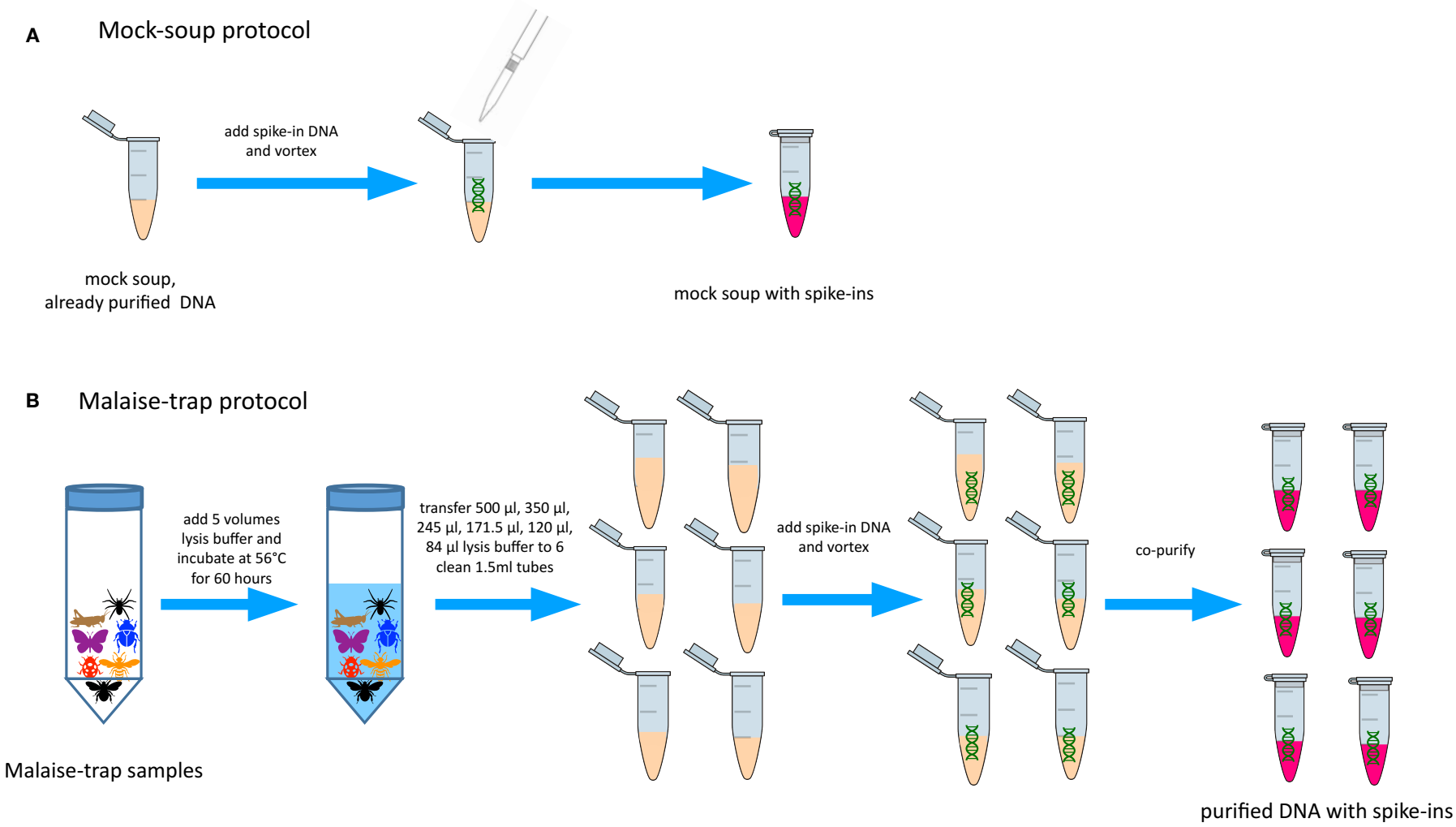



**Figure 5**. **DNA spike-in protocols**. A. Mock-soup protocol. Each mock soup was constructed by pooling purified DNA from individually extracted insects, after which a fixed aliquot of spike-in DNA was added to each pool. B. Malaise-trap-sample protocol. Each bulk sample of arthropods was non-destructively DNA-extracted by soaking in 5X volume of lysis buffer. From each of the 7 samples, 500 µl, 350 µl, 245 µl, 171.5 µl, 120 µl, and 84 µl lysis buffer was used to create 6 dilution soups, a fixed amount of spike-in DNA was added, and the mixture was co-purified. (**94 words**)

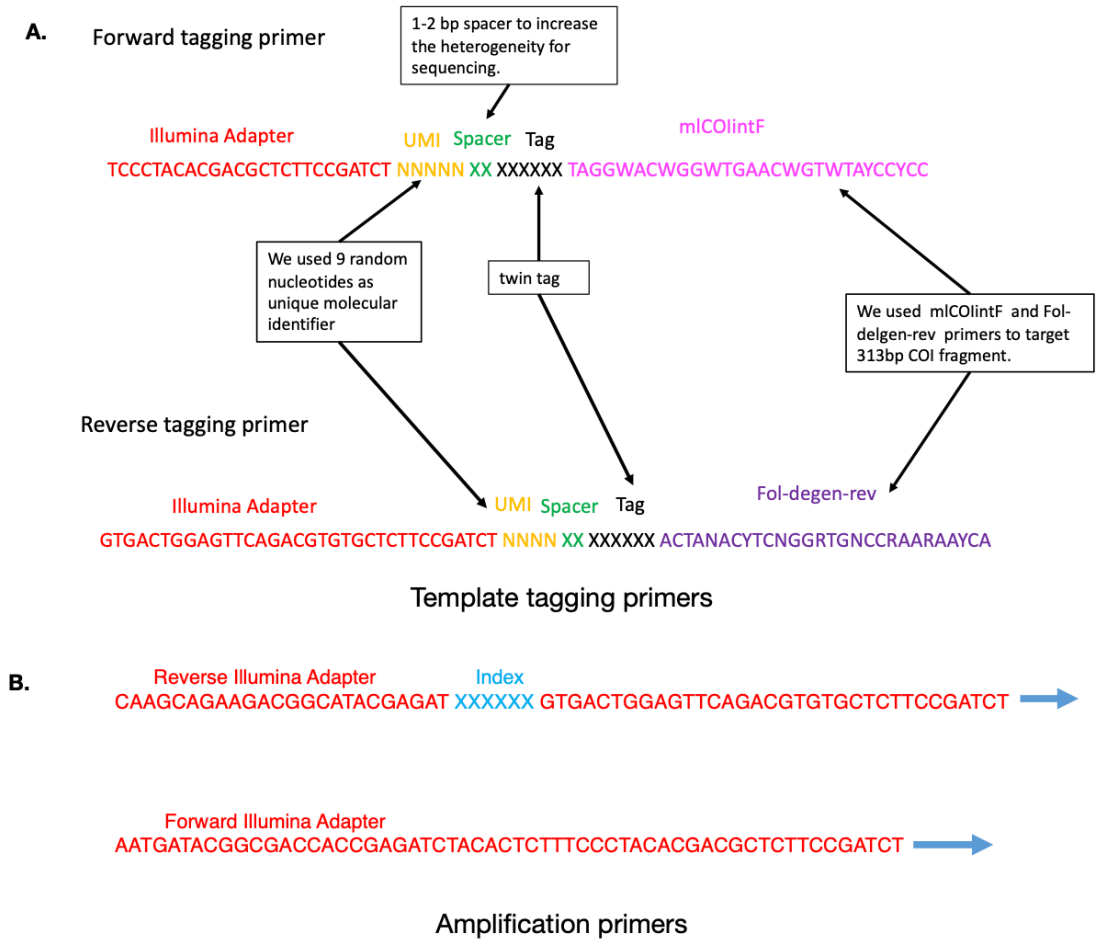



**Figure 6**. **Two-step PCR protocol**. A. In the first step, we amplified a 313 bp COI amplicon using the Leray-FolDegenRev primers. To both the forward and reverse primers, we added a 1-2 nucleotide heterogeneity spacer, the same 6-nucleotide sample tag (twin tags), a 9-random-nucleotide UMI (unique molecular identifier, split as 5 Ns on the forward and 4 Ns on the reverse), and a part of Illumina adapter sequence to anneal to the PCR primers in the second step. B. In the second step, we concatenated a library index sequence and the forward and reverse Illumina adapter sequences. Additional details on primer design (other than the UMI portion) can be found in Yang et al. (2021).(**113 words**)

394 *2.5 PCR and the Begum pipeline*. – The first PCR amplifies COI and concatenates
395 sample tags and UMIs and runs for only two cycles using KAPA 2G Robust HS PCR
396 Kit (Basel, Roche KAPA Biosystems). We used the mlCOIintF-FolDegenRev primer
397 pair (Yu et al. 2012, p. 2012; Leray et al. 2013), which amplifies a 313-bp fragment of
398 the COI barcode; and we followed the Begum protocol (Zepeda-Mendoza et al. 2016;
399 Yang et al. 2021), which is a wet-lab and bioinformatic pipeline that combines multiple
400 independent PCR replicates per sample, twin-tagging and false positive controls to
401 remove tag-jumping and reduce erroneous sequences. Twin-tagging means using the
402 same tag sequence on both the forward and reverse primers in a PCR. Tag jumps
403 almost always result in non-matching tag sequences, and these are identified and
404 removed in the Begum pipeline. We performed 3 PCR replicates per sample, which
405 means we used 3 different twin-tags to distinguish the 3 independent PCR replicates.
406 Begum removes erroneous sequences by filtering out the reads that appear in a low
407 number of PCR replicates (e.g. only one PCR) at a low number of copies per PCR (e.g.
408 only 2 copies), because true sequences are more likely to appear in multiple PCRs with
409 higher copy numbers per PCR. The 20 µl reaction mix included 4 µl Enhancer, 4 µl
410 Buffer A, 0.4 µl dNTP (10 mM), 0.8 µl per primer (10 mM), 0.08 µl KAPA 2G HotStart
411 DNA polymerase (Basel, Roche KAPA Biosystems), 5 µl template DNA and 5 µl water.
412 PCR conditions were initial denaturation at 95°C for 3 minutes, followed by two cycles
413 of denaturation at 95°C for 1 minute, annealing at 50°C for 90 seconds, and extension
414 at 72°C for 2 minutes. Then the products were purified with 14 µl of KAPA pure beads
415 (Roche KAPA Biosystems, Switzerland) to remove the primers and PCR reagents and
416 were eluted into 16 µl of water.

417 The second PCR amplifies the tagged templates for building the libraries that can be
418 sequenced directly on Illumina platform. The 50 µl reaction mix included 5 µl TAKARA
419 buffer, 4 µl dNTP (10 mM), 1.2 µl per primer (10 mM), 0.25 µl TAKARA Taq DNA
420 polymerase, 15 µl DNA product from the first PCR, and 23.35 µl water. PCR conditions
421 were initial denaturation at 95°C for 3 minutes, 5 cycles of denaturation at 95°C for 30
422 seconds, annealing at 59°C for 30 seconds (-1 °C per cycle), extension at 72°C for 30
423 seconds, followed by 25 cycles of denaturation at 95°C for 30 seconds, annealing at
424 55°C for 30 seconds, extension at 72°C for 30 seconds; a final extension at 72°C for 5
425 minutes, and cool down to 4°C.

426 From all second PCR products, 2 µl was roughly quantified on 2% agarose gel with
427 Image Lab 2.0 (Bio-Rad, USA). For each set of PCR reactions with the same index,
428 amplicons were mixed at equimolar ratios to make a pooled library. One PCR negative
429 control were set for each library. We sent our samples to Novogene (Tianjin, China) to

do PE250 sequencing on Illumina NovaSeq 6000, requiring a 0.8 GB raw data from each PCR reaction.

*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 V181 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 07 Sep 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 4 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.

*2.6 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. Model syntax is given in the legend of Figure 7. We used the {MuMIn} 1.43.17 package (CRAN.R-project.org/package=MuMIn, accessed 2 Jan 2022) to calculate the

466 variance explained by fixed effects only (marginal $R^2$). To carry out spike-in correction,
467 we first calculated a weighted mean from the added spike-ins (e.g. mean(Bombycidae
468 + Elateridae/2 + Mordellidae/4)), rescaled the new mean spike-in so that the smallest
469 value is equal to 1, and divided each row's OTU size and UMI number by the weighted,
470 scaled spike-in.

## 471 Results

472 *3.1 Bioinformatic processing of the Malaise-trap samples and the mock soups.* – Five
473 libraries yielded a total of 283,319,770 paired-end reads, of which 247,285,097 were
474 merged successfully in Pandaseq. After *Begum* sorting and demultiplexing, which
475 removed a large number of tag-jumped reads and some reads <300 bp length, we
476 retained 106,649,397 reads. After *Begum*'s filtering of erroneous reads, we retained
477 76,289,802 reads, and after *de-novo* chimera removal, we retained 73,818,971 reads.
478 Sequences were clustered at 97% similarity into 1,188 OTUs, and LULU combined the
479 OTUs of the Malaise-trap samples into 435 OTUs. After removing the spike-in OTUs,
480 the seven Malaise-trap samples contained a total of 432 OTUs. All 52 OTUs of the 7
481 mock soups were recovered.

482 *3.2 Mock soups, COI copy number.* – Without spike-in correction, OTU size (numbers
483 of reads per OTU) predicts almost none of the within-species (dilution-gradient-caused)
484 variation in COI copy number ($R^2$ = 0.04, all values marginal $R^2$), but with spike-in
485 correction, OTU size predicts 42.0% of the variation (Figure 7 AB). As expected, UMI
486 number by itself does not predict input COI copy number ($R^2$ = 0.05), but with spike-in
487 correction, UMI number does predict COI copy number ($R^2$ = 0.42) (Figure 7 CD). Also
488 as expected, spike-in correction does not achieve across-species quantification, as
489 shown by the orders of magnitude variation in intercepts across the 52 OTUs. Note that
490 this experiment pooled DNA extracts with equalised concentrations of genomic DNA
491 mass per species, which suggests that PCR bias is the main source of species bias in
492 this dataset.

493 *3.4 Mock soup within-species abundance in input genomic DNA mass.* – Of course, our
494 goal is to estimate not COI copy number but specimen biomass. We thus tested how
495 well OTU size and UMI numbers predicted genomic DNA concentration. Non-spike-
496 corrected OTU size and UMI number both failed to predict input genomic DNA mass
497 ($R^2$ < 0.02 for both, Figure 8 AC), but spike-corrected OTU size and UMI number again
498 both successfully predicted input genomic DNA mass ($R^2$ = 0.52 and 0.53, Figure 8
499 BD).

*3.5 Malaise-trap within*-species *abundance recovery.* – Recall that each of the 7 selected Malaise-trap samples was serially diluted by 0.7X to create six soups per sample. Non-spike-corrected OTU size did not predict within-species variation in input genomic-DNA mass (p = 0.33) (Figure 9A), but spike-corrected OTU size again did predict within-species variation in input genomic-DNA mass ($R^2$ = 0.53) (Figure 9B).

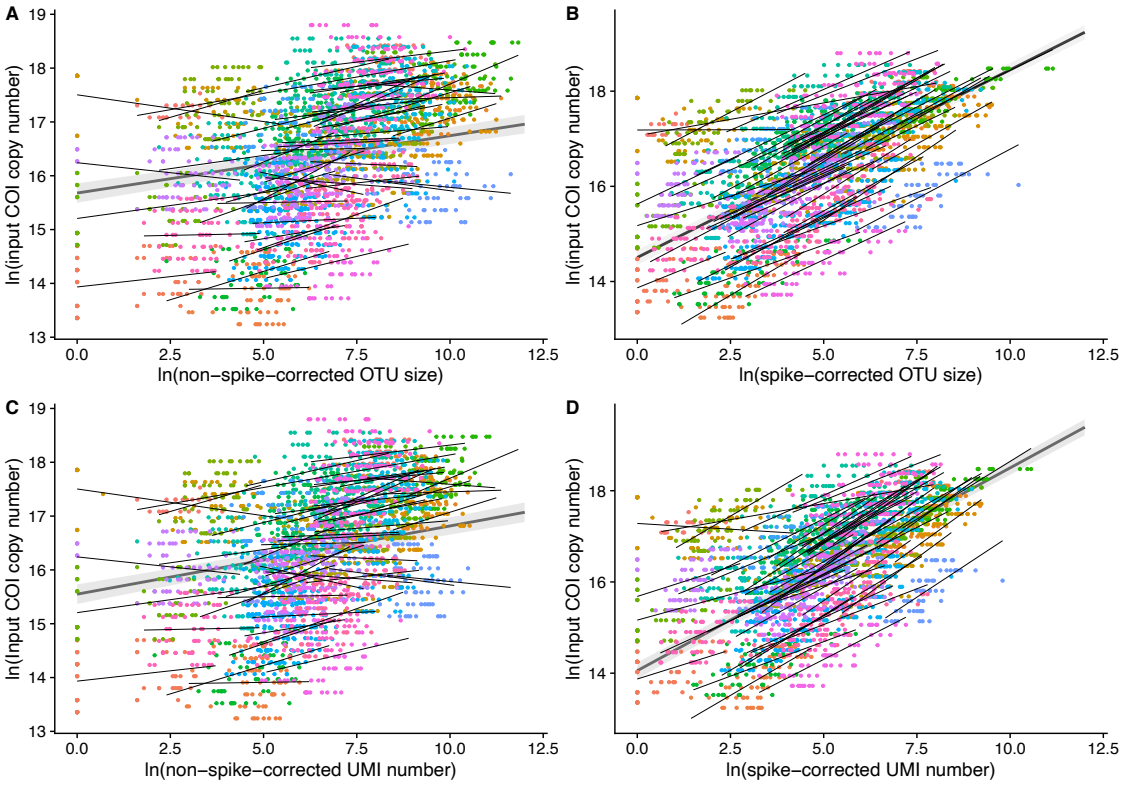



**Figure 7**. **Prediction of within-species variation in COI copy number in the mock-soup experiment**. For visualisation, all data points are shown (including all soup and PCR replicates), each thin line is fit to one of the OTUs across the seven serially diluted mock-soup samples, and the thick line represents the fitted model in which OTUs were treated as a random factor. **A**. Non-spike-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 $R^2$ = 0.04) (Bates et al. 2015). **B**. Spike-corrected OTU size successfully predicts within-species variation in input COI copy number (mixed-effects linear model, marginal $R^2$ = 0.42), but species bias remains, as can be seen in the orders-of-magnitude variation in intercepts (Figure 1). **C**. UMI number (per OTU per soup) poorly predicts within-species variation in input COI copy number (marginal $R^2$ = 0.05). **D**. Spike-corrected UMI number successfully predicts within-species variation in input COI copy number (marginal $R^2$ = 0.42), and again species bias remains, as shown by the different intercepts. Model syntax: `lme4::lmer(log.input_COI ~ log.OTUsize_or_UMINumber + (log.OTUsize_or_UMINumber | OTUID) + (1 |soupRep/pcrRep))`. (**186 words**)

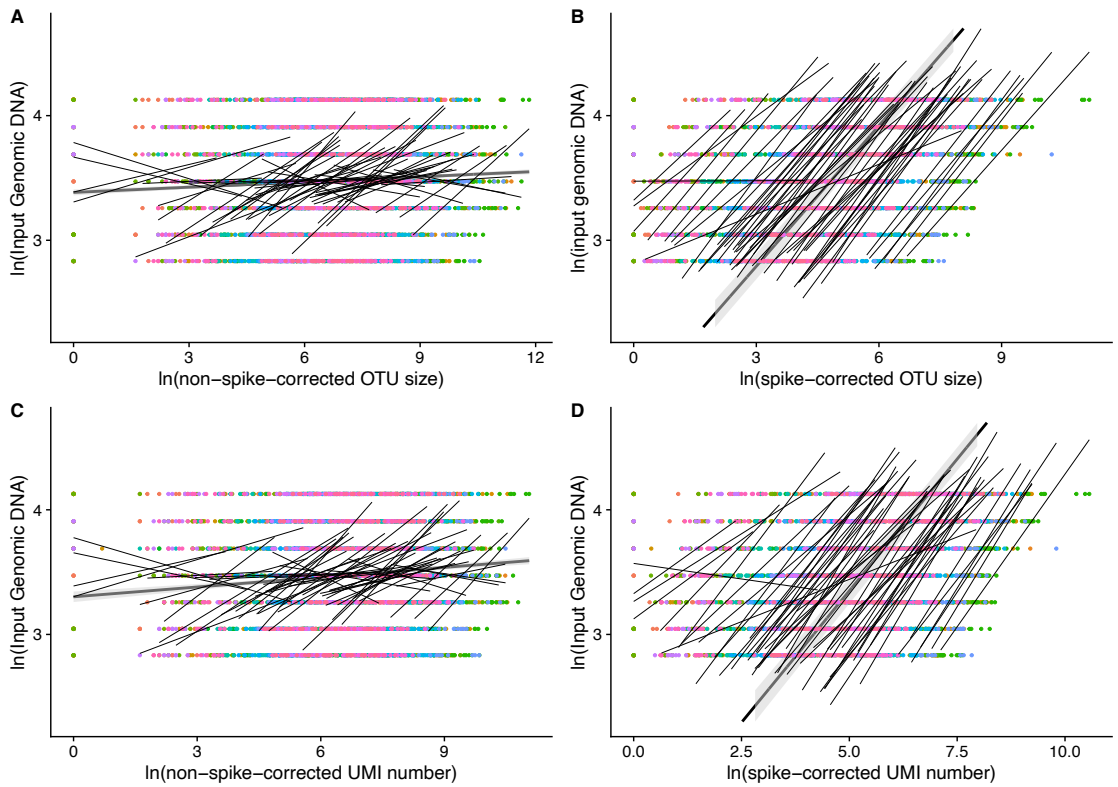



**Figure 8**. **Prediction of within-species variation in genomic DNA concentration in the mock-soup experiment**. For visualisation, all data points are shown (including all soup and PCR replicates), each thin line is fit to one of the OTUs across the seven serially diluted mock-soup samples, and the thick line represents the fitted model in which OTUs were treated as a random factor. **A**. Non-spike-corrected read number poorly predicts within-species variation in input genomic DNA concentration (linear mixed-effects model, marginal $R^2$ = 0.01). B. Spike-corrected read number successfully predicts within-species variation in input genomic DNA concentration (linear mixed-effects model, marginal $R^2$ = 0.52) despite species bias (Figure 1). C. UMI number poorly predicts within-species variation in input genomic DNA concentration (marginal $R^2$ = 0.02). D. Spike-corrected UMI number successfully predicts within-species variation in input genomic DNA concentration (marginal $R^2$ = 0.53) despite species biases. Model syntax: `lme4::lmer(log.input_gDNA ~ log.OTUsize_or_UMINumber + (log.OTUsize_or_UMINumber | OTUID) + (1 |soupRep/pcrRep))`. (**152 words**)

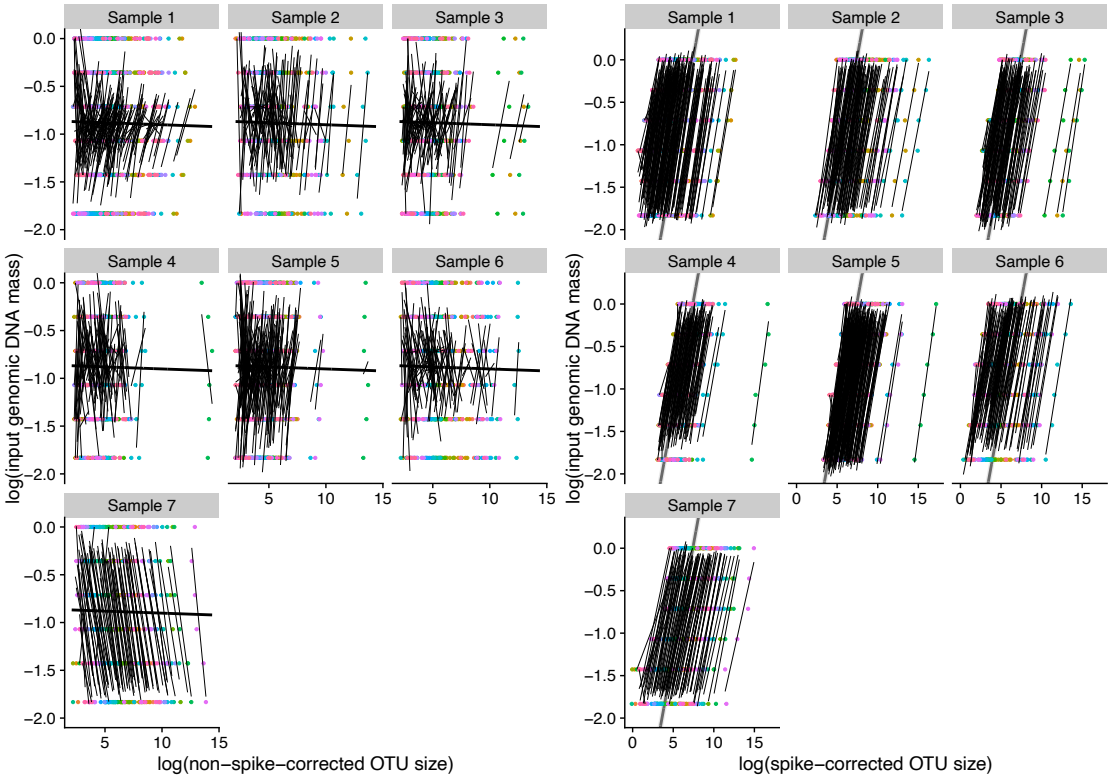



**Figure 9**. **Prediction of within-species variation in genomic DNA concentration in the Malaise-trap samples**. For visualisation, each thin line is fit to an OTU's serial dilution made from each of the seven Malaise-trap samples, and the thick lines are the fitted model with sample and OTU as random factors. There are 176, 113, 111, 104, 196, 110, and 82 OTUs in samples 1-7, respectively. **A**. Non-spike-corrected read number (OTU size) does not predict within-species variation in genomic DNA concentration ($R^2$ = 0.0). **B**. Spike-corrected read number successfully predicts within-species variation in genomic DNA concentration (marginal $R^2$ = 0.53) 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))`. (**130 words**)

## Discussion

We propose that there is a useful distinction to be made between *within*-species abundance change and *across*-species abundance differences (Figures 1, 2). Within-species abundance information is sufficient for inferring species interactions, modelling population dynamics and species distributions, detecting environmental state and change, and inferring false positives and negatives (Figure 3). We experimentally show that spike-ins allow the recovery of within-species abundance change, which is expected since spike-ins remove pipeline noise.

In our first test, we serially diluted 52 OTUs into seven mock soups, and after spike-in correction, we were able to recover within-species abundance change in both input COI copy number (Figure 7) and input genomic DNA (Figure 8), 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, and we were able to recover within-species abundance change in input genomic DNA (Figure 9). Finally, our experimental protocol included Unique Molecular Identifiers (UMIs), which required a laborious two-step PCR protocol, but we found no additional quantification benefit over the spike-in. Note that UMIs have other benefits that could recommend them over a physical spike-in, such as contamination detection and error correction (Fields et al. 2021).

In most cases, practitioners will need to add exogenous spike-ins to their samples (reviewed in Harrison et al. 2021), but in the special but important case of host-associated microbiomes, Lundberg et al. (2021) describe a two-step PCR method to use a single-copy host gene as a built-in spike-in. In the Supplementary Information code for Figure 7 and 8, we apply the model-based pipeline-noise estimator to the mock-soup dataset and achieve an $R^2$=11.8% for prediction of COI copy number, which lies between the $R^2$ values achieved for the non-physical-spike-corrected ($R^2$=0.04) and physical-spike-corrected values ($R^2$=0.042) (Figure 7), and we achieve an $R^2$=21.3% for prediction of genomic DNA, also intermediate between the non-physical-spike-corrected ($R^2$=0.0) and physical-spike-corrected values ($R^2$=0.53 (Figure 8). In the Malaise-trap data, the model-based approach performed poorly at 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 dataset, consistent with a dilution gradient. This behaviour is as expected for the model-based method – it will recover relative not absolute DNA concentrations, hence is a tool best used to study effects on compositional not total abundance.

Statistical analysis of DNA-based datasets 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 utilise abundance data that ranges continuously from [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 available in the {paramedic} *R* package (Williamson et al. 2021) and in the metagenomic protocols of Lang et al. (2019) and 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.

## Data availability statement

Data and R scripts for Figures 7, 8, 9, and the model-based estimator are posted at https://github.com/LoreMJ/quantitive_metabarcoding. All sequence data, reference files, folder structure, output files and scripts (32.5 GB) on DataDryad. (https://datadryad.org/stash/share/0rJ5Yy2PRIv5UpVrCS95Wf7pY0J2R_Hqic6DWyMeaD8)

## Acknowledgments

We thank Sarah Bourlat and Nathan Geraldi for very helpful comments on the manuscript. The authors were supported by the Key Research Program of Frontier Sciences, CAS (QYZDY-SSW-SMC024), Supported by the Strategic Priority Research Program of Chinese Academy of Sciences, Grant No. XDA20050202, the University of East Anglia, the State Key Laboratory of Genetic Resources and Evolution (GREKF19-01, GREKF20-01, GREKF21-01) at the Kunming Institute of Zoology, the University of East Anglia, and the University of Chinese Academy of Sciences.

## References

Abrego, N., T. Roslin, T. Huotari, Y. Ji, N. M. Schmidt, J. Wang, D. W. Yu, et al. 2021. Accounting for species interactions is necessary for predicting how arctic arthropod communities respond to climate change. Ecography 44:885–896.

Amend, A. S., K. A. Seifert, and T. D. Bruns. 2010. Quantifying microbial communities with 454 pyrosequencing: does read abundance count? Molecular Ecology 19:5555–5565.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. Fitting Linear Mixed-Effects Models Using **lme4**. Journal of Statistical Software 67:1–48.

Bell, K. L., J. Fowler, K. S. Burgess, E. K. Dobbs, D. Gruenewald, B. Lawley, C. Morozumi, et al. 2017. Applying Pollen DNA Metabarcoding to the Study of Plant–Pollinator Interactions. Applications in Plant Sciences 5:1600124.

628 Brys, R., D. Halfmaerten, S. Neyrinck, Q. Mauvisseau, J. Auwerx, M. Sweet, and J.
629 Mergeay. 2021. Reliable eDNA detection and quantification of the European weather
630 loach ( *Misgurnus fossilis* ). Journal of Fish Biology 98:399–414.

631 Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K.
632 Costello, N. Fierer, et al. 2010. QIIME allows analysis of high-throughput community
633 sequencing data. Nature Methods 7:335–336.

634 Carraro, L., E. Mächler, R. Wüthrich, and F. Altermatt. 2020. Environmental DNA
635 allows upscaling spatial patterns of biodiversity in freshwater ecosystems. Nature
636 Communications 11:3585.

637 Carraro, L., J. B. Stauffer, and F. Altermatt. 2021. How to design optimal eDNA
638 sampling strategies for biomonitoring in river networks. Environmental DNA 3:157–172.

639 Creedy, T. J., H. Norman, C. Q. Tang, K. Qing Chin, C. Andujar, P. Arribas, R. S.
640 O'Connor, et al. 2020. A validated workflow for rapid taxonomic assignment and
641 monitoring of a national fauna of bees (Apiformes) using high throughput DNA
642 barcoding. Molecular Ecology Resources 20:40–53.

643 Deagle, B. E., L. J. Clarke, J. A. Kitchener, A. M. Polanowski, and A. T. Davidson.
644 2018. Genetic monitoring of open ocean biodiversity: An evaluation of DNA
645 metabarcoding for processing continuous plankton recorder samples. Molecular
646 Ecology Resources 18:391–406.

647 Deagle, B. E., A. C. Thomas, J. C. McInnes, L. J. Clarke, E. J. Vesterinen, E. L. Clare,
648 T. R. Kartzinel, et al. 2019. Counting with DNA in metabarcoding studies: How should
649 we convert sequence reads to dietary data? Molecular Ecology 28:391–406.

650 deWaard, J. R., V. Levesque-Beaudin, S. L. deWaard, N. V. Ivanova, J. T. A.
651 McKeown, R. Miskie, S. Naik, et al. 2019. Expedited assessment of terrestrial
652 arthropod diversity by coupling Malaise traps with DNA barcoding. Genome 62:85–95.

653 Doi, H., K. Fukaya, S. Oka, K. Sato, M. Kondoh, and M. Miya. 2019. Evaluation of
654 detection probabilities at the water-filtering and initial PCR steps in environmental DNA
655 metabarcoding using a multispecies site occupancy model. Scientific Reports 9:3581.

656 Dorazio, R. M., and R. A. Erickson. 2018. ednaoccupancy : An R package for
657 multiscale occupancy modelling of environmental DNA data. Molecular Ecology
658 Resources 18:368–380.

659 Edgar, R. C. 2010. Search and clustering orders of magnitude faster than BLAST.
660 Bioinformatics 26:2460–2461.

Elbrecht, V., and F. Leese. 2015. Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass—Sequence Relationships with an Innovative Metabarcoding Protocol. PLOS ONE 10:e0130324.

Erickson, R. A. 2019. Sampling Designs for Landscape-level eDNA Monitoring Programs. Integr Environ Assess Manag 12.

Fields, B., S. Moeskjær, V. Friman, S. U. Andersen, and J. P. W. Young. 2021. MAUI-seq: Metabarcoding using amplicons with unique molecular identifiers to improve error correction. Molecular Ecology Resources 21:703–720.

Frøslev, T. G., R. Kjøller, H. H. Bruun, R. Ejrnæs, A. K. Brunbjerg, C. Pietroni, and A. J. Hansen. 2017. Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. Nature Communications 8:1188.

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 30:3057–3067.

Garrido-Sanz, L., M. À. Senar, and J. Piñol. 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 1755–0998.13464.

Griffin, J. E., E. Matechou, A. S. Buxton, D. Bormpoudakis, and R. A. Griffiths. 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) 69:377–392.

Harrison, J. G., W. John Calder, B. Shuman, and C. Alex Buerkle. 2021. The quest for absolute abundance: The use of internal standards for DNA-based community ecology. Molecular Ecology Resources 21:30–43.

Hebert, P. D. N., T. W. A. Braukmann, S. W. J. Prosser, S. Ratnasingham, J. R. deWaard, N. V. Ivanova, D. H. Janzen, et al. 2018. A Sequel to Sanger: amplicon sequencing that scales. BMC Genomics 19:219.

Hindson, B. J., K. D. Ness, D. A. Masquelier, P. Belgrader, N. J. Heredia, A. J. Makarewicz, I. J. Bright, et al. 2011. High-Throughput Droplet Digital PCR System for Absolute Quantitation of DNA Copy Number. Analytical Chemistry 83:8604–8610.

Hoshino, T., and F. Inagaki. 2017. Application of Stochastic Labeling with Random-Sequence Barcodes for Simultaneous Quantification and Sequencing of Environmental 16S rRNA Genes. PLOS ONE 12:e0169431.

695 Hoshino, T., R. Nakao, H. Doi, and T. Minamoto. 2021. Simultaneous absolute
696 quantification and sequencing of fish environmental DNA in a mesocosm by
697 quantitative sequencing technique. Scientific Reports 11:4372.

698 Ji, Y., T. Huotari, T. Roslin, N. M. Schmidt, J. Wang, D. W. Yu, and O. Ovaskainen.
699 2020. SPIKEPIPE: A metagenomic pipeline for the accurate quantification of eukaryotic
700 species occurrences and intraspecific abundance change using DNA barcodes or
701 mitogenomes. Molecular Ecology Resources 20:256–267.

702 Krehenwinkel, H., M. Wolf, J. Y. Lim, A. J. Rominger, W. B. Simison, and R. G.
703 Gillespie. 2017. Estimating and mitigating amplification bias in qualitative and
704 quantitative arthropod metabarcoding. Scientific Reports.

705 Lang, D., M. Tang, J. Hu, and X. Zhou. 2019. Genome-skimming provides accurate
706 quantification for pollen mixtures. Molecular Ecology Resources 19:1433–1446.

707 Leray, M., J. Y. Yang, C. P. Meyer, S. C. Mills, N. Agudelo, V. Ranwez, J. T. Boehm, et
708 al. 2013. A new versatile primer set targeting a short fragment of the mitochondrial COI
709 region for metabarcoding metazoan diversity: application for characterizing coral reef
710 fish gut contents. Frontiers in Zoology 10:34.

711 Levi, T., J. M. Allen, D. Bell, J. Joyce, J. R. Russell, D. A. Tallmon, S. C. Vulstek, et al.
712 2019. Environmental DNA for the enumeration and management of Pacific salmon.
713 Molecular Ecology Resources 19:597–608.

714 Li, H. 2015. BFC: correcting Illumina sequencing errors. Bioinformatics 31:2885–2887.

715 Lundberg, D. S., P. P. N. Ayutthaya, A. Strauß, G. Shirsekar, W.-S. Lo, T. Lahaye, and
716 D. Weigel. 2021. Host-associated microbe PCR (hamPCR) enables convenient
717 measurement of both microbial load and community composition. eLife 10:e66186.

718 Lundberg, D. S., S. Yourstone, P. Mieczkowski, C. D. Jones, and J. L. Dangl. 2013.
719 Practical innovations for high-throughput amplicon sequencing. Nature Methods
720 10:999–1002.

721 Lyet, A., L. Pellissier, A. Valentini, T. Dejean, A. Hehmeyer, and R. Naidoo. 2021.
722 eDNA sampled from stream networks correlates with camera trap detection rates of
723 terrestrial mammals. Scientific Reports 11:11362.

724 Masella, A. P., A. K. Bartram, J. M. Truszkowski, D. G. Brown, and J. D. Neufeld. 2012.
725 PANDAseq: paired-end assembler for illumina sequences. BMC Bioinformatics 13:31.

726 McLaren, M. R., A. D. Willis, and B. J. Callahan. 2019. Consistent and correctable bias
727 in metagenomic sequencing experiments. eLife 8:e46923.

Meier, R., W. Wong, A. Srivathsan, and M. Foo. 2016. $1 DNA barcodes for reconstructing complex phenomes and finding rare species in specimen-rich samples. Cladistics 32:100–110.

Nielsen, M., M. T. P. Gilbert, T. Pape, and K. Bohmann. 2019. A simplified DNA extraction protocol for unsorted bulk arthropod samples that maintains exoskeletal integrity. Environmental DNA 1:144–154.

Pauvert, C., M. Buée, V. Laval, V. Edel-Hermann, L. Fauchery, A. Gautier, I. Lesur, et al. 2019. Bioinformatics matters: The accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline. Fungal Ecology 41:23–33.

Peel, N., L. V. Dicks, M. D. Clark, D. Heavens, L. Percival-Alwyn, C. Cooper, R. G. Davies, et al. 2019. Semi-quantitative characterisation of mixed pollen samples using MinION sequencing and Reverse Metagenomics (RevMet). Methods in Ecology and Evolution 10:1690–1701.

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 15:819–830.

Piñol, J., M. A. Senar, and W. O. C. Symondson. 2019. The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. Molecular Ecology 28:407–419.

Pochardt, M., J. M. Allen, T. Hart, S. D. L. Miller, D. W. Yu, and T. Levi. 2020. Environmental DNA facilitates accurate, inexpensive, and multiyear population estimates of millions of anadromous fish. Molecular Ecology Resources 20:457–467.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ratnasingham, S. 2019. mBRAVE: The Multiplex Barcode Research And Visualization Environment. Biodiversity Information Science and Standards 3:e37986.

Rognes, T., T. Flouri, B. Nichols, C. Quince, and F. Mahé. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584.

Rojahn, J., L. Pearce, D. M. Gleeson, R. P. Duncan, D. M. Gilligan, and J. Bylemans. 2021. The value of quantitative environmental DNA analyses for the management of invasive and endangered native fish. Freshwater Biology fwb.13779.

Schenk, J., S. Geisen, N. Kleinboelting, and W. Traunspurger. 2019. Metabarcoding data allow for reliable biomass estimates in the most abundant animals on earth. Metabarcoding and Metagenomics 3:e46704.

Schubert, M., S. Lindgreen, and L. Orlando. 2016. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Research Notes 9:88.

Shelton, A. O., J. L. O'Donnell, J. F. Samhouri, N. Lowell, G. D. Williams, and R. P. Kelly. 2016. A framework for inferring biological communities from environmental DNA. Ecological Applications 26:1645–1659.

Smets, W., J. W. Leff, M. A. Bradford, R. L. McCulley, S. Lebeer, and N. Fierer. 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.

Srivathsan, A., L. Lee, K. Katoh, E. Hartop, S. N. Kutty, J. Wong, D. Yeo, et al. 2021. MinION barcodes: biodiversity discovery and identification by everyone, for everyone. BioRXiv doi:10.1101/2021.03.09.434692.

Stauffer, S., M. Jucker, T. Keggin, V. Marques, M. Andrello, S. Bessudo, M.-C. Cheutin, et al. 2021. How many replicates to accurately estimate fish biodiversity using environmental DNA on coral reefs? BioRxiv doi:10.1101/2021.05.26.445742.

Steinke, D., T. W. Braukmann, L. Manerus, A. Woodhouse, and V. Elbrecht. 2021. Effects of Malaise trap spacing on species richness and composition of terrestrial arthropod bulk samples. Metabarcoding and Metagenomics 5:e59201.

Tang, M., C. J. Hardman, Y. Ji, G. Meng, S. Liu, M. Tan, S. Yang, et al. 2015. High-throughput monitoring of wild bee diversity and abundance via mitogenomics. Methods in Ecology and Evolution 6:1034–1043.

Thalinger, B., A. Rieder, A. Teuffenbach, Y. Pütz, T. Schwerte, J. Wanzenböck, and M. Traugott. 2021. The Effect of Activity, Energy Use, and Species Identity on Environmental DNA Shedding of Freshwater Fish. Frontiers in Ecology and Evolution 9:623718.

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 16:714–726.

Tkacz, A., M. Hortala, and P. S. Poole. 2018. Absolute quantitation of microbiota abundance in environmental samples. Microbiome 6:110.

Verkuil, Y. I., M. Nicolaus, R. Ubels, M. M. Dietz, J. M. Samplonius, A. Galema, K. Kiekebos, et al. 2020. DNA metabarcoding successfully quantifies relative abundances

of arthropod taxa in songbird diets: a validation study using camera-recorded diets. BioRxiv doi:10.1101/2020.11.26.399535.

Wang, Y., U. Naumann, S. T. Wright, and D. I. Warton. 2012. mvabund - an R package for model-based analysis of multivariate abundance data. Methods in Ecology and Evolution 3:471–474.

Warton, D. 2022. Eco-Stats - Data Analysis in Ecology. Methods in Statistical Ecology (1st ed.). Springer International Publishing, Switzerland.

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

Wührl, L., C. Pylatiuk, M. Giersch, F. Lapp, T. von Rintelen, M. Balke, S. Schmidt, et al. 2021. DiversityScanner: Robotic discovery of small invertebrates with machine learning methods. BioRxiv doi:10.1101/2021.05.17.444523.

Yang, C., K. Bohmann, X. Wang, W. Cai, N. Wales, Z. Ding, S. Gopalakrishnan, et al. 2021. Biodiversity Soup II: A bulk-sample metabarcoding pipeline emphasizing error reduction. Methods in Ecology and Evolution 12:1252–1264.

Yates, M. C., M. E. Cristescu, and A. M. Derry. 2021*a*. Integrating physiology and environmental dynamics to operationalize environmental DNA (eDNA) as a means to monitor freshwater macro-organism abundance. Molecular Ecology mec.16202.

Yates, M. C., D. M. Glaser, J. R. Post, M. E. Cristescu, D. J. Fraser, and A. M. Derry. 2021*b*. The relationship between eDNA particle concentration and organism abundance in nature is strengthened by allometric scaling. Molecular Ecology 30:3068–3082.

Yu, D. W., Y. Ji, B. C. Emerson, X. Wang, C. Ye, C. Yang, and Z. Ding. 2012. Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods in Ecology and Evolution 3:613–623.

Zepeda-Mendoza, M. L., K. Bohmann, A. Carmona Baez, and M. T. P. Gilbert. 2016. DAMe: a toolkit for the initial processing of datasets with PCR replicates of double-tagged amplicons for DNA metabarcoding analyses. BMC Research Notes 9:255.

Zhou, X., Y. Li, S. Liu, Q. Yang, X. Su, L. Zhou, M. Tang, et al. 2013. Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. GigaScience 2:4.

# Supplementary Information: Unabridged Methods

*1 Mock soup construction*

*1.1 Input species*. – We used Malaise traps to collect 286 arthropods in Kunming, China (25°8'23" N, 102°44'17" E). DNA was extracted from each individual using the DNeasy Blood & Tissue Kit (Qiagen GmbH, Germany). Mean genomic DNA per species concentration was quantified from three replicates using PicoGreen fluorescent dye. We DNA-barcoded the individuals using the Folmer primer pair LCO1490 5'-GGTCAACAAATCATAAAGATATTGG -3', and HC02198 (5'-TAAACTTCAGGGTGACCAAAAAATCA -3') (Folmer et.al 1994) , with PCR parameters of initial denaturation at 95°C for 3 minutes, followed by 34 cycles of denaturation at 94°C for 30 seconds, annealing at 50°C for 30 seconds and extension at 72°C for 1 minute, a final extension at 72° C for 5 minutes, and a cooldown to 4°C. After the 658-bp COI sequences were cut to 313 bp based on our metabarcoding primers (see below), 286 arthropods were clustered into 168 OTUs at 97% similarity. To create mock soups, we selected 52 individuals with genomic DNA >20 ng/µl, representing 52 OTUs.

*1.2 OTU quantification*. – We quantified COI concentrations of the 52 OTUs using the mean of three qPCRs with the *Leray-FolDegenRev* primer pair (Yu et al. 2012; Leray et al. 2013). First, to create a standard DNA curve, we used a purified and sequenced COI amplicon (Lepidoptera Bombycidae *Bombyx mori*) and quantified its concentration with a Qubit, from which we calculated COI copy number from the COI concentration (ng/µl), molar mass (233530.15 g/mol), and the Avogadro constant.

$$COI\ copy\ number\ = \frac{COI\ concentration \times 10^{-9}}{molar\ mass} \times 6.02 \times 10^{23} \quad \text{(Eq. 1)}$$

This was 0.5X serially diluted into six concentration levels, as follows:

1. Start by diluting the above PCR amplicon to 100x by transferring 5 µl of amplicon to 495 µl of Ultragrade water. Vortex for 20 sec. Label as Tube 1.
2. Label 5 tubes from 2 to 6 and add 250 µl of water into each tube.
3. Transfer 250 µl from Tube 1 to Tube 2, and vortex for 20 sec.
4. Transfer 250 µl from Tube 2 to Tube 3, and vortex for 20 sec.
5. Continue the dilution series to Tube 6.
6. Use Tubes 1-6 as the standard DNA gradient for qPCR.

For each of the 52 OTUs, we used qPCR to quantify COI concentration (copy/µl). qPCRs were performed in a total volume of 20 µl with 10 µl of TAKARA SYBR premix Ex Taq (TaKaRa Biosystems, Dalian, China), 0.4 µl of Rox, 0.8 µl each of 10 µM of Leray and FolDegenRev primers (synthesized by Invitrogen, Shanghai, China), 2 µl of template DNA, and 6 µl of PCR-grade water. The qPCR conditions were 30 sec of initial denaturation at 95°C, followed by 35 cycles of denaturation at 95°C for 10 sec, 30 sec of annealing at 50°C, and 30 sec of extension at 72°C. Each OTU was quantified twice.

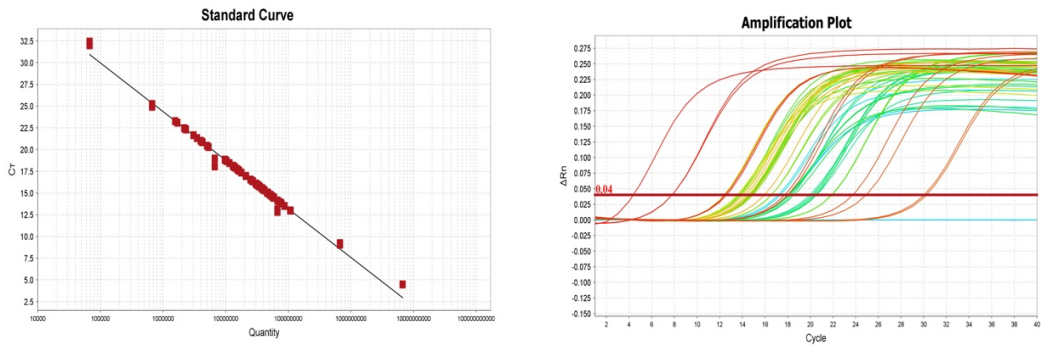



Figure **qPCR**. Left. qPCR standard curve (CT ~ input-COI concentration). Right. qPCR amplification curves, with the 12 (= 2 replicates x 6 concentrations) red lines representing the standard-DNA curves. The other colours represent the 52 OTUs, each amplified twice. All OTU curves lie within the bounds of the standard-DNA curves, allowing interpolation of absolute COI concentrations.

*1.3 Creation of mock soups*. – Each mock soup was constructed with equal masses of purified genomic DNA from 52 OTUs. Input genomic DNA masses of each OTU from soup 1 to soup 7 decreased by 0.8X to create seven mock soups of differing absolute abundances. The individual-OTU genomic DNA masses for the seven mock soups were 61, 48.8, 39, 31.2, 25, 20, and 16 ng and were independently added to the mock soups. Each of the seven soups was made in triplicate ($n_{tot}$ = 21).

*2 Preparation of Malaise-trap samples*

2.1 *Malaise trap sampling*. – 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, United States in July 2018. Traps were left for 7 non-rainy 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.

886 2.2 *DNA extraction*. – The samples were air dried individually on filter papers for less
887 than an hour and then transferred to 50 ml tubes or 5 ml tubes according to sample
888 volume. The samples were then weighed. DNA was non-destructively extracted by
889 soaking the samples in lysis buffer, using the protocol from Ji et al. (2020) and Nielsen
890 et al. (2019). In short, we added lysis buffer to each sample at a 5 ml:1 mg volume:air-
891 dried-mass ratio and incubated the samples in an incubator at 56°C, shaking for 60 hrs.
892 The recipe to make 50 ml of lysis buffer is 2 ml of 1M Tris-HCl buffer (PH 8.0), 1 ml of
893 5M sodium chloride, 10 ml of 10% Sodium Dodecyl Sulfate, 150 µl of 1M calcium
894 chloride, 34.225 ml of PCR-grade water, 2 ml of 1M Dithiothreitol, and 625 µl of
895 Proteinase K solution (20 mg/ml). After extraction, visual checks confirmed that the
896 extracted insects largely retained their body shapes, although they became darker.
897 Note that non-destructive DNA extraction increases the effects of species biases since
898 species with hard exoskeletons likely release less tissue.

899 For this study, we selected seven samples spread over the study area. After
900 completion of lysis, we serially diluted the 7 samples by using 0.7X lysis buffer volume
901 (500 µl, 350 µl, 245 µl, 171.5 µl, 120 µl and 84 µl) to create six soups per sample ($n_{tot}$ =
902 42). We used QIAquick PCR purification kit (Qiagen GmbH, Germany) by following the
903 manufacturer instructions to purify lysis buffer on one spin column per soup.

904 *3 Adding spike-in DNA*

905 *3.1 Spike-in DNA*. – For our spike-ins, we used three insect species from China, none
906 of which is expected to appear in the Oregon Malaise-trap samples (Table S1). An
907 alternative is to use one or more synthetic, random DNA sequences (Tkacz et al.
908 2018). Each of our three spike-ins is represented by a 658-bp COI fragment (Table 1)
909 with primer binding sites that match the Folmer primers HCO2198 and LCO1490. For
910 long-term storage, we inserted the COI fragments as plasmids into monoclonal
911 bacteria. Plasmids were extracted using TIANprep Mini Plasmid Kit (Beijing, China)
912 following manufacturer's instructions.

913 qPCR was used to quantify the three spike-in DNA COI concentrations (ng/µl) using
914 TAKARA SYBR Premix Ex Taq (TaKaRa Bio, Japan). COI copy numbers of the three
915 spike-in DNA were calculated using equation 1.

916 In our plan, we will mix the three spike-ins in a ratio of 1:2:4
917 (Bombycidae:Elateridae:Mordellidae), to check for error and degradation; the reads
918 from the three spike-ins should be found in the ratio of 1:2:4, but if, for instance, one of

919 the spike-ins is not at the correct ratio with the other two, we can omit that species'
920 reads and recalculate the spike-in correction.

| species | full COI barcode sequences |
|---|---|
| Lepidoptera Bombycidae<br>*Bombyx mori* | AACATTATATTTTATTTTTGGTATTTGATCAGGAATAATTGGAACATCTTTAAGACTT<br>TTAATTCGAGCTGAATTAGGAAATCCAGGATCATTAATTGGAGATGATCAAATTTAT<br>AATACTATTGTAACAGCACATGCTTTTATTATAATTTTTTTTATAGTTATACCTATTAT<br>AATTGGAGGATTTGGAAATTGATTAGTTCCTCTTATACTAGGAGCACCAGATATAG<br>CATTCCCACGAATAAATAATATAAGATTTTGACTCCTACCCCCCTCCCTTATATTAT<br>TAATTTCAAGAAGAATTGTAGAAAATGGTGCAGGAACAGGATGAACAGTTTACCCC<br>CCACTTTCATCTAATATCGCACATAGAGGAAGATCCGTAGATCTTGCTATTTTTTCA<br>CTACATTTAGCAGGTATTTCATCAATTATAGGAGCAATTAATTTTATTACAACAATA<br>ATTAATATACGATTAAATAATATATCATTTGATCAATTACCCTTATTTGTATGAGCTG<br>TAGGGATTACAGCATTTTTATTATTATTATCACTACCTGTTTTAGCTGGAGCTATTA<br>CAATATTATTAACAGATCGAAACTTAAATACATCATTTTTTGATCCTGCTGGAGGAG<br>GAGACCCAATTTTATATCAACATTTATTT |
| Coleoptera Elateridae sp. | AACTCTATACTTCATTTTTGGTGCCTGATCAGGAATACTAGGAACATCCTTAAGATT<br>ACTAATCCGTGCAGAACTAGGAAACCCAGGGGCACTAATTGGAAATGACCAAATT<br>TACAATGTTGTTGTTACAGCCCATGCATTTATCATAATTTTCTTCATAGTTATACCTA<br>TCATAATTGGTGGATTCGGAAATTGGTTAGTCCCTTTAATGCTAGGGGCACCAGAT<br>ATAGCTTTCCCTCGAATAAATAACATAAGATTCTGACTTCTCCCACCTTCCCTCAG<br>ATTGCTTTTAATGAGTAGAATTGTTGAAAGAGGGGCTGGGACTGGATGAACTGTTT<br>ATCCACCTTTAGCAGCAAACATTGCCCACAGAGGATCTTCTGTAGATTTAGCTATT<br>TTTAGTCTCCATTTAGCAGGAATTTCATCAATTTTAGGCGCTGTAAATTTCATCTCA<br>ACTGTAATCAACATGCGATCCACAGGGATAACTTTTGACCGTATACCTCTATTTGT<br>TTGAGCAGTTGCCATTACAGCCCTCCTTCTTCTTTTGTCTCTTCCTGTCCTTGCAG<br>GAGCAATTACAATACTTTTAACAGACCGAAACTTAAACACCTCATTTTTTGACCCAG<br>CAGGAGGAGGGGACCCAATTCTTTACCAACACTTATTC |
| Coleoptera Mordellidae sp. | AACACTTTATTTTATTTTTGGGGCCTGGGCAGGCATATTGGGAACGTCTCTAAGAT<br>TATTAATTCGATCAGAATTAGGAACCCCAGGAACTCTGATTGGAGATGACCAAATT<br>TACAACGTAATTGTTACAGCTCATGCTTTCGTAATAATTTTTTTTATAGTTATGCCT<br>GTAATAGTAGGGGGATTCGGCAATTGATTAGTTCCTCTTATATTAGGAGCACCTGA<br>CATAGCATTCCCCCGACTCAATAATATAAGATTTTGATTATTGCCCCCCTCTCTTTC<br>TTTATTACTAATGAGAAGATTAGTCGAAAACGGTGCCGGAACAGGATGTACAGTAT<br>ACCCCCCTTTATCTGCAAATCTAGCCCACGGAGGAGCTTCTGTAGATCTAGCCAT<br>CTTCAGTCTACATCTAGCAGGAATCTCATCAATTTTAGGAGCAATTAATTTTATCTC<br>AACTATACTTAATATACGTCCAATAGGAATAACATTAGATCGACTACCTTTATTTGT<br>TTGAGCTATTGTAATTACAGCAGTCCTATTGCTTTTATCCCTACCCGTTCTCGCAG<br>GGGCTATTACCATATTATTAACAGATCGAAATCTAAATACCTCATTTTTTGACCCTG<br>CAGGAGGGGGAGATCCTATTCTTTATCAACACTTATTC |

921 Table S1. DNA-barcode sequences of the three spike-in species.

| species | COI concentration (ng/µl) | COI copy number (copy/µl) |
|---|---|---|
| Lepidoptera Bombycidae<br>*Bombyx mori* | 0.0002 | $5.17 * 10^5$ |
| Coleoptera Elateridae sp. | 0.3914 | $1.02 * 10^9$ |
| Coleoptera Mordellidae sp. | 0.05807 | $1.51 * 10^8$ |

922 Table S2. COI concentrations of the three spike-in species

923 *3.2 Adding spike-in to the mock soups*. – Adding too much spike-in wastes sequencing
924 data, while adding too little risks loss of abundance information in at least some
925 samples when the number of spike-in reads is too low to use as a reliable correction
926 factor. Thus, we quantified the COI copy numbers of the mock soups by qPCR and
927 chose a volume so that spike-in reads should make up 1% of the total number of COI
928 copies in the lowest-concentration mock soups, balancing efficiency with reliability.

929 Because we mixed 52 individual OTUs to create the mock soups, we can sum the input
930 COI copy numbers of all 52 OTUs to calculate the COI copy number of the lowest-
931 concentration mock soup ($5.25 * 10^8$ copies). We used 1% of this number to calculate
932 the mass of total spike-in to add to all 21 soups ($5.25 * 10^6$ copies). To mix spike-ins
933 DNA conveniently, the Elateridae spike-in was diluted 1000-fold to $1.02 * 10^6$ copies
934 and the Mordellidae spike-in was diluted 100-fold to $1.51 * 10^6$ copies before mixing.
935 Then we added 1.45 µl of *Bombyx* spike-in ($1.45 * 5.17 *10^5 = 0.75 * 10^6$), 1.5 µl of
936 diluted Elateridae ($1.5 * 1.02 * 10^6 = 1.5 * 10^6$) and 2 µl of diluted Mordellidae ($2 * 1.51$
937 $* 10^6 = 3 * 10^6$) to each soup.

938 The sum of the three spike-ins was thus: $(0.75+1.5+3)\times 10^6 = 5.25\times 10^6$, and their ratio
939 was: $0.75 \times 10^6 :1.5 \times10^6 :3 \times 10^6$=1:2:4.

940 Spike-in DNA was added directly to the mock soups' DNA since they were already
941 purified (Figure 5).

942 *3.3 Adding spike-in to the Malaise-trap samples*. – From the 244 Malaise-trap samples,
943 we first extracted 17 Malaise-trap samples without adding spike-ins, and then in order
944 to decide how much spike-in should be added, we used qPCR to quantify the COI
945 concentrations of these 17 Malaise-trap samples, using the same protocol that we used
946 for the mock-soup OTUs. The mean COI concentration per sample was 0.006 ng/µl, for
947 each of which we added 200 µl volume. Mean COI mass per aliquot was thus 1.2 ng
948 (0.006 ng/µl x 200 µl), and 1% is 0.012 ng.

949 Because the Bombycid DNA spike-in had degraded after we arrived in University of
950 Oregon where we extracted DNA from the Malaise-trap samples, we used only the
951 other two spike-in species for the Malaise trap samples, at a ratio of 1:9
952 (Mordellidae:Elateridae). Mordellidae and Elateridae were re-quantified by qPCR and
953 were diluted 100-fold. We added 2.7 µl of diluted Elateridae (2.7 µl * 0.004 ng/µl =
954 0.0108 ng) and 2 µl of diluted Mordellidae (2 µl * 0.0006 ng/µl = 0.0012 ng) to make
955 spike-ins' mixture for each Malaise-trap sample.

956 The sum of the two spike-ins: 0.0012 ng + 0.0108 ng = 0.012 ng, and their ratio was:
957 0.0012:0.0108 = 1:9.

958 We used the same spike-in mass for all the Malaise-trap samples except for the 17
959 samples that had already been extracted. For the selected 7 samples used in this
960 study, lysis buffer (500 µl, 350 µl, 245 µl, 171.5 µl, 120 µl and 84 µl) from each sample
961 was transferred into a clean 1.5 ml tube, and the spike-in DNA was added and
962 vortexed for 10 sec. We then extracted DNA with the Qiagen QIAquick PCR purification
963 kit, following the manufacturer instructions. DNA was eluted with 200 µl of elution
964 buffer. In this way, the spike-in DNA was co-purified, co-amplified, and co-sequenced
965 along with the sample DNA (Figure 5). We also recorded the total lysis buffer volume of
966 each sample, for downstream correction.

967 *4 Primer design*

968 For this study, we simultaneously tested two methods for extracting abundance
969 information: spike-ins and UMIs (Unique Molecular Identifiers). UMI tagging requires a
970 two-step PCR procedure, first using tagging primers and then using amplification
971 primers (Figure 6). The tagging primers include (1) the Leray-FolDegenRev primer pair
972 to amplify the 313-bp COI amplicon of interest, (2) a 1- or 2-nucleotide heterogeneity
973 spacer on both the forward and reverse primers to increase sequence entropy for the
974 Illumina sequencer, (3) the same 6-nucleotide sequence on both the forward and
975 reverse primers to 'twin-tag' the samples for downstream demultiplexing, (4) a 5N
976 random sequence on the forward primer and a 4N random sequence on the reverse
977 primer (9N total) to act as the UMI tags for across-species abundance estimation, (5)
978 and parts of Illumina universal adapter sequences to anneal to the 3' ends of the
979 forward and reverse primer regions for the second PCR. By splitting the 9N UMI into
980 5N + 4N over the forward and reverse primers, we avoid primer dimers. The
981 amplification primers include (1) the annealing primer pair to bind to the amplicons of
982 the first PCR, (2) the full length of the Illumina adapter sequences. For further
983 explanation of the design of the tagging primers (except for the UMI sequences), see
984 Yang et al. (2021).

985 *5 PCR and the Begum pipeline*

986 We performed a two-step PCR (Lundberg et al. 2013).

987 The first PCR amplifies COI and concatenates sample tags and UMIs and runs for only
988 two cycles using KAPA 2G Robust HS PCR Kit (Basel, Roche KAPA Biosystems). We
989 used the mlCOIintF-FolDegenRev primer pair (Yu et al. 2012; Leray et al. 2013), which

amplifies a 313-bp fragment of the COI barcode; and we followed the *Begum* protocol (Zepeda-Mendoza et al. 2018, Yang et al. 2021), which is a wet-lab 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 primer in a PCR. We performed 3 PCR replicates per sample, which means we used 3 different twin-tags to distinguish the 3 independent PCR replicates. *Begum* removes erroneous sequences by filtering out the sequences that appear in a low number of PCR replicates (e.g. one) at a low number of copies per PCR (e.g. 4 copies), because true sequences are more likely to appear in multiple PCRs and with high copy numbers. The 20 µl reaction mix included 4 µl Enhancer, 4 µl Buffer A, 0.4 µl dNTP (10 mM), 0.8 µl of 10 mM forward primer, 0.8 µl of 10 mM reverse primer, 0.08 µl KAPA 2G HotStart DNA polymerase (Basel, Roche KAPA Biosystems), 5 µl template DNA and 5 µl water. PCR conditions were initial denaturation at 95°C for 3 minutes, followed by two cycles of denaturation at 95°C for 1 minute, annealing at 50°C for 90 seconds, and extension at 72°C for 2 minutes. Then the products were purified with 14 µl of KAPA pure beads (Roche KAPA Biosystems, Switzerland) 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 which can be sequenced directly on Illumina platform. The 50 µl reaction mix included 5 µl TAKARA buffer, 4 µl dNTP (10 mM), 1.2 µl of 10mM forward primer, 1.2 µl of 10mM reverse primer, 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 minutes, 5 cycles of denaturation at 95°C for 30 seconds, annealing at 59°C for 30 seconds (-1 °C per cycle), extension at 72°C for 30 seconds, followed by 25 cycles of denaturation at 95°C for 30 seconds, annealing at 55°C for 30 seconds, extension at 72°C for 30 seconds; a final extension at 72°C for 5 minutes, and cooldown to 4°C.

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

*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* v181 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 07 Sep 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 4 reads per PCR, with minimum length of 300 bp. This stringency minimized the false positive reads in the negative PCR control. *vsearch* 2.14.1 (Rognes et al. 2016) was used to remove chimeras (`--uchime_denovo`). *Sumaclust* 1.0.2 was used to cluster the sequences into 97% similarity OTUs. The python script tabulateSumaclust.py from the DAMe toolkit was used 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). We also removed any OTUs in which we found stop codons. The OTU table and OTU representative sequences were used for downstream analysis.

Begum removed UMIs while sorting tags. Because of our complicated primer structure, there is no software available for our data to count the number of UMIs per OTU. We wrote our own bash scripts to process data from the Pandaseq-merged files, which includes all the UMIs, tags, and primers. We first used *Begum*-filtered sequences as a reference to filter reads in each PCR set and put the UMI information on the read headers. Then we carried out reference-based OTU clustering for each PCR with QIIME 1.9.1(`pick_otus.py -m uclust_ref -s 0.99`), using the OTU representative sequences as the reference, counted UMIs and reads for each OTU in each PCR set, and generated the UMI and READ OTU tables.