

RESEARCH ARTICLE

Estimating absolute DNA abundance from replicated metabarcoding

Andrew Olaf Shelton¹  | Shannon M. Kachel^{2,3}  | Megan R. Shaffer² |
 Pedro F. P. Brandão-Dias²  | Eiren K. Jacobson⁴  | Len Thomas⁴ | Ryan P. Kelly² 

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

²School of Marine and Environmental
 Affairs, University of Washington, Seattle,
 Washington, USA

³Woodland Park Zoological Society,
 Seattle, Washington, USA

⁴Centre for Research into Ecological
 and Environmental Modelling, School of
 Mathematics and Statistics, University of
 St Andrews, St Andrews, Scotland, UK

Correspondence

Andrew Olaf Shelton
 Email: ole.shelton@noaa.gov

Funding information

Office of Naval Research, Grant/Award
 Number: N00014-22-1-2719; NOAA
 Fisheries Omics Strategic Initiative

Handling Editor: Huijie Qiao

Abstract

1. Datasets derived from mixed-template DNA amplification and sequencing ('metabarcoding') are common in environmental DNA (eDNA), microbiome or microbial ecology analyses. Such metabarcoding data have been broadly used to inform estimates of biodiversity, distribution and abundance across a wide range of species and habitats. Generally, data from metabarcoding consist of a table of read-counts assigned to each of a set of taxa or sequence variants. As such, they appear to contain quantitative information about the environment under study, immediately eligible for traditional community ecology analyses.
2. In fact, metabarcoding data are compositional, can be biased by variation in PCR amplification rates among taxa, and have other statistical attributes that are distinct from many other common ecological observations. The difficulty in understanding and accounting for the complexity of metabarcoding observations has hindered their ecological interpretation and limited their integration with other ecological data types.
3. Here, we use both simulations and empirical observations to show that the variability among replicate metabarcoding observations is closely tied to the absolute quantities of DNA present in a sample. We then develop a general statistical model for estimating DNA abundance from metabarcoding observations using the relationship between mean and variance of metabarcoding reads.
4. We demonstrate that the model can effectively estimate DNA abundances using simulations and from experimentally created DNA samples. The model can be applied to a single marker or to multiple markers simultaneously, allowing the direct integration of information from multiple genetic markers derived from a common biological sample.
5. Estimates of absolute DNA abundance are the logical unit to connect DNA observations to organismal abundance or biomass. Our methods can be used to link together multiple types of DNA observations or to connect DNA observations with non-genetic ecological methods for surveying ecosystems. However, important aspects of the 'ecology of DNA' remain unresolved and may limit our ability to link observed concentrations to organismal abundances in a general way.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2026 The Author(s). *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society. This article has been contributed to by U.S. Government employees and their work is in the public domain in the USA.

KEYWORDS

compositional analysis, DNA concentration, environmental DNA, metabarcoding

1 | INTRODUCTION

Residual genetic information collected from environmental media such as water, air, or soil (environmental DNA, eDNA) derives from an ever-increasing variety of habitats and is used for a broad spectrum of purposes including monitoring biodiversity (Beng & Corlett, 2020; Bohmann & Lynggaard, 2023; Iacarus et al., 2025), detecting invasive species (Egan et al., 2015; Dougherty et al., 2016; Thomas et al., 2020; Duprey et al., 2023; Sepulveda et al., 2023), and providing abundance indices (Keller et al., 2022; Nevers et al., 2018; Shelton et al., 2019). These applications use a steadily expanding set of molecular tools (Blasko & Phelps, 2024; Tsuru et al., 2021; Williams et al., 2019), but most are derived from either quantitative PCR (qPCR) and related approaches—which measure the concentration of DNA in a sample that matches a specific DNA primer sequence—or DNA metabarcoding—which uses PCR amplification and DNA sequencing to observe DNA sequences from multiple taxa in a sample simultaneously. Metabarcoding techniques are general and flexible, with different primers designed to recover DNA sequences ranging from a narrow selection of closely related species (Ledger et al., 2025) up to large swaths of the tree of life (Leray et al., 2013; Mori et al., 2014). Analytical methods have advanced in parallel, expanding the inferences that can defensibly be drawn from qPCR and metabarcoding. In particular, substantial advances have come from developing methods that integrate multiple eDNA data types (e.g. combining qPCR and metabarcoding data; Allan et al., 2023; Guri et al., 2024; Pont, 2024) or formally linking eDNA data with data from traditional survey methods (e.g. Guri et al., 2024; Keller et al., 2022).

Metabarcoding data are rich and for this reason broadly used, but they are statistically challenging. During metabarcoding, PCR generates an enormous number of amplicons, of which only a small and variable fraction are sequenced and observed. The resulting observations are counts of copies for each unique genetic sequence, referred to as 'reads' or 'read-counts'. These sequences can then be compared against databases of genetic sequences and assigned to taxa, providing a table of read-counts by taxa for each sample. One consequence of the technology is that metabarcoding data are compositional (Gloor et al., 2017; McLaren et al., 2019): observations of multiple taxa within a sample are not independent of one another and information about the link between starting DNA abundance and observed read-counts is lost during laboratory processing. Despite this fact, many authors have sought to use metabarcoding observations to infer the quantity of DNA present in a sample for multiple taxa, proposing ways of summarizing metabarcoding observations and linking them to organismal or DNA abundance. For example, Suren et al. (2024) and many authors used summaries of read-counts as the relevant metabarcoding metric; others, such as Harper et al. (2018), used read-proportions; Laporte et al. (2021) advocated square-root or other transforms of

read-counts; Kelly et al. (2019) used a dual normalization procedure and assumed a constant underlying species community, and so forth (see Bylemans et al., 2025 for a review of some analytical choices). Other researchers have used laboratory methods such as adding internal controls of known concentration ('spike-ins', Ushio et al., 2018) or unique molecular identifiers (Fields et al., 2021) to samples to aid quantification. For 'spike-ins,' a known quantity of DNA from a non-target organism is added to each sample, allowing for the remaining taxa in the sample to be quantified (Ushio et al., 2018). However, ensuring the amount of spike-in DNA is enough to be observed but not so much as to swamp other species in a sample remains a technical challenge. This wide array of methods reveals the complexity of the laboratory processes underlying metabarcoding observations and the difficulty of linking observed patterns to insights about abundance.

Other researchers have acknowledged the difficulty of linking metabarcoding read-counts to measures of abundance and turned to the statistical machinery of occupancy models (Fukaya & Hasebe, 2025; McClenaghan et al., 2020; Tetzlaff et al., 2024) to investigate patterns of DNA occurrence (e.g. Ficetola et al., 2015; Johnsen et al., 2020). While this approach is common and appealing for its quantitative framework, it is not ideal for several reasons. First, converting to presence-absence does not eliminate the fact that metabarcoding data are compositional; it only obscures the dependence among observations. Second, converted metabarcoding data may not meet a common assumption of occupancy modelling—that the probability of detection for a taxon is constant among observations (or otherwise modelled via covariates or appropriate functional forms, see Royle, 2006). Unmodeled heterogeneity in detection probabilities can have severe impacts on the reliability of occupancy parameters (Miller et al., 2015). Third, converting to presence-absence necessarily results in the loss of quantitative information contained in the magnitude of observed read-counts (but see Fukaya & Hasebe, 2025 for an approach that uses read-count data). This means that converting to presence-absence makes connecting to a biologically relevant measure—like taxon density or abundance—difficult. There is value in occupancy models derived from eDNA and statistically rigorous approaches are certainly possible and useful (e.g. Griffin et al., 2020), but occupancy models are not the only available option for metabarcoding data.

Recent developments have advanced sophisticated statistical approaches for metabarcoding data derived directly from read-counts (Diana et al., 2025; Willis & Martin, 2022), but these approaches have not been widely applied in the ecological literature and deserve further attention and development. Here we work to advance methods that help bridge the gap between these statistical tools and ecological applications.

The challenges of metabarcoding analysis only grow as researchers begin to query samples for information using more than one primer region. Such multi-marker analyses result in a far fuller

picture of the underlying biota (Espinosa Prieto et al., 2024; Kelly et al., 2017), but leave the researcher with multiple compositional datasets often having only partial taxonomic overlap and no obvious way to combine them for joint ecological inference. Venn diagrams or equivalent tables can be useful as summaries of observations (Kelly et al., 2017), but are not themselves substitutes for quantitative analysis linking multiple markers. The result is a situation in which richer information can paradoxically yield less meaning.

As eDNA metabarcoding continues to gain traction as a source of ecological data, researchers are increasingly likely to query samples with multiple markers (either for single- or multi-taxon work) and/or use replicate sequencing of individual markers to make the most out of often-precious environmental samples (Crowley et al., 2024; Djurhuus et al., 2020; Eastwood et al., 2024; Wilkinson et al., 2024). It is therefore increasingly important to develop tools to jointly analyse multi-marker and replicated metabarcoding datasets for ecological inference.

Here we set out a framework for understanding the causes of common patterns observed in metabarcoding data and present methods to derive more information from common metabarcoding data. We advance methods for linking and combining information from metabarcoding observations in two parts. In Sections 2.1 and 3.1, we summarize several features and patterns that have long been observed but poorly understood in metabarcoding data and illustrate patterns of detection and variability in observations from computer simulations and experimentally created metabarcoding samples. We use a simple process model to illustrate how several biological and sampling processes interact to generate the observed metabarcoding patterns. In Sections 2.2 and 3.2, we show how this understanding of the metabarcoding process leads to a statistical model with general utility for estimating DNA abundance from metabarcoding data. We test and validate our approach against simulated and empirical data. By providing a path to estimate DNA abundance directly from metabarcoding data, this work expands methods for making joint inferences from disparate eDNA data (e.g. qPCR and metabarcoding data or data from multiple metabarcoding markers) or from eDNA data and more traditional ecological survey data.

2 | MATERIALS AND METHODS

2.1 | Simulation and mock samples

Identifying a common unit shared among different potential observations (e.g. across biological or technical replicates, or across different qPCR or metabarcoding markers used to query the same sample) allows us to analyse those observations in a common analytical framework. The concentration of DNA for each taxon is the logical choice for this shared unit, whether in a single- or multi-taxon context.

Specifically, for a given eDNA sample that has been collected and processed (i.e. filtered, extracted, cleaned, and preserved), the DNA present has concentration F_{ij} for taxon i in sample j (see

Appendix S1 for a discussion of sampling processes for eDNA). Many eDNA samples focus on eukaryotic mitochondrial DNA (mtDNA), and so for exposition purposes we assume there is a single concentration of mtDNA for a given gene in each sample with units of copies μL^{-1} . The same principles apply to bacterial genomes or plasmids or nuclei, so long as they can be linked via an underlying concentration across markers (i.e. the approach would also work with some modification for taxa with multiple gene copies). For the remainder of this paper, we focus on developing methods that stem from this shared value. We do not connect estimates of DNA copies to the abundance or biomass of source organisms in this manuscript. In the discussion, we highlight how extensions to this work could advance investigations of extracting abundance and biomass estimates from field eDNA samples.

We use simulations (Sections 2.1.1 and 2.1.2) and experimentally created samples (Section 2.1.3) to understand how information about absolute DNA concentration is reflected in the read-counts produced by metabarcoding analyses. In particular, we focus on how patterns of detection and variability among replicate sequencing of individual samples provide information about absolute abundance.

All data and code are available via <https://doi.org/10.5281/zenodo.22905265> (Shelton et al., 2026).

2.1.1 | Simulation

We first describe a two-stage approach to simulating metabarcoding data that underlies our analysis. We provide additional description with functions for implementing simulations in Appendix S2. As above, we are interested in a 'sample' representing the DNA extract resulting from a single bottle of water (or equivalent, from air, soil, etc.) taken from the environment. For any analysis, an aliquot of volume $V \mu\text{L}$ is taken from F_{ij} and therefore a discrete number of DNA copies, G_{ijr} , will be included in subsequent replicate analysis r . This stochastic process can be described using a Poisson distribution with expectation $\lambda_{ij} = F_{ij}V$,

$$G_{ijr} \sim \text{Poisson}(\lambda_{ij}). \quad (1)$$

The realized copies are treated independently among taxa as is appropriate for sampling DNA molecules floating in solution. We assumed that the volume of an aliquot is small relative to the total sample volume. When λ_{ij} is small, there is non-negligible probability that exactly zero copies will be included in the aliquot; from the definition of the Poisson distribution, $p(G_{ijr} = 0 | \lambda_{ij}) = e^{-\lambda_{ij}}$. For a given aliquot even two taxa with identical λ may have different realized copies—one taxon may be entirely absent while another will have non-zero copies simply due to stochastic sampling. As long as V is constant among metabarcoding replicates (as is typical), λ_{ij} and F_{ij} contain equivalent information and we write notation in terms of λ hereafter.

Next, we used a log-linear model for PCR amplification (Gold et al., 2023; Kelly et al., 2019) which produces a number of amplicons A as a taxon- and marker-specific (I) function of per-PCR-cycle amplification efficiency, a_{ij} , and number of PCR cycles, N_{PCR} :

$$A_{ijlr} = G_{ijr} (1 + a_{ij})^{N_{PCR}}. \quad (2)$$

The amplification rate for a taxon may not be equivalent between markers; one may not amplify a taxon at all ($a_{i1} = 0$) while another might perfectly amplify ($a_{i2} = 1$; 1 corresponds to amplicons doubling each PCR cycle; $0 \leq a_{ij} \leq 1$). After generating a number of amplicons A_{ijlr} for each taxon, sample, marker, and aliquot, we simulated the process of DNA sequencing and assignment via a bioinformatics pipeline as a compositional process in which reads were drawn from a multinomial distribution with a sampling probability proportional to the abundance of their amplicons, $\pi_{ijlr} = \frac{A_{ijlr}}{\sum_i A_{ijlr}}$,

$$Y_{jlr} \sim \text{Multinomial}(N_{jlr}, \pi_{jlr}), \quad (3)$$

with bold indicating vectors. Y_{ijlr} is the observed number of reads and $N_{jlr} = \sum_i Y_{ijlr}$ is the total number of reads observed ('read depth'). In practice, because of the stochastic processes affecting each aliquot, the observed reads after sequencing will be overdispersed relative to a multinomial distribution and will vary as a function of λ ; we exploit this fact in the statistical approaches that follow (Appendix S1).

Amplification dynamics distinct from those provided in Equations (2) and (3) may also be reasonable. For example, a saturating model instead of the log-linear model for PCR in Equation (2) may be appropriate (Moinard et al., 2023). Similarly, different PCR protocols may have multiple amplification and subsampling steps that may warrant inclusion of additional deterministic or stochastic processes (Gold et al., 2023). These modifications do not materially change the qualitative patterns presented below; we have found that the exponential model is a good approximation for empirical observations.

Finally, the above simulation assumes that there is a single, shared concentration of DNA for that is detectable across multiple DNA markers. While this is a reasonable assumption for many commonly targeted marker regions (e.g. genes with single copies in the mitochondria), there are circumstances where DNA concentration will vary between markers. For example, when (i) one marker targets mitochondrial DNA and the other nuclear DNA, (ii) one marker targets a gene with a single copy in the genome and the other a gene with multiple copies, or (iii) the markers amplify regions with significantly different numbers of base pairs, which may have significantly different degradation rates and therefore different DNA concentrations. We discuss modifications that may address such differences in Section 1.3 of Appendix S1.

2.1.2 | Simulation variables

We used this simulation framework to explore how the observed reads change in response to three main variables: (i) the DNA expected copies per aliquot of each taxon (λ_{ij}), (ii) the total expected DNA copies per aliquot across all taxa ($\Lambda_j = \sum_i \lambda_{ij}$) and (iii) the read depth associated with each analysed aliquot (N_{jlr}). We simulated a single marker here. Throughout, the phrase 'total DNA copies'

applies to the total DNA copies amplifiable by a specific primer set, not a measure of all DNA present in a sample.

For each simulation, we specified the expected total DNA copies present across all taxa (Λ_j , ranging from 10 to 1 million copies by factors of 10) and the number of reads observed by the sequencer ($N_{jlr} = 1000, 3000, 10,000, 30,000$, or $100,000$). For each combination, we simulated 200 samples of 30 taxa with proportions specified by a Dirichlet distribution with a shared symmetric concentration parameter for each taxon, $\omega_i = 0.5$. This produced a wide range of proportions within a given sample that sum to one. We used a single shared amplification efficiency across all taxa ($a_i = 0.9$), 40 cycles of PCR ($N_{PCR} = 40$), and five replicate aliquots taken for each sample (see Appendix S2 for additional simulation scenarios with among-species amplification variability).

We first explored how probabilities of detection (defined as the probability of observing at least one read for a particular taxon from an aliquot) varied among simulation scenarios. While we were not interested in determining the true occurrence or absence of each taxon (see Section 1), the presence and frequency of zero observations can be highly informative about DNA abundance. Detection probabilities are often viewed as the fundamental output of metabarcoding analyses, but we focused on them as a first step in linking read-counts to DNA copies. A secondary goal was to understand whether and how the probability of detection changes in response to variability in both field and laboratory methods (Bylemans et al., 2025; Gold et al., 2023), and thereby assess the validity of some common assumptions of occupancy models applied to metabarcoding data.

Second, we explored how patterns of variability in read-counts among metabarcoding aliquots within samples were related to simulation variables. This variability among replicate aliquots is closely related to the number of DNA molecules in the sample, and therefore understanding it is vital to extracting DNA concentrations from metabarcoding data. We therefore investigate the relationship between the mean and standard deviation of the observed proportion across taxa and simulation scenarios, as a way to infer the total concentration of the sample.

2.1.3 | Mock sample construction and laboratory approaches

To test whether our simulation results adequately captured trends seen in empirical data, we constructed in the laboratory a series of multispecies eDNA samples each with known (precisely measured) concentrations of DNA molecules and sequenced these samples (Appendix S3). Artificial species assemblages with known inputs are typically referred to in eDNA texts as 'mock communities' (e.g. Shaffer et al., 2025; Shelton et al., 2023), but here we refer to these as "mock samples" to align the use of the word 'sample' between the simulation and estimation components in this paper. We used DNA extracts from 34 marine species (32 marine fish and two marine mammals) found in Washington state, USA, to make DNA mixtures

containing 15, 15 and 23 species respectively (referred to as 'mock 1', 'mock 2' and 'mock 3'). Within each mixture, we constructed two samples with identical taxa but distinct proportional DNA compositions (mocks 1A and 1B, 2A and 2B, 3A and 3B; Figure S3.1). The composition of mock samples reflects relatively diverse marine communities with overlapping, varying species compositions (see Shaffer et al., 2025). These samples were created at relatively high DNA abundance (Λ_j range: 66,530 to 159,146 expected total copies per aliquot; Appendix S3) and so we created a dilution series for each of the mock samples (undiluted, 1:10, 1:100 and 1:1000; Figure S3.1) resulting in 24 total mock samples. For all 24 samples, we drew triplicate replicate aliquots, PCR amplified using two mitochondrial markers from the 12S region (MiFish Universal [MFU], Miya et al., 2015; MarVer1 [MV1], Valsecchi et al., 2020), and independently sequenced each replicate (Appendix S3). This allowed us to analyse observations across a 1000-fold range of DNA abundance and compare empirical data with simulations and across six mock samples.

Results from simulation and mock samples were crucial for motivating subsequent statistical developments. Therefore, we suggest reading results in Section 3.1 before continuing to the methods in Section 2.2.

2.2 | Approaches for estimating DNA copies from metabarcoding

2.2.1 | Model description

Simulation and mock samples results suggested that the variability among replicate metabarcoding datasets contains information about the absolute abundance of DNA in the source sample (see Section 3.1). Here we build statistical tools to make use of this information. For brevity, we describe the final statistical models in the main text and present a full derivation of the model and statistical implementation details in Appendix S1. Directly estimating parameters using the model presented in the two-step simulation of Equations (1–3) is computationally infeasible. We develop faster and more general approaches to estimate DNA concentrations from metabarcoding data. We then apply these methods to assess their usefulness using both simulated and empirical results from the mock samples. We demonstrate how our approach leads to methods both for extracting information from replicated metabarcoding data and for combining information across multiple metabarcoding markers.

We begin with model and data assumptions. First, and most crucially, all models considered here assume that multiple metabarcoding observations are available for a given sample; most often, such data will take the form of replicate aliquots amplified via PCR and sequenced individually. This is necessary because, as shown in Figures 1 and 2, information about DNA concentration is found in the variability among aliquots, not in the point estimate of relative read abundance (Gloor et al., 2017; McLaren et al., 2019; Shelton et al., 2023). In the absence of replicate observations, there will be

little information about DNA abundance in metabarcoding observations (but see Diana et al., 2025). Second, these models assume that there are trivial amounts of contamination for the taxa of interest in the samples. Contamination from non-target taxa (e.g. human) may affect the precision of estimates, but can be identified and removed during data cleaning (see Griffard-Smith et al., 2025; McKnight et al., 2019); we view contamination as a laboratory challenge that can only be marginally ameliorated statistically.

The models presented here build on the work of Shelton et al. (2023) but explicitly include the stochastic sampling of molecules before amplification (Equation 1). This allows for zero observations—that is, zero reads from a taxon in a metabarcoding dataset—due either to the taxon's DNA being absent from an aliquot or from being relatively rare within the multi-taxon DNA mixture following PCR amplification. We derive a model for observed reads Y_{ijr} from a single metabarcoding marker but the results immediately generalize to multiple markers (Appendix S1). Because we develop the model using a single marker, we omit the l subscript in the following model description.

We begin by developing an expression for the probability of observing zero reads due to a taxon's absence from an aliquot. This is a function solely of the expected DNA copies in an aliquot, λ_{ij} . We create a latent variable, d_{ijr} , that indicates whether DNA copies are present ($d_{ijr} = 1$) or not ($d_{ijr} = 0$) in an aliquot. Following Equation (1), we model the probability of getting zero DNA copies in an aliquot using a Poisson distribution and so $p(d_{ijr} = 0 | \lambda_{ij}) = e^{-\lambda_{ij}}$ and therefore $p(d_{ijr} = 1 | \lambda_{ij}) = 1 - e^{-\lambda_{ij}}$. This accounts for one possible source of zero reads observed from a taxon—there were zero DNA copies in the PCR—but not the second case where there were non-zero DNA copies present in the PCR reaction, yet zero reads were observed after sequencing.

For this second case, we require a more complicated model to link the main parameter of interest, λ_{ij} , to observed metabarcoding reads. The simulations and mock samples reveal that the observed reads will be a function of λ_{ij} and the total expected copies across all taxa (Λ_j), with the relationship between the mean and standard deviation among replicate observations providing information about the absolute abundance of DNA in the sample (Figures 1 and 2). In addition, we know from previous work that variable amplification efficiency can bias estimates of proportional abundance (Kelly et al., 2017; Shelton et al., 2023), and thus, we need to quantify the amplification rate for each taxon relative to a reference taxon. We let α_i be the amplification rate relative to reference species R , and τ be the standard deviation of α_i among taxa. Furthermore, simulations demonstrate that amplification rates also affect the variability in observed reads; taxa with relatively high amplification rates have higher variability than taxa with relatively low amplification rates (see Figure S1.2). Together, these facts motivate a relatively complex statistical model that can incorporate all of the relevant factors.

To accommodate these factors, we developed an elaboration of the Shelton et al. (2023) model for compositional DNA sequence data. Conditional on $d_{ij} = 1$, we used a beta-binomial likelihood for the observed DNA read-counts. For $d_{ij} = 1$, the number of DNA

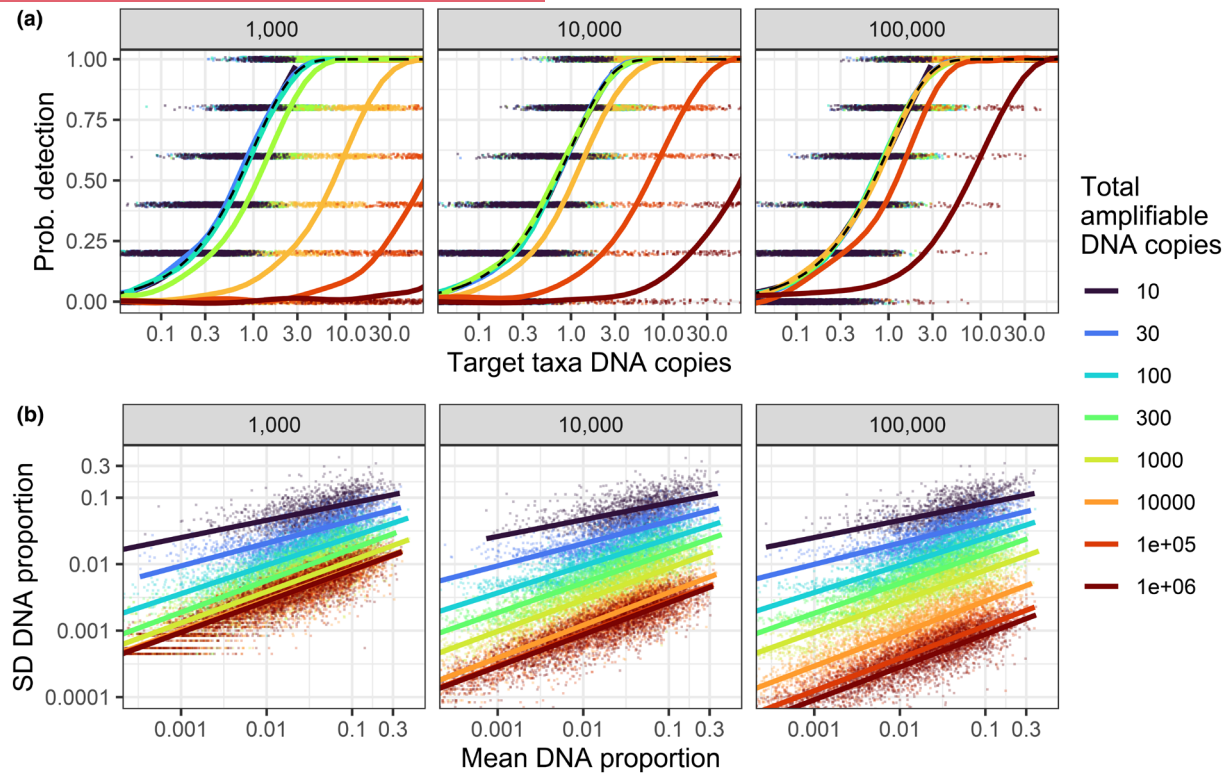


FIGURE 1 Results for simulated independent metabarcoding samples each with five replicates across a range of total amplifiable DNA concentrations. (a) Probability of detection as a function of DNA copies for different total (among taxa) number of DNA copies in a sample (colours) and different sequencing read depths (columns). Dashed line shows expected relationship between occurrence and DNA copies if detection is driven exclusively by the Poisson sampling of DNA molecules. Lines for each total DNA copy scenario are displayed but low total copy scenarios largely follow the dashed line and are difficult to differentiate. (b) Standard deviation of the estimated proportion among replicate simulations as a function of the mean proportion for each taxon. Here high total copy scenarios can have very similar relationships and lines may overlap. For all panels, individual points show individual taxa in a single simulation while lines show general additive model smooths (a) or linear model fits (b) across points.

copies in the aliquot is by definition non-zero, and so we defined λ_{kij} as the rate parameter of a zero-truncated Poisson distribution defined by λ_{ij} (Appendix S1). Following Equation (2) and Shelton et al. (2023), we use the known number of PCR cycles (N_{PCR}) and observed the read depth N_{jr} . Then, we can define the full model as

$$\begin{aligned}
 v_{ij} &= \log(\lambda_{kij}) - \log(\lambda_{KRj}) + N_{PCR}\alpha_i \\
 \alpha_i &\sim N(0, \tau) \\
 \pi_{ijr} &= \frac{d_{ijr}e^{v_{ij}}}{\sum_i d_{ijr}e^{v_{ij}}} \\
 \log(\phi_{ij}) &= \beta_0 + \log(\Lambda_j) + \beta_1\kappa_{ij} + \gamma_0e^{-\gamma_1(\log(\lambda_{kij}))} + \epsilon_{ij} \\
 \epsilon_{ij} &\sim N(0, \sigma) \\
 Y_{ijr} &\sim \text{Beta-Binomial}(\pi_{ijr}, \phi_{ij}, N_{jr})
 \end{aligned} \quad (4)$$

The parameter κ_{ij} is the centered log-ratio transform of amplification rates for taxa present in a particular sample and is a deterministic function of α (see Appendix S1). Note that in Equation (4), we define the beta-binomial distribution in terms of an expected proportion π_{ijr} and dispersion parameter ϕ_{ij} instead of the canonical α, β parameterization. One unusual component of the

model form for ϕ_{ij} is the non-linear component in Equation (4) (i.e. $\gamma_0e^{-\gamma_1(\log(\lambda_{kij}))}$). This allows ϕ_{ij} to increase (and therefore variability among replicate read-counts to decrease) as the conditionally positive expected DNA copies (λ_{kij}) decline towards 1 copy (see Figures S1.1 and S1.2).

We use the beta-binomial instead of the multivariate Dirichlet-multinomial distribution to allow for flexibility in accounting for both the presence or absence of each taxa from each aliquot. Additionally, it allows for variability in read-counts that are a function of taxon- and taxon-sample-specific traits (i.e. ϕ_{ij} is a function of λ_{ij} and κ_{ij} ; Equation 4). Such structure cannot be accounted for in the Dirichlet-multinomial likelihood. However, because we are modelling all taxa simultaneously, we maintain appropriate constraints on the estimated proportions for each sample (i.e. $\sum_i \pi_{ijr} = 1$).

As a final step, we needed to combine the information for the cases when zero or non-zero read-counts are observed conditioned on d . For notational simplicity, we introduce a vector of parameters that informs the variability among metabarcoding replicates ($\theta \in \{\beta_0, \beta_1, \gamma_0, \gamma_1, \sigma\}$). Then, Equation (4) is the likelihood for observed reads when DNA is present in an aliquot,

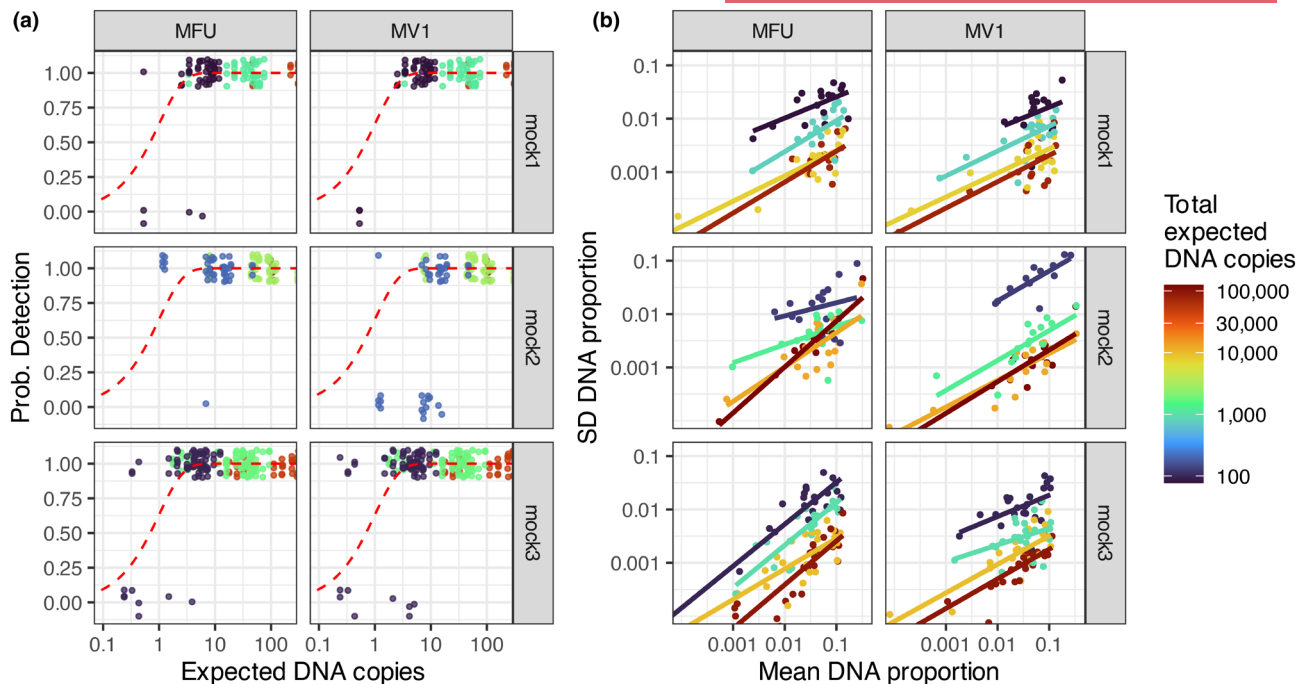


FIGURE 2 Results from three mock samples analysed with four dilutions and two metabarcoding markers (MiFishU [MFU] and MarVer1 [MV1]). Only 'A' samples for each mock sample are shown. (a) Detections (1) or absences (0) as a function of amplifiable DNA copies for individual replicates. Points have been jittered along the y-axis to reduce overlap. Dashed line shows expected relationship between detection and DNA copies if the relationship is driven exclusively by Poisson sampling of DNA molecules (Equation 1). (b) Standard deviation of the proportion of reads among replicate aliquots as a function of the mean estimated proportion for each taxon. Lines show least-squares linear model fits across taxa.

$p(Y_{ijr} = y | \lambda_j, \alpha, \tau, N_{jr}, d_{ijr} = 1, d_{-ijr}, \theta)$ with bolding indicating vectors across taxa and $-i$ indicating all taxa but taxon i . The full likelihood of the data given the parameters, $p(Y_{ijr} | \lambda_j, \alpha, \tau, N_{jr}, N_{PCR}, d_{ijr}, d_{-ijr}, \theta)$ is then

$$p(Y_{ijr} | \lambda_j, \alpha, \tau, N_{jr}, N_{PCR}, d_{ijr}, d_{-ijr}, \theta) = \begin{cases} (1 - e^{-\lambda_{ij}}) p(Y_{ijr} = y | \lambda_j, \alpha, \tau, N_{jr}, N_{PCR}, d_{ijr} = 1, d_{-ijr}, \theta) & \text{if } y > 0. \\ e^{-\lambda_{ij}} + (1 - e^{-\lambda_{ij}}) p(Y_{ijr} = y | \lambda_j, \alpha, \tau, N_{jr}, N_{PCR}, d_{ijr} = 1, d_{-ijr}, \theta) & \text{if } y = 0 \end{cases} \quad (5)$$

The two key innovations in this model are (i) incorporating the latent variable d into the softmax transformation (the expression for the expected proportion of reads, π , in Equation 4) so that each technical replicate can have a distinct value that reflects the DNA copies of all other taxa in each aliquot and (ii) linking the parameter ϕ to the individual species DNA copies (λ), the total DNA copies across all taxa (Λ), and the relative amplification for each taxon (α_i ; see Appendix S1). Furthermore, the parameters informing ϕ_{ij} provide a pathway for including additional, unmodeled heterogeneity in the amplification and sequencing processes that is not fully captured in Equation (2). We implement the above model in a Bayesian framework using R (R Core Team, 2025) and Nimble (de Valpine et al., 2017). Appendix S1 contains details of prior distributions (Table A.3), implementation and model variants including the generalization of the model for multiple markers and an approximation using Stan (Stan Development Team, 2025).

2.2.2 | Application to simulations

To test whether we could estimate DNA copies from replicated metabarcoding data, we simulated a 30-taxon sample over seven DNA samples that ranged from very low to high DNA copies (the seven samples span Λ_j of 10 to 10,000 copies per aliquot). We randomly assigned an amplification efficiency to each species (a_i drawn from a beta distribution with mean of 0.9 and SD of 0.01; such variation is consistent with existing estimates of amplification variability among taxa; Shelton et al. 2023). We simulated three independent aliquot replicates taken from each sample and metabarcoded, yielding the taxon-by-replicate table of read-counts common to metabarcoding sequencing datasets. For each replicate, we assigned a read depth randomly drawn from a uniform distribution between 50,000 and 100,000 reads. We then used the model from Section 2.2.1 to estimate the expected DNA copies for each taxon in each sample and the total expected DNA copies across all 30 taxa. We repeated the simulation-estimation exercise 100 times and reported the estimated number of DNA copies for each taxon in each sample and the total DNA copy estimate for each. For these simulations, we assumed amplification parameters were known a priori, as the purpose of this exercise is to understand the new estimation components, not to focus on estimating relative amplification efficiency (see Shelton et al., 2023).

We repeat a similar simulation-estimation exercise using an example with two markers and 50 species where the two markers

detect different but overlapping sets of species in [Appendix S1](#) ([Figures S1.9](#) and [S1.10](#)).

2.2.3 | Application to mock samples

We applied our model to triplicate observations from our 24 mock samples assayed with two markers (MFU and MV1) to test whether we could recover species copies from metabarcoding observations alone. Unlike the simulation models, here we estimated the relative amplification efficiency for each species within the model by splitting the 24 samples into two groups. We treated three samples (undiluted mock 1A, 2A and 3A) as having known DNA copy numbers for all taxa and used these three samples to inform the relative amplification efficiency among taxa. In the absence of including some samples with known starting DNA abundance, estimates of absolute or relative DNA abundance from metabarcoding are confounded with estimates of relative amplification efficiency (see Shelton et al., 2023). The remaining 21 samples were considered to be unknown and therefore served as an out-of-sample measure of model accuracy. The model estimated parameters using all 24 samples simultaneously. In [Appendix S1](#), a parallel analysis using three different samples to inform estimates of relative amplification efficiency is presented.

3 | RESULTS

3.1 | Simulation and mock samples

3.1.1 | Probability of detection

In the simulations, the probability of detecting a taxon depended on both the absolute number of DNA copies available (top row of [Figure 1](#)) as well as the relative abundance (proportion) of each taxon (see [Figure S2.1](#) for a general schematic representation of sampling and metabarcoding processes). For example, the blue line in [Figure 1](#) shows a total sample containing 100 copies while the orange line corresponds to 100,000 copies; thus, a taxon with ten expected copies in each sample would result in expected frequencies of 0.1 and 0.00001, respectively. The effect of relative abundance was strongly mediated by the read depth. When many reads were observed, the probability of detection was largely driven by the absolute, not relative, abundance except at very high total DNA copies (see also [Figure S2.5](#)). This pattern also occurred when there was variation in amplification rates among taxa ([Figure S2.2](#)), although there was additional noise from amplification variability. A general conclusion of these simulations is that non-detections (dropouts) were more likely to occur as a result of low absolute DNA abundance than from relative rarity as long as reasonable read depths are achieved (i.e. >10,000 total reads per sample in our simulation). Relative rarity becomes more important in the case of low read depths or in samples with high total DNA copies (i.e. when a taxon represents 0.001 or less of the proportion of DNA amplified from a sample; [Figure S2.5](#)). The results presented here were representative of patterns observed across the wide range of simulations ([Figures S2.2](#) and [S2.3](#)).

Empirical observations from mock samples closely mirrored the simulations. Among empirical observations, dropouts only occurred in samples with less than 100 total copies and only in taxa with less than about 10 copies in an aliquot ([Figure 2](#); [Figure S3.1](#)). Relative abundances for an individual taxon vary among mock samples, allowing for a limited comparison of the effect of relative abundance on detection. For example, Pacific sardine (*Sardinops sagax*) was present in all six mock samples and ranged among mock samples from a relative abundance of 0.004 (mock 3B) to 0.096 (mock 1A). Across the 1000-fold dilution series, individual mock samples have identical relative abundances. We observe no dropouts in the higher-concentration samples (undiluted and 1:10 dilutions) and very few dropouts at the 1:100 dilution ([Figures S3.3](#) and [S3.4](#)). This demonstrates that dropouts coincide with changes in absolute DNA abundance. We see no clear evidence of an effect of relative rarity on dropouts. For example, Pacific sardine was detected across the range of relative abundances present in all six mock communities in the undiluted, 1:10 and 1:100 replicates ([Figures S3.3](#) and [S3.4](#)). However, the effect of relative rarity on detection should become strongest at very low relative abundance (values of 0.001 or less; see [Figure S2.5](#)) and our mock samples do not contain very rare taxa. We targeted and achieved around 100,000 reads per aliquot (average within a sample ranged from about 28,000 to 216,000 reads with most around 100,000 reads; [Figure S3.2](#)) indicating broad similarity between the simulation and the experimental results. For such read depths, range of relative abundances, and total DNA concentrations, we would not expect relative rarity to play a large role in detections and indeed it did not. We saw very similar patterns across all mock samples ([Figures S3.5](#) and [S3.6](#)).

3.1.2 | Variability among replicates

We found that there were strong signals of expected total DNA copies in the relationship between mean and standard deviation of proportion of reads among technical replicates ([Figure 1](#)). Specifically, samples with few total expected DNA copies are highly variable among aliquots, while samples with large numbers of expected copies were more consistent and had lower variability among aliquots. As with detections, there was a clear effect of read depth, as samples where all replicates had only 1000 reads and high expected DNA copies had similar mean-standard deviation relationships. In contrast, with 100,000 reads, differentiating among larger total copy numbers was possible, as each had a distinct relationship between the mean and the standard deviation of observations. Overall, the simulation suggested that the mean-standard deviation relationship among metabarcoding replicates contained usable information about the total expected copies of amplifiable DNA.

Within the empirical mock samples, we saw the same ordering of mean-standard deviation relationships with DNA abundance as the simulations ([Figure 2b](#)). Samples with few DNA copies were highly variable while samples with many DNA copies were less variable. Similar patterns occurred in both markers across all samples ([Figures S3.7](#) and [S3.8](#)). As expected, the empirical patterns were

somewhat noisier than the simulation (Figure 2; least-squares fit lines for different total copy numbers crossed for some samples). This was likely because the empirical data represent a single sample with three aliquots and varying read depth while the simulation represented the pattern averaged across hundreds of simulated samples. These results indicate that the inclusion of additional stochastic processes to the process model may be warranted.

Together, the results suggested that the concentration of DNA in a sample can be assessed by measuring the variability of metabarcoding reads among replicate aliquots. While such patterns of metabarcoding variability have been qualitatively described in a range of studies (Gold et al., 2023; Leray & Knowlton, 2017; Mata et al., 2019), we devote the remainder of the paper to developing statistical estimators for DNA abundance within a sample using the variability of metabarcoding reads.

3.2 | Estimating DNA copies

3.2.1 | Simulations

For the simulated scenario, concentrations were estimable from replicated metabarcoding observations alone (Figure 3). Figure 3a shows that estimates for each taxon broadly reflect the true expected copies, with estimates closely following the 1:1 line. Unsurprisingly, low copy taxa were uncertain, with estimates particularly uncertain below one copy. This reflects the fact that with only three aliquots, rare taxa will sometimes be entirely absent from all aliquots while others will be present in multiple aliquots due to stochastic sampling. Differentiating among taxa having fewer than approximately three copies was difficult. However, it was easy to distinguish these rare taxa from higher-concentration taxa (i.e. differentiating a one

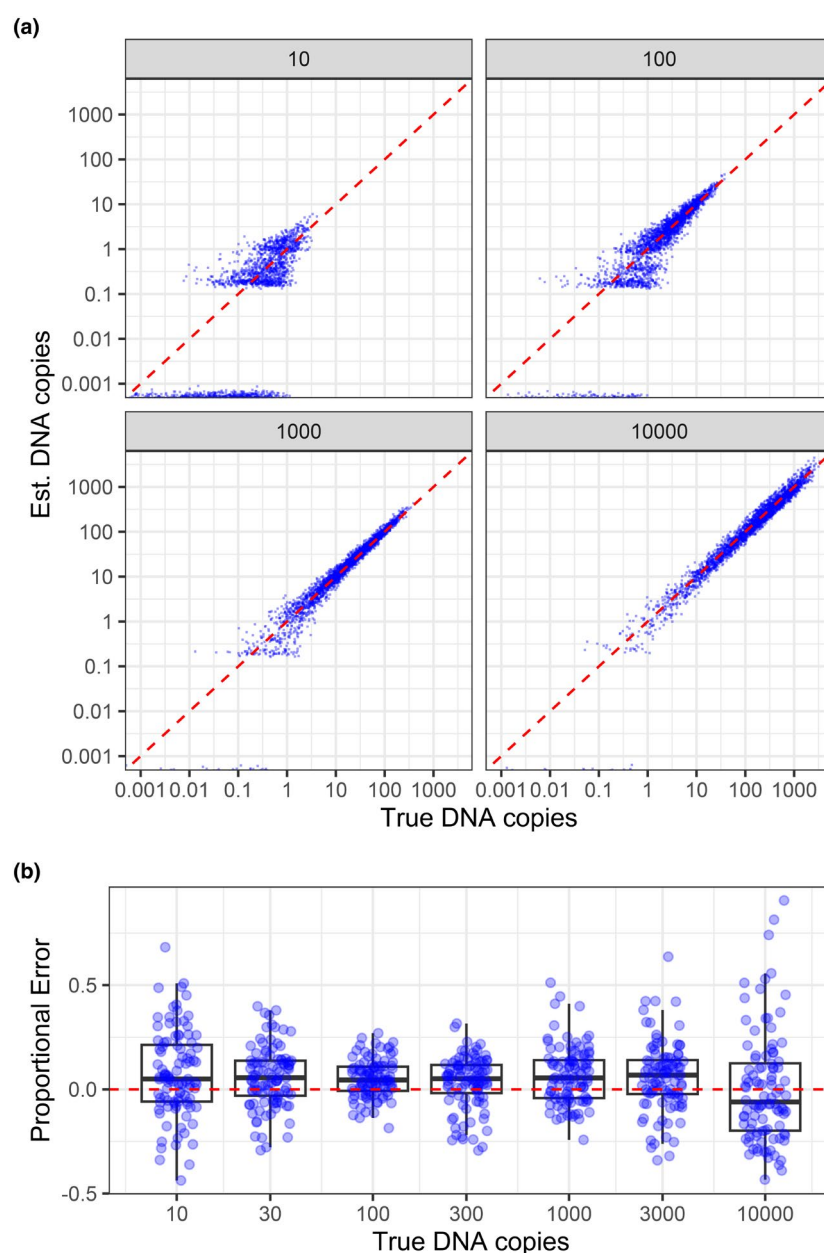


FIGURE 3 Estimation of simulated metabarcoding data for 100 replicate simulations of a 30 taxa sample assayed across seven sites with a single marker and three replicate aliquots. (a) Expected DNA copies for taxa across all simulation replicates for four of the seven sites (with 10, 100, 1000, or 10,000 total expected DNA copies as facets). Points along the bottom of the y-axis (estimated DNA copies < 0.001) reflect taxa that were not observed in any simulated metabarcoding observations despite being present in the simulated sample at low copy number. (b) Proportional error of posterior mean estimates for total expected DNA copies across all taxa.

copy taxon from a 100 copy taxon was straightforward). We saw a modest positive bias in the model estimate (median proportional error of the posterior mean estimate of about 0.05 across 100 simulations; Figure 3).

3.2.2 | Mock samples

Our model provided accurate estimates of both total DNA and individual taxon DNA copies (Figure 4). These estimates integrated

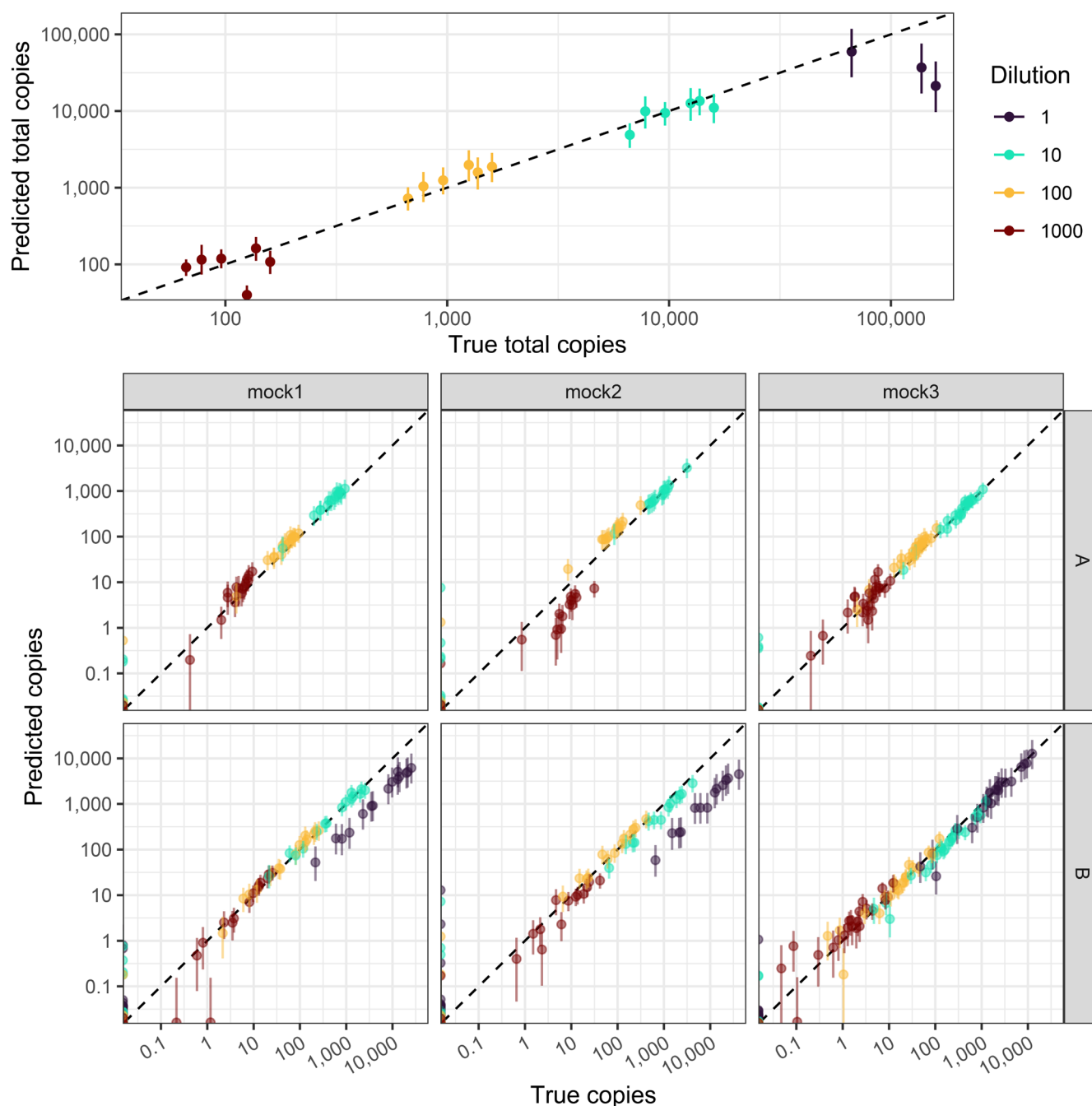


FIGURE 4 Cross-validation of model estimates (predicted expected DNA copies) against true expected DNA copies from mock samples. Data from both markers (MFU and MV1) inform estimates. Top: Estimated total expected DNA concentration for each of six samples across four dilutions (undiluted, 1:10, 1:100, and 1:1000). Note that three samples in the undiluted category are not displayed as they were used to inform amplification efficiencies during model estimation. Posterior mean and 90% intervals are shown. Perfect predictions would fall along the dashed, 1:1 line. Bottom: Estimated expected DNA copies for each taxon with true expected copies for each of the six mock samples and four dilutions (1:1 line and mean and 90% CI shown). Points along the bottom of the x- and y-axis (estimated or predicted expected DNA copies <0.002) reflect taxa that were not observed in metabarcoding observations despite being present in the simulated sample at low copy number or observed at very low levels in a mock sample where a taxon was nominally absent.

information from both the MFU and MV1 data and thus represent the consensus estimates for DNA from all available data from two metabarcoding markers. For both total expected DNA copies and expected DNA copies for each taxon, estimates closely track the inputs (Figure 4). For total expected copies, the 90% credible interval includes the true estimate in 16 of 21 samples; three of the 1000-fold dilutions were slightly off the 1:1 line and two of the undiluted samples were underestimated. The underestimated samples were the undiluted samples from mock 2 and mock 3 (see also Figure 4). Despite extensive investigation, we do not have a good explanation for why these particular estimates appear to be somewhat low, though the undiluted samples were sequenced on a different plate than the remaining observations (see Appendix S1). Overall, the estimation remained quite good. The model provided similar fit when different samples were used for estimating relative amplification efficiency (Figures S1.11–S1.13). These results revealed an ability to combine metabarcoding data from independent markers and recover DNA concentrations from replicated metabarcoding experiments while accounting for variable amplification efficiency among taxa.

4 | DISCUSSION

We have created a unified framework for (i) deriving quantitative information from replicate metabarcoding samples, and (ii) meaningfully combining results from two or more metabarcoding loci. The key to the statistical model is recognizing that replicated metabarcoding data—either from sequencing replicates from a single locus or from multiple loci—derives from a shared quantity, the concentration of amplifiable copies of DNA in the environmental sample. By modelling the processes of sampling and amplifying DNA molecules via PCR, we illustrate the mechanisms leading to metabarcoding observations, and thereby link observations to underlying processes. We conducted tests using experimentally created mock samples and found that our statistical methods performed well on empirical data. These approaches represent a substantial advance over currently available approaches.

This work expands the utility of metabarcoding data for ecologists and similar sequencing-derived data used by microbiologists, epidemiologists, and others by estimating amplifiable DNA copies, a biologically meaningful measure. Because metabarcoding can measure the DNA of hundreds of species simultaneously (depending upon the marker and samples), the statistical approaches presented greatly expand the potential impact of metabarcoding as a useful ecological tool. Moreover, this kind of process-based modelling may help to move the field away from ad hoc data transformations and occupancy models in the analysis of metabarcoding data. By showing how observations from multiple markers can be joined in a single framework, we can move from Venn diagrams toward unified, quantitative assessments of the number of copies of DNA present in biological samples.

Here we have focused on the mechanisms and analytical processes that lead to metabarcoding observations within individual biological samples. A separate and important question that we do not

address is the relationship among source organisms, DNA present in the environment, and DNA collected in biological samples (e.g. bottles of water). These connections are particularly important for large-bodied eukaryotic species that may not be present when their DNA was collected (i.e. not bacteria, phytoplankton or other small organisms that are directly captured in a biological sample). The relationship between DNA sources and collected samples is likely to depend strongly on the species and environmental contexts of interest; DNA concentrations arise from the combined processes of shedding, transport, and decay (Andruszkiewicz et al., 2019; Barnes & Turner, 2016; Brasseale et al., 2025). In all cases, a key step in connecting patterns of DNA in the environment to organismal abundance and distribution is deriving estimates of taxon-specific DNA concentration in individual samples. We have shown how to derive such estimates from metabarcoding observations alone. We provide a general observation model for metabarcoding data that can be connected to almost any underlying spatial or temporal model of DNA concentration.

Our framework offers an explanation for a range of common patterns in ecological eDNA studies that lead to confusion among users. For example, metabarcoding data from aliquots taken from a single sample can sometimes look very different from one another, both in terms of taxa detected and their relative abundances. Our simulations and empirical results suggest that these patterns are likely a consequence of samples with low DNA concentrations of the target organisms—and hence, concomitantly high variability—consistent with discussions elsewhere in the literature (see Leray & Knowlton, 2017; Mata et al., 2019). Rather than indicating a failure of molecular data as a valid source of information, these kinds of differences among replicate aliquots are instead the expected behaviour. We have shown how to make use of this variability to provide estimates of DNA concentration that were otherwise unavailable.

As made clear by our analysis of mock samples, metabarcoding data remains challenging. Two potential limitations stand out. First, the spectre of amplification variability among taxa remains a major concern for the interpretation of metabarcoding data. Developing experimental mixtures of DNA from multiple taxa can clearly be used to estimate relative amplification rates (see also Shaffer et al., 2026; Shelton et al., 2023). Although estimating relative amplification from the attributes of the primer and template shows significant promise (e.g. Shaffer et al., 2025), quantifying relative amplification rates for large communities remains a difficult component of metabarcoding. Second, it is possible that PCR inhibition or stochastic lab errors could produce patterns of high variability that could lead to errant estimates of DNA concentrations. We speculate that modest inhibition may have led to an underestimate of DNA concentration in our highest concentration mock samples; high DNA concentrations are known to be associated with PCR inhibition (Opel et al., 2010). However, many field samples will contain substantially less DNA than the mock samples we used here.

We also note that this framework could be easily extended to incorporate additional information from individual taxa within a sample. For example, a spike-in—in which a taxon of known concentration and amplification efficiency is added to each sample (Deagle

et al., 2018)—provides a single taxon with precisely estimated copies against which the abundance of other taxa could be calibrated. However, determining an appropriate amount of DNA added during a spike-in to each sample remains a challenge (Deagle et al., 2018) and, like all other DNA observed during metabarcoding, the relative amplification efficiency of any spike-in needs to be known or estimable for proper inference. Measurements of DNA concentration for a single taxon in each sample via qPCR or ddPCR could similarly help calibrate the abundance of remaining taxa. In practice, a single taxon may not be present in all samples of interest and therefore estimates of DNA copies from multiple taxa may be helpful for calibration.

We are confident in our results but our analyses are limited to simulated data and an application to mock samples created from 34 marine species from the NE Pacific Ocean. The methods presented deserve further testing and application to a variety of markers and systems. We motivated our work based on the PCR, sequencing, and bioinformatic protocols that we use but future testing should investigate the model's efficacy on data derived from other markers and laboratory protocols. Similarly, we expect the statistical model can be improved. Our estimators have a mild estimation bias that further investigation may be able to resolve. In particular, we note that the form for the dispersion parameter, ϕ , is complex; alternate forms for ϕ are worth exploration. In the appendices, we provide simulation tools and statistical estimation tools that can form the basis for advancing this work.

Finally, there are clear avenues for connections with other common ecological data and statistical approaches. Most obviously, DNA concentrations can also be measured directly by single taxon qPCR and ddPCR, meaning that multiple kinds of environmental genetic data can now be linked together (e.g. Guri et al., 2024). For focused studies of individual or a few species, qPCR or ddPCR methods will still be preferred to metabarcoding approaches due to cost and ease of analysis, but for larger ecological communities of interest, the statistical approach provided here provides an intriguing path forward. Recent decades have seen a rapid expansion in spatio-temporal statistics for ecological survey data (Cressie & Wikle, 2011) including some eDNA applications (Diana et al., 2025; Shelton et al., 2022). In particular, it is possible to envision building quantitative spatial or spatio-temporal models of DNA concentration for tens of taxa simultaneously and linking such information to other available observation types (e.g. visual or acoustic survey data). The resulting joint estimates of abundance and distribution may expand the value of expensive field surveys—for example, allowing natural-resources agencies to optimize the yield of information about species of interest from a mix of molecular and more traditional sampling methods.

AUTHOR CONTRIBUTIONS

Andrew Olaf Shelton conceptualized the study and developed the statistical methods with input from Shannon M. Kachel, Ryan P. Kelly, Eiren K. Jacobson and Len Thomas. Andrew Olaf Shelton and Shannon M. Kachel wrote the code to implement the methods. Megan R. Shaffer and Pedro F. P. Brandão-Dias collected, processed and organized the data. All authors contributed critically to drafts of the manuscript and gave final approval for publication.

ACKNOWLEDGEMENTS

We thank E.A. Allan, G. Guri, A. Ip, O. Liu, K. Nichols, K. Parsons, and O. Scott for discussions during the development of this manuscript. D. Baetscher, K. Nichols, J. Samhoury, and two anonymous reviewers provided helpful comments on manuscript drafts. This material is based upon research supported by the Office of Naval Research under Award Number (N00014-22-1-2719). Additional support was provided by the NOAA Fisheries Omics Strategic Initiative.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

PEER REVIEW

The peer review history for this article is available at <https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210x.70427>.

DATA AVAILABILITY STATEMENT

Data, simulation code and estimation code used in the paper are archived at Zenodo: <https://doi.org/10.5281/zenodo.22905265> (Shelton et al., 2026) and at the GitHub repository: https://github.com/nwfs-cb/eDNA_replicated_metabarcoding.

ORCID

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

Shannon M. Kachel  <https://orcid.org/0000-0003-3694-6634>

Pedro F. P. Brandão-Dias  <https://orcid.org/0000-0001-8583-0617>

Eiren K. Jacobson  <https://orcid.org/0000-0003-0147-8367>

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

REFERENCES

- Allan, E. A., Kelly, R. P., D'Agnese, E. R., Garber-Yonts, M. N., Shaffer, M. R., Gold, Z. J., & Shelton, A. O. (2023). Quantifying impacts of an environmental intervention using environmental DNA. *Ecological Applications*, 33, e2914.
- Andruszkiewicz, E. A., Koseff, J. R., Fringer, O. B., Ouellette, N. T., Lowe, A. B., Edwards, C. A., & Boehm, A. B. (2019). Modeling environmental DNA transport in the coastal ocean using Lagrangian particle tracking. *Frontiers in Marine Science*, 6, 477.
- Barnes, M. A., & Turner, C. R. (2016). The ecology of environmental DNA and implications for conservation genetics. *Conservation Genetics*, 17, 1–17.
- Beng, K. C., & Corlett, R. T. (2020). Applications of environmental DNA (eDNA) in ecology and conservation: Opportunities, challenges and prospects. *Biodiversity and Conservation*, 29, 2089–2121.
- Blasko, T., & Phelps, M. (2024). Streamside detection of Chinook Salmon (*Oncorhynchus tshawytscha*) environmental DNA with CRISPR technology. *Environmental DNA*, 6, e549.
- Bohmann, K., & Lynggaard, C. (2023). Transforming terrestrial biodiversity surveys using airborne eDNA. *Trends in Ecology & Evolution*, 38, 119–121.
- Brasseale, E., Adams, N., Allan, E. A., Jacobson, E. K., Kelly, R. P., Liu, O. R., Moore, S., Shaffer, M., Xiong, J., & Parsons, K. (2025). Marine eDNA production and loss mechanisms. *Journal of Geophysical Research: Oceans*, 130, e2024JC021643.
- Bylemans, J., Everts, T., Brys, R., & Duncan, R. P. (2025). From anarchy to clarity, data pre-processing and statistical choices influence

- quantitative environmental DNA (eDNA) analyses. *Methods in Ecology and Evolution*, 16, 1322–1333.
- Cressie, N., & Wikle, C. K. (2011). *Statistics for spatio-temporal data*. John Wiley & Sons.
- Crowley, S. E., Bentzen, P., Kess, T., Duffy, S. J., Messmer, A. M., Watson, B., Dempson, J. B., Keefe, D. G., Perry, R. C., Marquis, B., Hajibabaei, M., Fahner, N., Berghuis, L., Hobrecker, K., & Bradbury, I. R. (2024). eDNA metabarcoding reveals riverine fish community structure and climate associations in northeastern Canada. *Environmental DNA*, 6, e517.
- de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., & Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, 26, 403–413.
- Deagle, B. E., Clarke, L. J., Kitchenner, J. A., Polanowski, A. M., & Davidson, A. T. (2018). Genetic monitoring of open ocean biodiversity: An evaluation of DNA metabarcoding for processing continuous plankton recorder samples. *Molecular Ecology Resources*, 18, 391–406.
- Diana, A., Matechou, E., Griffin, J., Yu, D. W., Luo, M., Tosa, M., Bush, A., & Griffiths, R. A. (2025). eDNAPlus: A unifying modeling framework for DNA-based biodiversity monitoring. *Journal of the American Statistical Association*, 120, 120–134. <https://doi.org/10.1080/01621459.2024.2412362>
- Djurhuus, A., Closek, C. J., Kelly, R. P., Pitz, K. J., Michisaki, R. P., Starks, H. A., Walz, K. R., Andruszkiewicz, E. A., Olesin, E., Hubbard, K., Montes, E., Otis, D., Muller-Karger, F. E., Chavez, F. P., Boehm, A. B., & Breitbart, M. (2020). Environmental DNA reveals seasonal shifts and potential interactions in a marine community. *Nature Communications*, 11, 254.
- Dougherty, M. M., Larson, E. R., Renshaw, M. A., Gantz, C. A., Egan, S. P., Erickson, D. M., & Lodge, D. M. (2016). Environmental DNA (eDNA) detects the invasive rusty crayfish *Orconectes rusticus* at low abundances. *Journal of Applied Ecology*, 53, 722–732.
- Duprey, J., Gallego, R., Klinger, T., & Kelly, R. P. (2023). Environmental DNA reveals patterns of biological invasion in an inland sea. *PLoS One*, 18, e0281525.
- Eastwood, N., Kissane, S., Campbell, L., Briscoe, A. G., Egeter, B., & Orsini, L. (2024). Single metabarcoding multiplex captures community-level freshwater biodiversity and beyond. *Environmental DNA*, 6, e515.
- Egan, S. P., Grey, E., Olds, B., Feder, J. L., Ruggiero, S. T., Tanner, C. E., & Lodge, D. M. (2015). Rapid molecular detection of invasive species in ballast and harbor water by integrating environmental DNA and light transmission spectroscopy. *Environmental Science & Technology*, 49, 4113–4121.
- Espinosa Prieto, A., Hardion, L., Debortoli, N., & Beisel, J.-N. (2024). Finding the perfect pairs: A matchmaking of plant markers and primers for multi-marker eDNA metabarcoding. *Molecular Ecology Resources*, 24, e13937.
- Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., Gielly, L., Lopes, C. M., Boyer, F., Pompanon, F., Rayé, G., & Taberlet, P. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, 15, 543–556.
- Fields, B., Moeskjær, S., Friman, V.-P., Andersen, S. U., & Young, J. P. W. (2021). MAUI-seq: Metabarcoding using amplicons with unique molecular identifiers to improve error correction. *Molecular Ecology Resources*, 21, 703–720. <https://doi.org/10.1111/1755-0998.13294>
- Fukaya, K., & Hasebe, Y. (2025). occumb: An R package for site occupancy modeling of eDNA metabarcoding data. *Population Ecology*, 67, 285–297. <https://doi.org/10.1002/1438-390X.12219>
- Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology*, 8, 2224.
- Gold, Z., Shelton, A. O., Casendino, H. R., Duprey, J., Gallego, R., Van Cise, A., Fisher, M., Jensen, A. J., D'Agnese, E., Andruszkiewicz Allan, E., Ramón-Laca, A., Garber-Yonts, M., Labare, M., Parsons, K. M., & Kelly, R. P. (2023). Signal and noise in metabarcoding data. *PLoS One*, 18, 1–21.
- Griffard-Smith, R., Schueddig, E., Mahoney, D. E., Chalise, P., Koestler, D. C., & Pei, D. (2025). micRoclean: An R package for decontaminating low-biomass 16S-rRNA microbiome data. *Frontiers in Bioinformatics*, 5, 1556361.
- Griffin, J. E., Matechou, E., Buxton, A. S., Bormpoudakis, D., & Griffiths, R. A. (2020). Modelling environmental DNA data; Bayesian variable selection accounting for false positive and false negative errors. *Journal of the Royal Statistical Society: Series C: Applied Statistics*, 69, 377–392.
- Guri, G., Shelton, A. O., Kelly, R. P., Yoccoz, N., Johansen, T., Præbel, K., Hanebrekke, T., Ray, J. L., Fall, J., & Westgaard, J. I. (2024). Predicting trawl catches using environmental DNA. *ICES Journal of Marine Science*, 81, 1536–1548.
- Harper, L. R., Lawson Handley, L., Hahn, C., Boonham, N., Rees, H. C., Gough, K. C., Lewis, E., Adams, I. P., Brotherton, P., Phillips, S., & Hänfling, B. (2018). Needle in a haystack? A comparison of eDNA metabarcoding and targeted qPCR for detection of the great crested newt (*Triturus cristatus*). *Ecology and Evolution*, 8, 6330–6341.
- Iacaruso, N. J., Reves, O. P., Merkelz, S. J., Waldrep, C. L., & Davis, M. A. (2025). A systematic review evaluating the performance of eDNA methods relative to conventional methods for biodiversity monitoring. *Ecography*, 2026, e07952.
- Johnsen, S. I., Strand, D. A., Rusch, J. C., & Vrålstad, T. (2020). Environmental DNA (eDNA) monitoring of noble crayfish *Astacus astacus* in lentic environments offers reliable presence-absence surveillance—but fails to predict population density. *Frontiers in Environmental Science*, 8, 612253.
- Keller, A. G., Grason, E. W., McDonald, P. S., Ramón-Laca, A., & Kelly, R. P. (2022). Tracking an invasion front with environmental DNA. *Ecological Applications*, 32, e2561.
- Kelly, R. P., Closek, C. J., O'Donnell, J. L., Kralj, J. E., Shelton, A. O., & Samhuri, J. F. (2017). Genetic and manual survey methods yield different and complementary views of an ecosystem. *Frontiers in Marine Science*, 3, 283.
- Kelly, R. P., Shelton, A. O., & Gallego, R. (2019). Understanding PCR processes to draw meaningful conclusions from environmental DNA studies. *Scientific Reports*, 9, 12133.
- Laporte, M., Reny-Nolin, E., Chouinard, V., Hernandez, C., Normandeau, E., Bougas, B., Côté, C., Behmel, S., & Bernatchez, L. (2021). Proper environmental DNA metabarcoding data transformation reveals temporal stability of fish communities in a dendritic river system. *Environmental DNA*, 3, 1007–1022.
- Ledger, K. J., Everett, M., Nichols, K. M., Larson, W. A., & Baetscher, D. S. (2025). Development and validation of an eDNA primer set for rockfishes in the Alaska Current provides a new tool for fisheries management. *Environmental DNA*, 7, e70176.
- Leray, M., & Knowlton, N. (2017). Random sampling causes the low reproducibility of rare eukaryotic OTUs in illumina COI metabarcoding. *PeerJ*, 5, e3006.
- Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., Boehm, J. T., & Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, 10, 34.
- Mata, V. A., Rebelo, H., Amorim, F., McCracken, G. F., Jarman, S., & Beja, P. (2019). How much is enough? Effects of technical and biological replication on metabarcoding dietary analysis. *Molecular Ecology*, 28, 165–175.
- McClenaghan, B., Compson, Z. G., & Hajibabaei, M. (2020). Validating metabarcoding-based biodiversity assessments with multi-species occupancy models: A case study using coastal marine eDNA. *PLoS One*, 15, e0224119.
- McKnight, D. T., Huerlimann, R., Bower, D. S., Schwarzkopf, L., Alford, R. A., & Zenger, K. R. (2019). microDecon: A highly accurate

- read-subtraction tool for the post-sequencing removal of contamination in metabarcoding studies. *Environmental DNA*, 1, 14–25.
- McLaren, M. R., Willis, A. D., & Callahan, B. J. (2019). Consistent and correctable bias in metagenomic sequencing experiments. *eLife*, 8, e46923.
- Miller, D. A. W., Bailey, L. L., Grant, E. H. C., McClintock, B. T., Weir, L. A., & Simons, T. R. (2015). Performance of species occurrence estimators when basic assumptions are not met: A test using field data where true occupancy status is known. *Methods in Ecology and Evolution*, 6, 557–565.
- Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J. Y., Sato, K., Minamoto, T., Yamamoto, S., Yamanaka, H., Araki, H., Kondoh, M., & Iwasaki, W. (2015). MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: Detection of more than 230 subtropical marine species. *Royal Society Open Science*, 2, 150088.
- Moinard, S., Piau, D., Laporte, F., Rioux, D., Taberlet, P., Gonindard-Melodelima, C., & Coissac, E. (2023). Towards quantitative DNA metabarcoding: A method to overcome PCR amplification bias. *bioRxiv: The Preprint Server for Biology*. <https://www.biorxiv.org/content/early/2023/11/28/2023.10.03.560640>
- Mori, H., Maruyama, F., Kato, H., Toyoda, A., Dozono, A., Ohtsubo, Y., Nagata, Y., Fujiyama, A., Tsuda, M., & Kurokawa, K. (2014). Design and experimental application of a novel non-degenerate universal primer set that amplifies prokaryotic 16S rRNA genes with a low possibility to amplify eukaryotic rRNA genes. *DNA Research*, 21, 217–227.
- Nevers, M. B., Byappanahalli, M. N., Morris, C. C., Shively, D., Przybyla-Kelly, K., Spoljaric, A. M., Dickey, J., & Roseman, E. F. (2018). Environmental DNA (eDNA): A tool for quantifying the abundant but elusive round goby (*Neogobius melanostomus*). *PLoS One*, 13, e0191720.
- Opel, K. L., Chung, D., & McCord, B. R. (2010). A study of PCR inhibition mechanisms using real time PCR. *Journal of Forensic Sciences*, 55, 25–33. <https://doi.org/10.1111/j.1556-4029.2009.01245.x>
- Pont, D. (2024). Predicting downstream transport distance of fish eDNA in lotic environments. *Molecular Ecology Resources*, 24, e13934.
- R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Royle, J. A. (2006). Site occupancy models with heterogeneous detection probabilities. *Biometrics*, 62, 97–102.
- Sepulveda, A. J., Dumoulin, C. E., Blanchette, D. L., McPhedran, J., Holme, C., Whalen, N., Hunter, M. E., Merkes, C. M., Richter, C. A., Neilson, M. E., Daniel, W. M., Jones, D. N., & Smith, D. R. (2023). When are environmental DNA early detections of invasive species actionable? *Journal of Environmental Management*, 343, 118216.
- Shaffer, M. R., Allan, E. A., Kelly, R. P., Adams, N. G., Moore, S. K., Brasseale, E., Brandão-Dias, P. F. P., Scott, O. M., Shelton, A. O., & Parsons, K. M. (2026). Non-target detections change the interpretation of environmental DNA communities. *Metabarcoding and Metagenomics*, 2026. <https://doi.org/10.3897/mbmg.10.174467>
- Shaffer, M. R., Andruszkiewicz Allan, E., Van Cise, A. M., Parsons, K. M., Shelton, A. O., & Kelly, R. P. (2025). Observation bias in metabarcoding. *Molecular Ecology Resources*, 25, e14119.
- Shelton, A. O., Gold, Z. J., Jensen, A. J., D'Agnes, E., Andruszkiewicz Allan, E., Van Cise, A., Gallego, R., Ramón-Laca, A., Garber-Yonts, M., Parsons, K., & Kelly, R. P. (2023). Toward quantitative metabarcoding. *Ecology*, 104, e3906.
- Shelton, A. O., Kachel, S. M., Shaffer, M. R., Brandão-Dias, P. F. P., Jacobson, E. K., Thomas, L., & Kelly, R. P. (2026). Data and code for "estimating absolute DNA abundance from replicated metabarcoding". *Zenodo*. <https://doi.org/10.5281/zenodo.22905265>
- Shelton, A. O., Kelly, R. P., O'Donnell, J. L., Park, L., Schwenke, P., Greene, C., Henderson, R. A., & Beamer, E. M. (2019). Environmental DNA provides quantitative estimates of a threatened salmon species. *Biological Conservation*, 237, 383–391.
- Shelton, A. O., Ramón-Laca, A., Wells, A., Clemons, J., Chu, D., Feist, B. E., Kelly, R. P., Parker-Stetter, S. L., Thomas, R., Nichols, K. M., & Park, L. (2022). Environmental DNA provides quantitative estimates of pacific hake abundance and distribution in the open ocean. *Proceedings of the Royal Society B: Biological Sciences*, 289, 20212613.
- Stan Development Team. (2025). *RStan: The R interface to Stan*. <https://mc-stan.org/>
- Suren, A. M., Burdon, F. J., & Wilkinson, S. P. (2024). eDNA is a useful environmental monitoring tool for assessing stream ecological health. *Environmental DNA*, 6, e596.
- Tetzlaff, S. J., Katz, A. D., Johnson, M. D., & Sperry, J. H. (2024). Community ecology in a bottle: Leveraging eDNA metabarcoding data to predict occupancy of co-occurring species. *Environmental DNA*, 6, e579.
- Thomas, A. C., Tank, S., Nguyen, P. L., Ponce, J., Sinnesael, M., & Goldberg, C. S. (2020). A system for rapid eDNA detection of aquatic invasive species. *Environmental DNA*, 2, 261–270.
- Tsuri, K., Ikeda, S., Hirohara, T., Shimada, Y., Minamoto, T., & Yamanaka, H. (2021). Messenger RNA typing of environmental RNA (eRNA): A case study on zebrafish tank water with perspectives for the future development of eRNA analysis on aquatic vertebrates. *Environmental DNA*, 3, 14–21.
- Ushio, M., Murakami, H., Masuda, R., Sado, T., Miya, M., Sakurai, S., Yamanaka, H., Minamoto, T., & Kondoh, M. (2018). Quantitative monitoring of multispecies fish environmental DNA using high-throughput sequencing. *Metabarcoding and Metagenomics*, 2, e23297. <https://doi.org/10.3897/mbmg.2.23297>
- Valsecchi, E., Bylemans, J., Goodman, S. J., Lombardi, R., Carr, I., Castellano, L., Galimberti, A., & Galli, P. (2020). Novel universal primers for metabarcoding environmental DNA surveys of marine mammals and other marine vertebrates. *Environmental DNA*, 2, 460–476.
- Wilkinson, S. P., Gault, A. A., Welsh, S. A., Smith, J. P., David, B. O., Hicks, A. S., Fake, D. R., Suren, A. M., Shaffer, M. R., Jarman, S. N., & Bunce, M. (2024). TICI: A taxon-independent community index for eDNA-based ecological health assessment. *PeerJ*, 12, e16963.
- Williams, M.-A., O'Grady, J., Ball, B., Carlsson, J., de Eyto, E., McGinnity, P., Jennings, E., Regan, F., & Parle-McDermott, A. (2019). The application of CRISPR-cas for single species identification from environmental DNA. *Molecular Ecology Resources*, 19, 1106–1114.
- Willis, A. D., & Martin, B. D. (2022). Estimating diversity in networked ecological communities. *Biostatistics*, 23, 207–222.

SUPPORTING INFORMATION

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

Appendix S1. Statistical model derivation, testing, and validation.

Appendix S2. Description of simulations.

Appendix S3. Construction and analysis of mock samples.

How to cite this article: Shelton, A. O., Kachel, S. M., Shaffer, M. R., Brandão-Dias, P. F. P., Jacobson, E. K., Thomas, L., & Kelly, R. P. (2026). Estimating absolute DNA abundance from replicated metabarcoding. *Methods in Ecology and Evolution*, 00, 1–14. <https://doi.org/10.1111/2041-210x.70427>