Distinguishing Signal from Noise: Understanding Patterns of Non-Detections to Inform Accurate Quantitative Metabarcoding

Zachary Gold[1,2*], Andrew Olaf Shelton[2], Helen R. Casendino[3], Joe Duprey[3], Ramón Gallego[2], Amy Van Cise[2], Mary Fisher[4], Alexander J. Jensen[2], Erin D'Agnese[3], Elizabeth Andruszkiewicz Allan[3], Ana Ramón-Laca[2], Maya Garber-Yonts[3], Michaela Labare[5], Kim M. Parsons[2], Ryan P. Kelly[3]

[1] Cooperative Institute for Climate, Ocean, & Ecosystem Studies, UW, Seattle, WA

[2] Northwest Fisheries Science Center, NMFS/NOAA, Seattle, WA

[3] School of Marine and Environmental Affairs, UW, Seattle, WA

[4] School of Aquatic Fisheries Science, UW, Seattle, WA

[5] Scripps Institution of Oceanography, UCSD, La Jolla

*Corresponding author email: Zachary.gold@noaa.gov

Conflicts of Interest: The authors have no conflicts of interest to report.

Key Words: quantitative metabarcoding, eDNA, amplicon sequencing, non-detection, zero inflation

Abstract

Correcting for amplification biases in genetic metabarcoding data can yield quantitative estimates of template DNA concentrations. However, a major source of uncertainty in metabarcoding data is the presence of non-detections, where a technical PCR replicate fails to detect a species observed in other replicates. Such non-detections are an important special case of variability among technical replicates in metabarcoding data, particularly in environmental samples. While many sampling and amplification processes underlie observed variation in metabarcoding data, understanding the causes of non-detections is an important step in distinguishing signal from noise in metabarcoding studies. Here, we use both simulated and empirical data to 1) develop a qualitative understanding of how non-detections arise in metabarcoding data, 2) outline steps to recognize uninformative data in practice, and 3) identify the conditions under which amplicon sequence data can reliably detect underlying biological signals. We show in both simulations and empirical data that, for a given species, the rate of non-detections among technical replicates is a function of both the template DNA concentration and species-specific amplification efficiency. Consequently, we conclude metabarcoding datasets are strongly affected by (1) deterministic amplification biases during PCR and (2) stochastic sampling of amplicons during sequencing — both of which we can model — but also by (3) stochastic sampling of rare molecules prior to PCR, which remains a frontier for quantitative metabarcoding. Our results highlight the importance of estimating species-specific amplification efficiencies and critically evaluating patterns of non-detection in metabarcoding datasets to better distinguish environmental signal from the noise inherent in molecular detections of rare targets.

Introduction

Metabarcoding, or DNA amplicon sequencing, is a powerful tool that can characterize biological communities without the need to physically observe individual organisms. The rise of metabarcoding via high-throughput sequencing is rapidly advancing human and wildlife health, ecology, and conservation science [1–8].

However, the power of metabarcoding applications lies in the ability to obtain reliable quantitative estimates of underlying communities [9–11]. In the case of metabarcoding and similar amplicon-based studies [12], it has become clear that 1) observations are non-linearly related to the underlying biology of interest [13,14], and 2) those observations are noisy, with many having relatively high variances as a function of expected values [11,15–17]. In order to make reliable quantitative estimates for any set of observations, we must be able to distinguish random variation from real signal. Thus, understanding the underlying signal-to-noise ratio is key to quantifying the power of detection for a given dataset [18].

Substantial efforts to correlate sequence reads and underlying community abundance have reported promising but largely equivocal results [10,15,19–23]. However, it is unsurprising that the application of simple linear correlations to non-linear and compositional datasets produce ambiguous results given the failure to model the underlying drivers of observed DNA sequence patterns and distributions. In response, recent mechanistic frameworks have begun to address the discrepancies between observed metabarcoding sequence counts and true underlying biological patterns by modeling the compounding processes that occur between DNA extraction and sequence observation [24–28]. These processes include DNA extraction, PCR, and multiple subsampling steps prior to sequencing [16,17,27,29,30]. Importantly, these mechanistic frameworks explicitly model the amplicon sequence-generating process by stating that observed

sequence reads are a function of both the species-specific amplification efficiency and the underlying abundance of each species' DNA within a sample [26]. Such models also reflect the inherent compositional nature of metabarcoding, acknowledging that metabarcoding data can only provide proportional (not absolute) abundances of a given species' DNA in a given sample [11]. This approach can reconstruct starting DNA proportions, prior to PCR (e.g. [16,25–27]) and, where metabarcoding data are combined with additional information on underlying DNA concentrations, can yield absolute abundance estimates of the sampled DNA concentrations (e.g. [31]).

Despite these advances in modeling the amplicon sequence-generating process, it is clear that the sequential molecular steps required to generate metabarcoding data result in highly variable sequence-read counts among technical replicates derived from the same DNA extract [13,32–36]: replicate samples yield somewhat different results. Thus, in practice, it can be difficult to distinguish signal from noise in metabarcoding datasets. In particular, zeros are frequently over-represented in metabarcoding data, contributing substantially to among-replicate variability [16,24,34]. For example, in three technical replicates, a unique amplicon sequence variant (ASV) may be represented by 3,897; 165; and 0 reads across replicates (132,731, 196,260, 55,400 read depth for each replicate respectively; [31]). This observed variability among technical replicates far exceeds the expected variability arising from binomial- or multinomial sampling, and so demands a different explanation [16].

Here we focus on the patterns and causes of non-detections (in which a species is unobserved in one technical replicate despite being observed in other replicates) in metabarcoding datasets. After first synthesizing previous research on patterns of sequence counts, we simulate the process of metabarcoding to develop a qualitative understanding of the

scenarios under which non-detections arise. We then use these results to generate predictions for the frequency of non-detections. Next, we use empirical observations to test these predictions using metabarcoding data derived from a set of ethanol-preserved fish larvae, in which both the underlying organismal abundances and the resulting metabarcoding dataset are well-characterized. Our empirical findings closely match the predictions and suggest a mechanism for non-detections and stochastic variability in general. Given this understanding of the sources of variability, we can more confidently distinguish signal from noise in metabarcoding datasets.

Methods

Conceptual Model and Simulating Metabarcoding Data

Our generating model for metabarcoding derives from Shelton et al. [27], building on the work of others [11,17,25,26,29]. Briefly, we envision a metabarcoding dataset as compositional, arising from a chain of sampling and amplification processes acting on individual DNA molecules.

We start with a sample of extracted DNA containing sequences from multiple species. From this starting point, there are many different metabarcoding laboratory protocols that lead to observed sequences from a sequencing instrument [37,38]. Here, we approximate this using three main stochastic processes following the commonly used two-step PCR library generation process (e.g., a target PCR followed by an indexing PCR). First, we assume a sample of DNA is extracted and included in the multi-taxon PCR reaction. Second, PCR amplification using a specific primer and protocol occurs, replicating the DNA molecules for each taxon. This second step includes the various target PCR, cleaning, indexing PCR, and pooling steps that occur during or following the main PCR reaction. Finally, the resulting mix of DNA copies is sampled

to generate a compositional sample of amplicons that are observed through the sequencing instrument.

Mathematically, we can write a simulation for this framework as a series of linked stochastic processes. Specifically, we start with a sample of extracted DNA containing λ_i copies μL^{-1} DNA from the *i*th taxon, $i = 1,2,...,I$. Let W_{ij} be the discrete number of DNA molecules for species *i* sampled (i.e., in the tube in which a given PCR reaction takes place) of technical replicate *j* at the beginning of PCR and

$$W_{ij} \sim Poisson(\lambda_i V) \quad (1)$$

Here, V is the volume (μL) of template DNA sampled from the DNA extract. This equation assumes each taxon is sampled independently. Note that different technical replicates (e.g., $j = 1$ and $j = 2$) arise from the same environmental sample but may contain different numbers of molecules for a given species due to sampling variability.

Next, we model a three-step PCR process. Most importantly, we assume the amplicons produced during a PCR reaction are influenced by a species-specific amplification efficiency a_i, which is characteristic of the interaction between the particular primer set, reaction chemistry, and template molecule of each species (*i*) being amplified [27]. For any species, X_{ij} is the expected number of amplicons present in a technical replicate at the end of PCR. X_{ij} is directly related to the efficiency of amplification and the starting number of DNA molecules, $W_{ij}(1 + a_i)^{N_{pcr}}$, where N_{pcr} is the number of PCR cycles and a_i is bounded on (0,1); $a_i = 1$ represents a perfect doubling of molecules with each PCR cycle. For the purpose of this paper, we assume a two-step PCR process with a sub-sampling and PCR cleaning process in between, and X_{ij} can be modeled at each step as:

$$X_{1ij} \sim Poisson(W_{ij}(1 + a_i)^{N_{pcr1}}) \quad (2)$$

$$X_{2ij} \sim Binomial(\pi, X_{1ij}) \quad (3)$$

$$X_{3ij} \sim Poisson(X_{2ij}(1+0.9)^{N_{pcr2}}) \quad (4)$$

where π is the proportion of the first PCR product used in the second PCR amplification. Note that during the indexing reaction (equation 4) all taxa share a single amplification efficiency (a_i=*0.9*) as we assume all indexing primers anneal to the indexing adapter sequences with equal efficiency. Here X_3 is the number of amplicons present after both PCR amplifications but before sequencing. Finally, the sequencing instrument generates a total number of reads within technical replicate j (N_{readsj}) and each replicate has a vector of observed read counts ($\boldsymbol{Y}_j$, bolding indicates vectors) for I species.

$$\boldsymbol{Y}_j \sim Multinomial\left(\boldsymbol{p}_j, N_{reads,j}\right) \quad (5)$$

where p is the proportion of reads from species i in technical replicate j, and $p_{ij} = \frac{X_{3ij}}{\sum_{i=1}^{I} X_{3ij}}$. Thus, observed read counts (Y_{ij}) are sampled stochastically based on their relative amplicon abundances, p_{ij}.

The above model provides a general framework for understanding the causes of variability in observed read counts (Y_{ij}), specifically the probability of non-detections as a factor of 1) the initial DNA concentration λ_i and 2) the amount of species-specific variation in amplification efficiency (a_i). The simulation allows us to identify two distinct causes of non-detection, $p(Y_{ij} = 0)$. First, non-detection may occur because there were no molecules of species i in the initial PCR ($W_{ij} = 0$, in which case we are interested in $p(W_{ij} = 0)$ because $p(Y_{ij} = 0|W_{ij} = 0) = 1$). Second, zeros can arise due to PCR amplification and sequence-sampling processes; thus, we are interested in $p(Y_{ij} = 0|W_{ij} > 0)$. While $p(W_{ij} = 0)$ is trivial to calculate

from equation 1, determining $p(Y_{ij} = 0)$ and $p(Y_{ij} = 0|W_{ij} > 0)$ are not. We turn to simulations to understand the contributions of variation in λ_i and a_i to the probability of non-detection.

We simulate four communities with different levels of richness (N = 4, 10, 30, and 50 taxa). For simplicity, we assume all taxa start with identical DNA concentrations regardless of the richness. DNA concentrations, λ_i are varied from 0.5 to 10,000 copies μL^{-1} (for simplicity we set $V = 1$ for all simulations). We further allow a range of amplification efficiencies (a_i) among taxa where $a \sim Beta(0.7\gamma, 0.3\gamma)$ with γ ranging from 5 (high variation among species) to 1 million (no variation among species), but with a constant average amplification efficiency of 0.7 for all scenarios. We simulated 50,000 realizations for each combination of richness (4 levels), λ (18 levels), and γ (6 levels: 5, 10, 20, 100, 100, 1 million)), for a total of 432 scenarios. For all the simulations, we allowed sequencing depth to vary among replicates ($N_{read,j}$ was uniformly drawn from discrete values between 60,000 and 140,000), used a fixed sampling fraction ($\pi = 0.20$), and set $N_{pcr2} = 35$ and $N_{pcr2} = 10$). We calculated a range of summary statistics for each scenario, including the overall probability of non-detection, $p(Y_{ij} = 0)$; the probability of non-detection due to the absence of the target molecule, $p(W_{ij} = 0)$; and summaries of the reads both in absolute terms and in terms of relative abundance.

Empirical Testing

We test these hypotheses against a real metabarcoding data set making use of three data streams generated from a common set of biological samples: organismal abundance (as a proxy for input DNA molecules), metabarcoding data, and amplification-efficiency estimates for the relevant species.

Study Design

As part of the California Cooperative Oceanic Fisheries Investigations (CalCOFI), Gold et al. [31] use morphological and molecular methods to analyze the response of ichthyoplankton in the California Current Large Marine Ecosystem to ocean warming. Ichthyoplankton samples were collected in oblique bongo net tows on CalCOFI research cruises over two decades (1996; 1998-2019). Once a sampling tow concluded, ichthyoplankton present on one side of the net were preserved in Tris-buffered 95% ethanol and stored in the Pelagic Invertebrate Collection at Scripps Institution of Oceanography [39]. The paired ichthyoplankton were preserved in sodium borate-buffered 2% formaldehyde for microscopy-derived species identification and abundance (number of larvae per species per jar). This dataset yields paired samples for both metabarcoding analysis and absolute abundance counts from the same sampling event.

Abundance Estimation from Microscopy

Formalin-preserved larvae were identified and enumerated following the methods of Thompson et al. [40]. The majority of taxa were identified to species level. Here we assume that the relationship of absolute abundance (counts of individual species) is proportional to the amount of species-specific DNA in the extraction. See the discussion for the merits of this assumption.

Metabarcoding Data Generation

DNA sequences were generated from 84 ethanol-preserved samples as described in Gold et al. [31]. Briefly, ethanol samples were filtered onto 0.2 µm PVDF filters and were extracted using a Qiagen DNeasy Blood and Tissue kit. We then amplified three technical PCR replicates using a touchdown PCR and the MiFish Universal Teleost specific primer [41]. Both a negative control (molecular grade water instead of DNA extract) and two positive controls (DNA extract from non-native, non-target species) were included alongside samples. Libraries were prepared using Illumina Nextera indices following the methods of Curd et al. [42] and sequenced on a NextSeq

2x 150 bp mid output. Sequencing data was then processed using the *Anacapa Toolkit* [42] to conduct quality control, ASV dereplication, and taxonomic assignment. Sequences were annotated with the California fish specific reference database and a bootstrap confidence cutoff score of 60 following the methods of Gold et al. [43]. Eight technical replicates with either low sequencing depth (n<30,000) or high dissimilarity (Bray Curtis dissimilarity > 0.7) were removed.

Amplification Efficiency Estimation from Mock Communities

We used a subset of the mock communities generated for Shelton et al. [27] to estimate amplification efficiencies of relevant fish species. Mock communities included DNA from 57 voucher fish tissue samples, 17 of which were detected in the CalCOFI metabarcoding data set, from the Scripps Institution of Oceanography Marine Vertebrate Collection. To accurately quantify input DNA for each species within the mock community, we used a nested PCR strategy in which mock communities were generated by pooling resultant longer fragment PCR products of each species rather than by pooling the total genomic DNA of each species (which includes variable amounts of nDNA as well as bacterial and other DNA sources). To implement our nested PCR strategy, we first amplified a 612 bp fragment of the *12S* rRNA gene that contains the MiFish Universal Teleost *12S* primer set [44], and quantified the resulting PCR products using the QuBit Broad Range dsDNA assay (Thermofisher Scientific, Inc.); this yielded measurements of species-specific, amplifiable DNA. Using this known-concentration DNA we generated 9 distinct mock communities by pooling long fragment PCR products comprising three distinct sets of species and three abundance distributions (See Table S1). Pooled mock communities were using the QuBit Broad Range dsDNA assay (estimated concentrations ranged from 8-12 ng μL^{-1}) and then diluted serially by a 1:10 dilution down to

10^{-8} original concentration. We then converted ng μL^{-1} to copies μL^{-1} using the following equation:

$$\text{Copies } \mu L^{-1} = \frac{\text{QuBit Concentration } [\text{ng } \mu L^{-1}] * 6.022 \text{x} 10^{23} \text{ [molecules mol}^{-1}]}{612 \text{ [bp]} * 650 \text{ [g mol}^{-1} \text{ bp}^{-1}] * (1 \text{x} 10^{9} \text{ [ng g}^{-1}])}$$

Finally, input concentrations of 380-600 DNA copies μL^{-1} for each total community were loaded in the MiFish Universal Teleost *12S* PCR step (Table S2). Given this design, each mock community had a different number of DNA molecules per species. We then amplified each of the mock communities in triplicate with the MiFish Universal Teleost *12S* following the methods of Curd et al. [42], targeting a 185 bp fragment within the larger 612 bp PCR fragment used to generate the mock communities. Each triplicate PCR technical replicate was then treated as a unique library and sequenced separately. Metabarcoding libraries were then prepared and sequenced on a MiSeq platform using a v3 600 cartridge following the methods of Gold et al. [43]. We note that one set of mock communities were re-sequenced on a separate run to generate usable data. Resulting sequences were processed using the *Anacapa Toolkit* using the global *CRUX* generated reference database given the broad geographic distribution of species from Gold et al. [43]. We also used a taxonomic cutoff score of 60 as above. Taxonomic assignment of ASVs was confirmed with BLAST using default settings. For the two observed discrepancies, we chose to use BLAST assignments with greater than 99% identity and 100% query length match as they matched our known vouchered specimen identifications.

We fit the model from Shelton et al. [27] to a third of the data (3 technical replicates of each evenly pooled mock community). Generated parameter estimates were then used to predict the starting proportions of DNA in the remaining two-thirds of the data, for an out-of-sample estimate of accuracy. We used the resulting model output to calculate the mean amplification

efficiency per species. The model, implementation, and code are detailed in Shelton et al. [27], but there are two particularly relevant points from the model for connecting the simulation and empirical results that we highlight here. While we simulate absolute amplification efficiencies (a_i), because metabarcoding data is compositional, the absolute amplification efficiency cannot be estimated from metabarcoding data. Instead, we estimate amplification efficiencies for each species relative to a reference efficiency (see also [25,26]). In our case we estimate α_i as the amplification efficiency of species i, a_i, relative to the efficiency of a reference species, a_R, therefore $\alpha_i = \frac{a_i}{a_R}$. Thus, for simulations we discuss a but for estimation, we discuss α. Note while values of α can be directly calculated from a, values of a are not uniquely identifiable from α.

All data and code for conducting analyses will be made publicly available upon acceptance via NCBI SRA, Dryad, and GitHub (https://github.com/zjgold/Metabarcodings_Signal_from_Noise).

Hypothesis Testing

From the above outlined simulations and empirical data, we generate a series of hypotheses. First, we expect fewer non-detections for more abundant DNA molecules, given the same species (and therefore the same amplification efficiency). Second, we hypothesize that species with higher amplification efficiencies will have fewer non-detections and higher observed sequence read counts than species with lower amplification efficiencies, given the same abundance of template DNA molecules. Third, we expect species with low amplification efficiencies will have a high rate of non-detection regardless of the abundance of template DNA molecules. We test each of these hypotheses using both simulation and empirical results.

Results

Simulation Results

We found a strong correlation between the probability of non-detections and both the absolute abundance of template DNA molecules and amplification efficiencies (Figure 1). The probability of non-detections (*p(Y=0)*) dramatically declines when concentrations of template DNA are greater than ~10 copies μL^{-1} per species, given an average amplification efficiency of 0.7 (Figures 1C & 1D). Likewise, our results demonstrate that species with low amplification efficiencies exhibit high probabilities of non-detections regardless of starting DNA concentrations (Figure 1A, B). Importantly, we demonstrate that even species with an amplification efficiency slightly below average (e.g., $a = 0.7$) exhibit high rates of non-detections at DNA concentrations far higher than from typical eDNA field samples (e.g. $\lambda > 100$ copies/µL; [45]). Together these simulations indicate that the probability of non-detection is dominated by the subsampling process at low template DNA concentrations while the probability of non-detection is driven primarily by the PCR process (i.e., differences in amplification efficiencies) at higher template DNA concentrations.

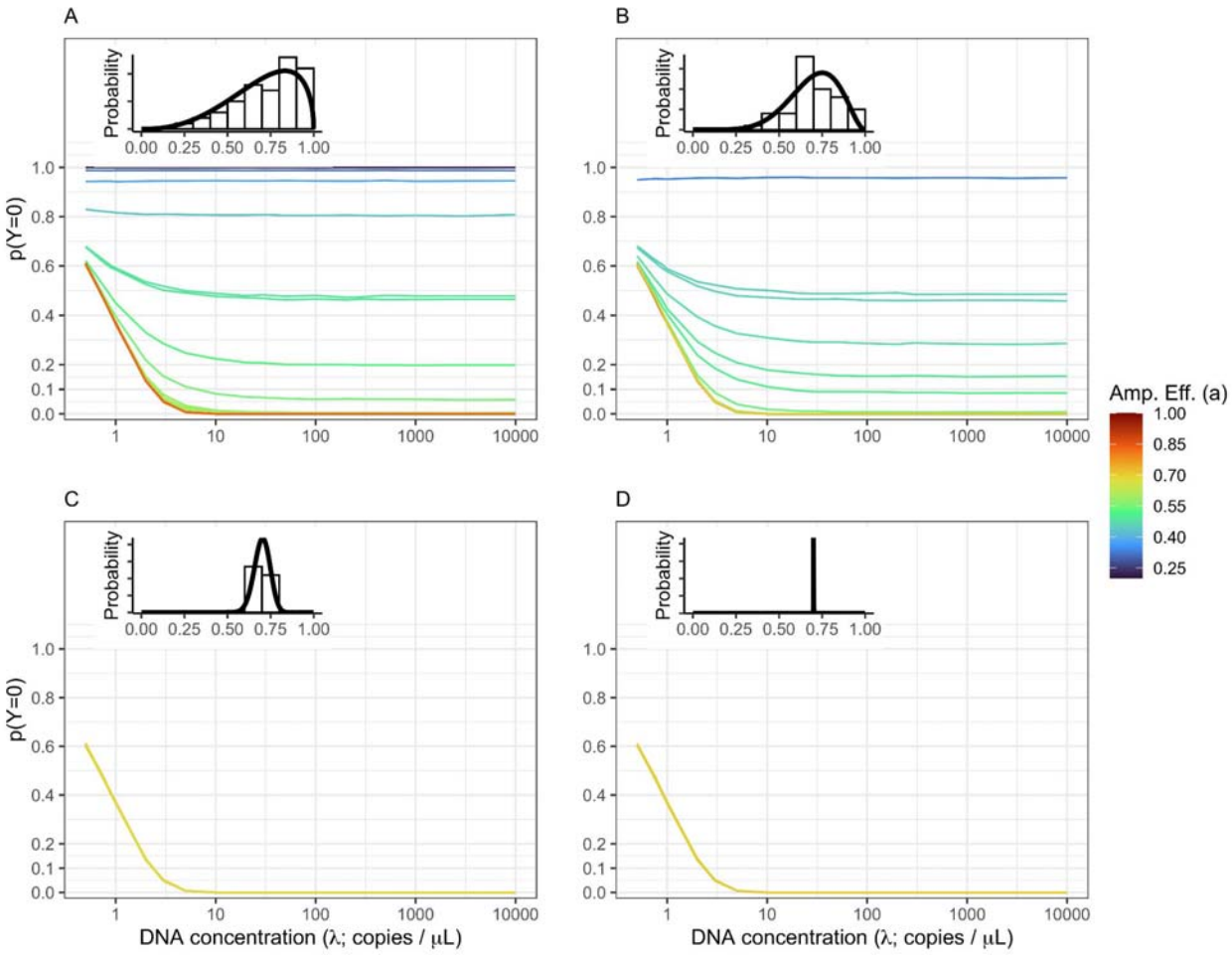

Figure 1. Non-detections Driven By Both DNA Concentration and Amplification Efficiency.

The probability of non-detection (*p(Y=0)*) is shown for a community of 50 equally abundant taxa with the amplification efficiency distribution shown inset in the upper left of each panel. The amount of among-taxa variation in amplification efficiency varies from high variation *(A;* γ=5) to moderate variation (*B:* γ=10) to low variation (*C:* γ=100) to effectively no variation (*D:* γ=1,000,000). Both subsampling and amplification efficiencies influence the rate of non-detection. The probability of observing no DNA in a given technical replicate is highest at low DNA concentrations (<10 copies /µL). However, non-detections are possible for species with below average amplification

297 efficiencies (in this case approximately a_i = 0.7) and very likely ($p(Y=0)$ > 0.5) for
298 amplification well below average ($a_i < 0.4$).
299

300 ***Empirical Results***

301 *Microscopy Results*

302 Independent estimates of abundance were generated from sorting 9,610 larvae from 84 jars (min
303 = 2, max =960). See Gold et al. [31] for a detailed description of the results.

304 *Metabarcoding Results*

305 The metabarcoding data set generated from ethanol-derived eDNA consisted of a total of 54.5
306 million amplicon sequence reads that passed through the *Anacapa Toolkit* quality control, ASV
307 dereplication, and decontamination processes. Sequencing depth ranged from 36,050 reads to 1.2
308 million reads per technical replicate. For our integrated Bayesian model of the probability of
309 non-detection in a technical replicate, we focused on the 17 species that had 1) sufficient
310 representation across the metabarcoding data set (observed in > 10 technical PCR replicates) to
311 achieve model convergence and 2) were represented in our mock communities. See Gold et al.
312 [31] for the full description of model implementation and results.

313 *Mock Community Results*

314 The mock community data set consisted of 4.0 million amplicon sequence reads that passed
315 through the *Anacapa Toolkit* quality control, ASV dereplication, and decontamination processes
316 across a total of 36 unique samples comprising three distinct community assemblages each with
317 three PCR technical replicates. Sequencing depth ranged from 9,872 reads to 206,900 reads per
318 technical replicate. Of the 57 voucher species represented, we classified 56 unique species, and

used the Shelton et al. [27] model to estimate amplification efficiencies for each species. One species, *Urobatis halleri,* was not detected in any technical replicate. *Citharichthys sordidus* was present in all mock communities and was selected as the reference species for estimating relative amplification efficiencies. Across all species, α_i ranged from -0.30 to 0.03 with a mean of -0.06 (Table S3). For presentation purposes, we label species with α_i values below -0.07 as a low amplification efficiency group (n=15) and the remaining species as a high amplification (n=41).

Hypothesis Testing

As with the simulation results, we found that the probability of non-detections is strongly correlated with both the abundance of DNA molecules for a given species within a sample and the species-specific amplification efficiency (Figures 2b, 3b). Non-detections occur more frequently at low DNA concentrations regardless of amplification efficiency (Figures 2b, 3b). Species exhibiting lower amplification efficiencies (α_i < -0.07) had higher rates of non-detections even at high input DNA concentrations (10^4 copies μL^{-1}) and larval counts (9 larvae per jar; Figures 2b, 3b).

Furthermore, from the mock community example, species with higher amplification efficiencies (α_i >-0.07) have higher observed sequence read counts for an equivalent template DNA concentration (Figure 2a). The 41 species with high amplification efficiencies have more reads sequenced per DNA molecule added (mean ± sd = 4.1 ± 6.31, range = 0.00-55.4) than the 15 species with low amplification efficiencies (mean ± sd = 0.1 ± 0.47, range = 0.00-5.5). Likewise, species with higher larval counts in ethanol-preserved samples from plankton tows also have higher observed sequence read counts (Figure 3a). From the CalCOFI example, the 15 species with high amplification efficiencies have more reads sequenced per larvae counted (mean

± sd = 6,689 ± 28,305, range = 0-79,454) than the two species in the low amplification efficiency group (mean ± sd = 524 ± 1,080, range = 0-7,101).

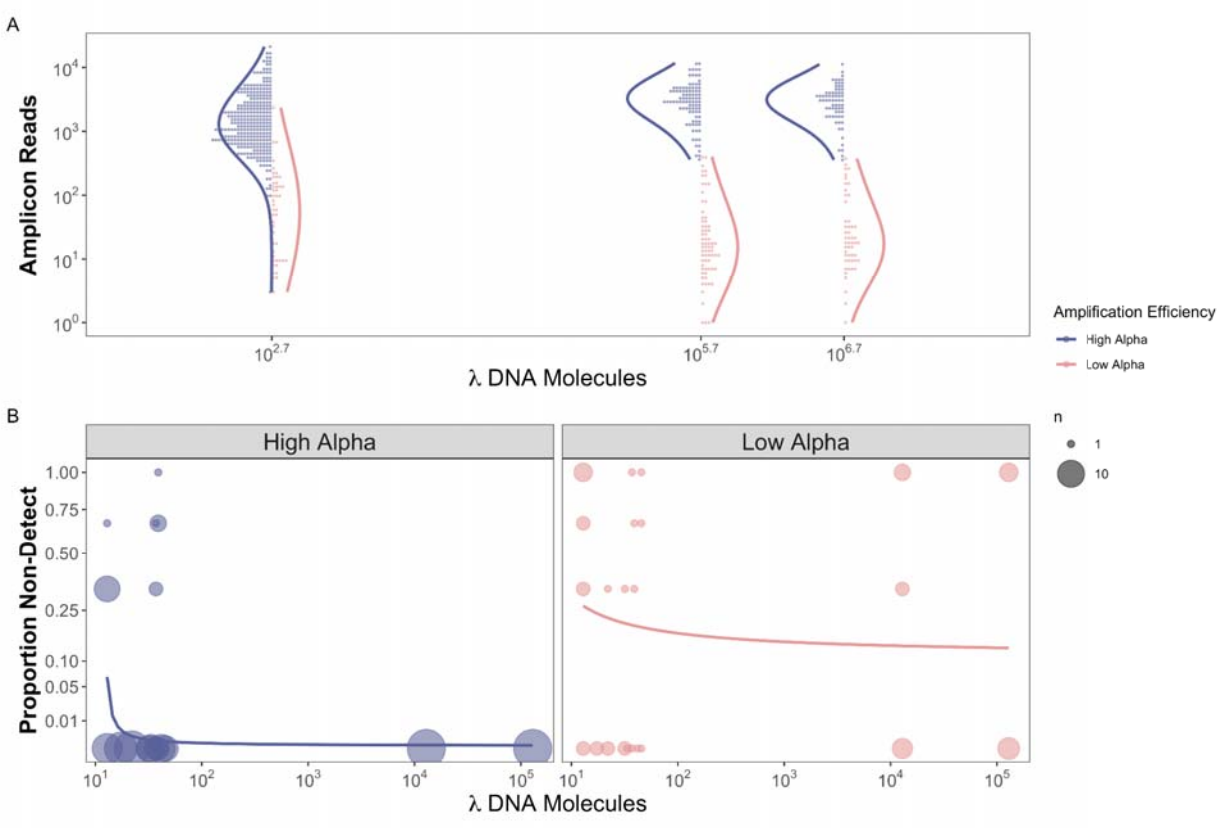

Figure 2: Observed Reads and Non-detections are a function of Amplification Efficiency and Input DNA Concentration in the Mock Community example

For species observed within a replicate, we find that species with higher amplification efficiencies (>0.7) have a greater number of observed reads for an equivalent template DNA concentration (a). We also find no difference in the total number of observed reads and increased DNA concentration as expected for a compositional data set. Furthermore, we find a greater proportion of non-detections when both DNA concentration and amplification efficiencies are lower (b). These results align well with our simulated data.

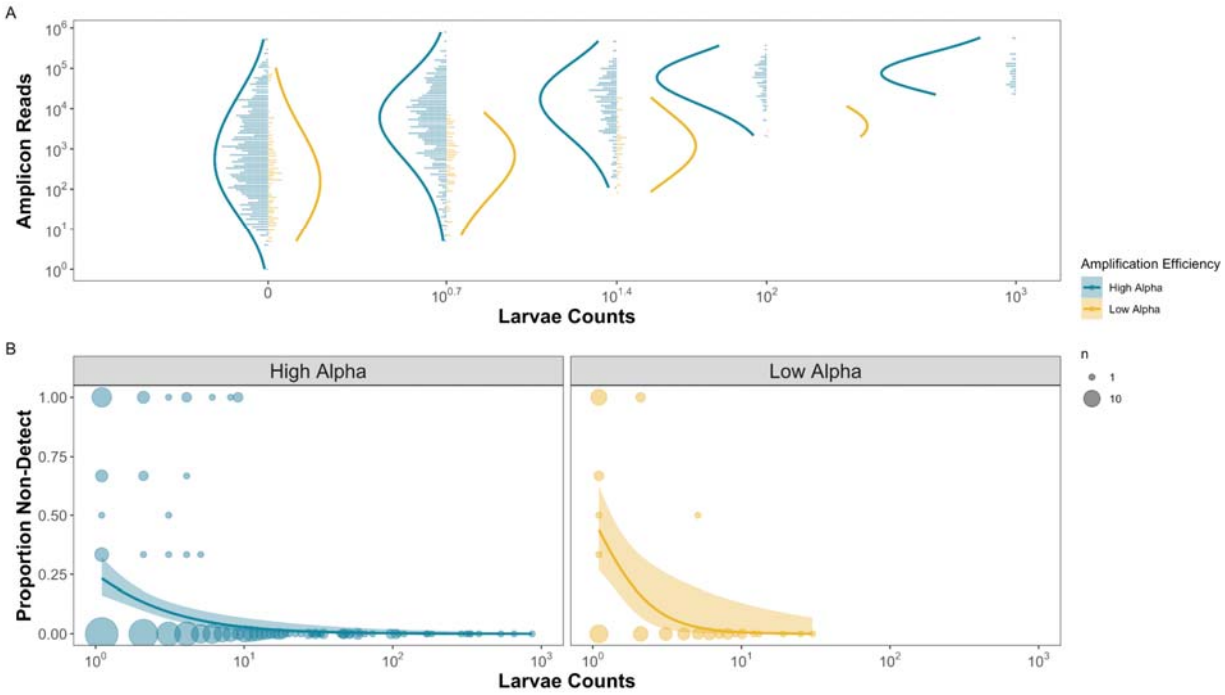

Figure 3. Observed Reads and Non-detections are a function of Amplification Efficiency and Larval Abundances in the CalCOFI example

For species observed within a replicate, we find that species with higher amplification efficiencies (>-0.07) have consistently greater numbers of observed reads for an equivalent template DNA concentration (a). We assume that the number of larvae in the jar is proportional to the number of DNA molecules present. We also find a greater proportion of non-detections when larvae are rare in the jars and have lower amplification efficiencies (b).

Discussion

Using both simulated and empirical data, we demonstrate that observed sequence read counts from metabarcoding data are a function of species-specific input DNA concentrations, subsampling, and species-specific amplification efficiencies. Variability among replicates in

368 detections of specific taxa – reflecting either rare targets or poor amplification efficiencies – are
369 a substantial source of noise in these data. Consequently, it can be difficult to distinguish signal
370 from noise in metabarcoding datasets. Our results illustrate several potential causes of non-
371 detections and suggest that metabarcoding data can provide reliable quantitative estimates for
372 species with abundant input DNA (> ~50 copies μL^{-1}) and high species-specific amplification
373 efficiencies. By characterizing underlying sources of sequence read count variability in
374 metabarcoding, we identify key sources of noise that impact our ability to derive quantitative
375 estimates of source DNA.

376 *Subsampling Rare Targets Results in Non-detections*

377 Consistent with expectation, our framework strongly suggests that all else being equal in a
378 metabarcoding assay (e.g., assuming even amplification efficiencies across species), rarer
379 template DNA molecules have a higher probability of non-detection across technical replicates.
380 These findings align well with observations of qPCR assays in which the probability of non-
381 detection increases as you approach the limit of detection, in terms of absolute copies of DNA
382 per reaction volume [46,47]. High rates of non-detections in qPCR assays are commonly
383 observed for input DNA concentrations between 1 and 10 copies [46,48,49] and are likely driven
384 by subsampling errors in which too few or no physical DNA molecules are transferred into a
385 given PCR reaction [36,50,51]. These observations from qPCR studies reflect the findings from
386 both simulated and empirical metabarcoding results reported here.

387 Importantly, subsampling rare target DNA molecules yields stark differences in observed
388 per-species read counts among technical replicates, non-detections being the most obvious case
389 of this phenomenon [15,16,34,46]. Together, these findings strongly support the hypothesis that
390 the concentration of target DNA within a sample influences the observed patterns of amplicon

read counts, particularly increasing the probability of non-detections for species with low template DNA concentrations. Such observations of high rates of non-detections also justify the use of over dispersed multinomial sampling approaches within metabarcoding models [27].

Amplification Efficiencies Drive Sequence Counts and Non-detections

While the relationship between template concentration and non-detections is well documented in the literature [10,12,16,24,34,46], the causes of non-detection among species with abundant DNA are not widely appreciated. Both simulation and empirical results demonstrate that species with higher amplification efficiencies have higher observed amplicon read counts, confirming the predictions of previous compositional modeling efforts [25–27]. Furthermore, we find a clear association between the probability of non-detections and amplification efficiencies, with species with higher amplification efficiencies exhibiting fewer non-detections. Here we observed an order of magnitude difference in average sequence reads per larvae collected in a given jar with a maximum observed difference in amplification efficiencies of 0.33 (n=17). Previous research using mock communities has similarly demonstrated that equal concentrations of DNA in a single extraction frequently results in amplicon read counts that differ by orders of magnitude [16,32,36,37,52]. Such dramatic differences in resulting read proportions are understandable given the exponential nature of PCR - even a subtle difference in amplification efficiency across 30+ PCR cycles can result in stark differences in sequence counts [27].

The observed variation in amplification efficiency among species in metabarcoding approaches arises from complex PCR processes, including primer specificity, DNA polymerase selectivity, annealing temperature, GC content, and higher-order dimensional structure of DNA, inhibition, and co-factors such as $MgCl_2$, among others [53–59]. This complexity makes designing metabarcoding assays that are highly specific for only target taxa challenging [60,61],

414 resulting in the amplification of off-target taxa as well as a range of amplification efficiencies
415 across target taxa [26,27,43,62,63]. As demonstrated by our simulations and empirical results,
416 such a range of amplification efficiencies can result in substantial noise in metabarcoding data
417 sets.

418 *Complex Relationship between Amplification Efficiencies and Abundance*

419 The above results highlight the cumulative importance of the variance in amplification efficiency
420 among species, as well as the abundance of template DNA for understanding the patterns of
421 metabarcoding non-detections. The interaction between these factors is key for disentangling the
422 signal from the noise of metabarcoding data. Here, we demonstrate that there are two ways to
423 obtain non-detections for a given species after sequencing: low initial DNA concentration or low
424 amplification efficiency. Both of these results are clear from our empirical CalCOFI fish larvae
425 dataset which captured the effects of species-specific amplification efficiency and DNA
426 concentrations on both sequence read counts and frequency of non-detections (Figure 2).
427 Importantly, our results demonstrate that noise in metabarcoding datasets, like signal, is non-
428 random and can be accounted for [16].

429 Alone, metabarcoding data is insufficient to tease apart these complex interactions.
430 However, distinguishing signal from noise in metabarcoding datasets is tractable using
431 independent estimates of amplification efficiencies and underlying DNA concentrations.
432 Amplification efficiencies can be estimated through either generating mock communities
433 [26,27], by amplifying a subset of samples multiple times at various numbers of PCR cycles
434 [25], or by including internal positive controls within each PCR [28]. Likewise, underlying DNA
435 concentrations can be estimated using qPCR or dPCR assays of key taxa or the metabarcoding
436 locus itself; or estimated using non-genetic independent abundance estimates such as the

437 microscopy counts presented above. As demonstrated here, and in Shelton et al. [27], McLaren et
438 al. [26], and Silverman et al. [25], the inclusion of independent estimates of amplification
439 efficiencies and DNA concentrations allow for the delineation of signal from noise from
440 metabarcoding data sets. Further modeling efforts incorporating stochastic sampling of rare
441 molecules prior to PCR will allow for accurate quantification and identification of true absences
442 in metabarcoding data sets, greatly enhancing biological and ecological interpretation.

443 Furthermore, our analysis also underscores the importance of technical PCR replicates to
444 quantify sequence variance in metabarcoding studies [64–66]. Without technical replicates, we
445 would not have been able to quantify the frequency of non-detections in our metabarcoding
446 datasets [17]. We demonstrate that non-detections may indicate low-relative-abundance starting
447 DNA concentrations regardless of observed read depth, and conversely, may indicate low
448 amplification efficiency regardless of starting concentration [27]. Thus, our results strongly
449 support the inclusion of technical replicates for metabarcoding studies, particularly for deriving
450 quantitative estimates.

451 Current best practices for qPCR and dPCR assays include numerous technical replicates
452 to help distinguish signal from noise [46,48]. However, we recognize that technical replication
453 dramatically increases the cost and effort of metabarcoding projects and may exhaust limited
454 DNA extracts and resources. Alternatively, technical replicates could be performed on a subset
455 of samples and the observed variance could be used to contextualize sequence read patterns in
456 the whole dataset. However, such approaches come with a suite of assumptions, particularly
457 whether the pattern of species' sequence counts behaves similarly across all samples and
458 environments/treatments. Future efforts to validate such approaches are clearly warranted.

459 In addition, given the importance of subsampling in driving non-detections, our results
460 strongly suggest that field and laboratory processes that increase the absolute abundance of DNA
461 molecules will reduce the noise in observed amplicon sequence reads [67]. For example, using a
462 greater volume of DNA template for PCR reactions (3 µL vs. 1 µL) will reduce subsampling
463 driven non-detections across samples. Likewise, increasing the total amount of water filtered for
464 eDNA samples (3 L vs. 1 L) acts to concentrate DNA from the environment, similarly reducing
465 subsampling driven non-detections [68]. These are two of many examples of laboratory
466 protocols that may serve to increase the available number of DNA molecules and reduce the
467 impacts of subsampling rare molecules, consequently improving quantitative estimates from
468 amplicon sequence data.
469 The above mechanistic frameworks focus on processes from DNA extraction through
470 sequencing, but do not approach the myriad of factors that influence the amount of DNA
471 collected from the environment, gut, or other starting communities for metabarcoding.
472 Substantial efforts have focused on understanding the effects of gene copy number, patchiness,
473 shedding and degradation rates, and the fate and transport of cellular DNA, among others, on the
474 amount/types of DNA collected from the environment [26,51,69]. Linking such research to the
475 growing body of work that quantifies sources of potential bias in the lab, including the present
476 study, is an important next step in understanding the relationship between biological signals and
477 observed sequence read counts.
478 We recognize that incorporating the additional laboratory analyses and technical
479 replicates to better characterize metabarcoding results may not be feasible for all metabarcoding
480 applications. Many metabarcoding efforts are exploratory in nature, primarily focused on the
481 characterization of biodiversity in under sampled habitats including the deep sea, polar regions,

remote alpine regions, etc. For such exploratory biodiversity surveys, the additional efforts needed to achieve quantitative metabarcoding outlined above may not be practicable given surveying and budget constraints. However, it is important to recognize that our framework extends not only to quantitative metabarcoding but detection rates of taxa from metabarcoding surveys. The expected detection rate (observed reads > 0) of a given taxon in metabarcoding data is a function of other species in the community, the amplification rate of the target species, the amplification rates of other species, the proportional abundance of the target species, and the absolute abundance of the target species as demonstrated in our empirical datasets above. Thus, estimating the probability of detection from metabarcoding data alone is difficult in the abstract, but is quite tractable given a set of estimated parameters for a particular sampled community. Conversely, interpreting metabarcoding results from exploratory applications within systems with limited ecological context is challenging as species detection rates are a function of multiple unsampled parameters.

Undoubtedly, addressing this shortcoming of compositional metabarcoding data requires increased field and laboratory efforts. Such challenges are acute in under studied systems where the creation of mock communities is particularly difficult with limited access to vouchered DNA samples, let alone known species lists. However, exploratory metabarcoding studies do not preclude the revisiting of quantitative metabarcoding approaches in the future, especially since DNA extracts can be archived. For example, metabarcoding data can be generated first to provide an initial perspective into community assemblages that then allows for the identification and development of single species qPCR/dPCR assays and mock communities or variable PCR targets. In summary, we argue that all future best practices of metabarcoding results incorporate additional independent estimates of amplification efficiency, independent estimates of DNA

concentrations, and technical replicates to better contextualize metabarcoding efforts. Given the rapid decline in sequencing costs and steady improvement in the development and implementation of molecular assays, such additional work is tractable, opening the door to adoption for routine application across metabarcoding studies to generate characterization of underlying biological communities.

Conclusion

Ultimately, we demonstrate that variation in amplification efficiencies and underlying template DNA concentration are responsible for a substantial portion of observed noise in metabarcoding datasets. This study demonstrates the value of incorporating additional independent estimates of amplification efficiencies and DNA concentration along with amplicon sequence data, providing for the application of routine statistical approaches and straightforward interpretation of observed read patterns. Together with Shelton et al. [27], we provide a framework for establishing reliable estimates of abundance from amplicon sequence data that will be critical for extending the application of this method to health and ecological questions.

Data Availability Statement

All data and code for analyses will be made publicly available at NCBI SRA, Dryad, and Github (https://github.com/zjgold/Metabarcodings_Signal_from_Noise) upon acceptance.

Conflict of Interest

Authors have no conflicts of interest to report.

References

1. Smeele ZE, Ainley DG, Varsani A. Viruses associated with Antarctic wildlife: From serology based detection to identification of genomes using high throughput sequencing. Virus Res. 2018;243: 91–105.
2. Sekse C, Holst-Jensen A, Dobrindt U, Johannessen GS, Li W, Spilsberg B, et al. High throughput sequencing for detection of foodborne pathogens. Front Microbiol. 2017;8: 2029.
3. Soon WW, Hariharan M, Snyder MP. High-throughput sequencing for biology and medicine. Mol Syst Biol. 2013;9: 640. doi:10.1038/MSB.2012.61
4. Poisot T, Péquin B, Gravel D. High-Throughput Sequencing: A Roadmap Toward Community Ecology. Ecol Evol. 2013;3: 1125–1139. doi:10.1002/ECE3.508
5. Manor O, Dai C, Kornilov S, … BS-N, 2020 undefined. Health and disease markers correlate with gut microbiome composition across thousands of people. nature.com. [cited 28 Jul 2022]. Available: https://www.nature.com/articles/s41467-020-18871-1
6. Taberlet P, Bonin A, Zinger L, Coissac E. Environmental DNA: For biodiversity research and monitoring. Environmental DNA: For Biodiversity Research and Monitoring. Oxford University Press; 2018. doi:10.1093/oso/9780198767220.001.0001
7. Beng KC, Corlett RT. Applications of environmental DNA (eDNA) in ecology and conservation: opportunities, challenges and prospects. Biodivers Conserv. 2020;29: 2089–2121. doi:10.1007/s10531-020-01980-0
8. Bohmann K, Evans A, Gilbert MTP, Carvalho GR, Creer S, Knapp M, et al. Environmental DNA for wildlife biology and biodiversity monitoring. Trends Ecol Evol. 2014;29: 358–367. doi:10.1016/j.tree.2014.04.003
9. Lacoursière-Roussel A, Rosabal M, Bernatchez L. Estimating fish abundance and biomass from eDNA concentrations: variability among capture methods and environmental conditions. Mol Ecol Resour. 2016;16: 1401–1414. doi:10.1111/1755-0998.12522
10. Yates MC, Fraser DJ, Derry AM. Meta-analysis supports further refinement of eDNA for monitoring aquatic species-specific abundance in nature. Environmental DNA. 2019;1: 5–13. doi:10.1002/edn3.7
11. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: And this is not optional. Front Microbiol. 2017;8: 2224. doi:10.3389/fmicb.2017.02224
12. Jia C, Hu Y, Kelly D, Kim J, Li M, Zhang NR. Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. Nucleic Acids Res. 2017.
13. Edgar RC. UNBIAS: An attempt to correct abundance bias in 16S sequencing, with limited success. BioRxiv. 2017; 124149.
14. Ficetola GF, Pansu J, Bonin A, Coissac E, Giguet-Covex C, de Barba M, et al. Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. Mol Ecol Resour. 2015;15: 543–556. doi:10.1111/1755-0998.12338
15. Stoeckle MY, Adolf J, Charlop-Powers Z, Dunton KJ, Hinks G, Vanmorter SM. Trawl and eDNA assessment of marine fish diversity, seasonality, and relative abundance in coastal New Jersey, USA. ICES Journal of Marine Science. 2021;78: 293–304. doi:10.1093/icesjms/fsaa225
16. Silverman JD, Roche K, Mukherjee S, David LA. Naught all zeros in sequence count data are the same. Comput Struct Biotechnol J. 2020;18: 2789–2798. doi:10.1016/j.csbj.2020.09.014

17. Egozcue JJ, Graffelman J, Ortego MI, Pawlowsky-Glahn V. Some thoughts on counts in sequencing studies. NAR Genom Bioinform. 2020;2: 1–10. Available: https://academic.oup.com/nargab/article/2/4/lqaa094/5996081
18. Bessey C, Jarman SN, Berry O, Olsen YS, Bunce M, Simpson T, et al. Maximizing fish detection with eDNA metabarcoding. Environmental DNA. 2020;2: 493–504.
19. Kelly RP, Port JA, Yamahara KM, Crowder LB. Using environmental DNA to census marine fishes in a large mesocosm. PLoS One. 2014;9: e86175. doi:10.1371/journal.pone.0086175
20. Evans NT, Olds BP, Renshaw MA, Turner CR, Li Y, Jerde CL, et al. Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. Mol Ecol Resour. 2016;16: 29–41. doi:10.1111/1755-0998.12433
21. Yates MC, Glaser DM, Post JR, Cristescu ME, Fraser DJ, Derry AM. The relationship between eDNA particle concentration and organism abundance in nature is strengthened by allometric scaling. Mol Ecol. 2021;30: 3068–3082.
22. di Muri C, Handley LL, Bean CW, Li J, Peirson G, Sellers GS, et al. Read counts from environmental DNA (eDNA) metabarcoding reflect fish abundance and biomass in drained ponds. Metabarcoding Metagenom. 2020;4: 97–112. doi:10.3897/MBMG.4.56959
23. Rourke ML, Fowler AM, Hughes JM, Broadhurst MK, DiBattista JD, Fielder S, et al. Environmental DNA (eDNA) as a tool for assessing fish biomass: A review of approaches and future considerations for resource surveys. Environmental DNA. 2022;4: 9–33.
24. Jiang R, Li WV, Li JJ. mbImpute: an accurate and robust imputation method for microbiome data. Genome Biology 2021 22:1. 2021;22: 1–27. doi:10.1186/S13059-021-02400-4
25. Silverman JD, Bloom RJ, Jiang S, Durand HK, Dallow E, Mukherjee S, et al. Measuring and mitigating PCR bias in microbiota datasets. PLoS Comput Biol. 2021;17: e1009113. doi:10.1371/journal.pcbi.1009113
26. McLaren MR, Willis AD, Callahan BJ. Consistent and correctable bias in metagenomic sequencing experiments. Elife. 2019;8: e46923. doi:10.7554/eLife.46923
27. Shelton AO, Gold ZJ, Jensen AJ, D'Agnese E, Allan EA, van Cise A, et al. Toward quantitative metabarcoding. bioRxiv. 2022.
28. McLaren M, Nearing J, Willis A, bioRxiv KL-, 2022 undefined. Implications of taxonomic bias for microbial differential-abundance analysis. biorxiv.org. [cited 21 Aug 2022]. Available: https://www.biorxiv.org/content/10.1101/2022.08.19.504330
29. Shelton AO, O'Donnell JL, Samhouri JF, Lowell N, Williams GD, Kelly RP. A framework for inferring biological communities from environmental DNA. Ecological Applications. 2016;26: 1645–1659.
30. Coissac E, Riaz T, Puillandre N. Bioinformatic challenges for DNA metabarcoding of plants and animals. Mol Ecol. 2012;21: 1834–1847. doi:10.1111/j.1365-294X.2012.05550.x
31. Gold Z, Kelly RP, Shelton AO, Thompson AR, Goodwin KD, Gallego R, et al. Message in a Bottle: Archived DNA Reveals Marine Heatwave-Associated Shifts in Fish Assemblages. biorxiv. 2022.
32. Braukmann TWA, Ivanova N v., Prosser SWJ, Elbrecht V, Steinke D, Ratnasingham S, et al. Metabarcoding a diverse arthropod mock community. Mol Ecol Resour. 2019;19: 711–727. doi:10.1111/1755-0998.13008

33. Port JA, O'Donnell JL, Romero□Maraccini OC, Leary PR, Litvin SY, Nickols KJ, et al. Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. Mol Ecol. 2016;25: 527–541.
34. Kaul A, Mandal S, Davidov O, Peddada SD. Analysis of microbiome data in the presence of excess zeros. Front Microbiol. 2017;8: 2114.
35. Guenay-Greunke Y, Bohan DA, Traugott M, Wallinger C. Handling of targeted amplicon sequencing data focusing on index hopping and demultiplexing using a nested metabarcoding approach in ecology. Scientific Reports 2021 11:1. 2021;11: 1–15. doi:10.1038/s41598-021-98018-4
36. Leray M, Knowlton N. Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina COI metabarcoding. PeerJ. 2017;2017: e3006. doi:10.7717/peerj.3006
37. Gohl DM, Vangay P, Garbe J, MacLean A, Hauge A, Becker A, et al. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nat Biotechnol. 2016;34: 942–949.
38. Bohmann K, Elbrecht V, Carøe C, Bista I, Leese F, Bunce M, et al. Strategies for sample labelling and library preparation in DNA metabarcoding studies. Mol Ecol Resour. 2021. doi:10.1111/1755-0998.13512
39. Gallo ND, Drenkard E, Thompson AR, Weber ED, Wilson-Vandenberg D, McClatchie S, et al. Bridging From Monitoring to Solutions-Based Thinking: Lessons From CalCOFI for Understanding and Adapting to Marine Climate Change Impacts. Front Mar Sci. 2019;6: 695. doi:10.3389/fmars.2019.00695
40. Thompson AR, Watson W, McClatchie S, Weber ED. Multi-scale sampling to evaluate assemblage dynamics in an oceanic marine reserve. Thrush S, editor. PLoS One. 2012;7: e33131. doi:10.1371/journal.pone.0033131
41. Miya M, Sato Y, Fukunaga T, Sado T, Poulsen JY, Sato K, et al. MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: Detection of more than 230 subtropical marine species. R Soc Open Sci. 2015;2: 150088. doi:10.1098/rsos.150088
42. Curd EE, Gold Z, Kandlikar GS, Gomer J, Ogden M, O'Connell T, et al. Anacapa Toolkit: An environmental DNA toolkit for processing multilocus metabarcode datasets. Methods Ecol Evol. 2019;10: 1469–1475. doi:10.1111/2041-210X.13214
43. Gold Z, Curd EE, Goodwin KD, Choi ES, Frable BW, Thompson AR, et al. Improving metabarcoding taxonomic assignment: A case study of fishes in a large marine ecosystem. Mol Ecol Resour. 2021;21: 2546–2564. doi:10.1111/1755-0998.13450
44. Collins RA, Trauzzi G, Maltby KM, Gibson TI, Ratcliffe FC, Hallam J, et al. Meta-Fish-Lib: A generalised, dynamic DNA reference library pipeline for metabarcoding of fishes. J Fish Biol. 2021;99: 1446–1454. doi:10.1111/jfb.14852
45. Shelton AO, Ramón-Laca A, Wells A, Clemons J, Chu D, Feist BE, et al. Environmental DNA provides quantitative estimates of Pacific hake abundance and distribution in the open ocean. Proceedings of the Royal Society B. 2022;289: 20212613.
46. Forootan A, Sjöback R, Björkman J, Sjögreen B, Linz L, Kubista M. Methods to determine limit of detection and limit of quantification in quantitative real-time PCR (qPCR). Biomol Detect Quantif. 2017;12: 1–6.
47. Lesperance ML, Allison MJ, Bergman LC, Hocking MD, Helbing CC. A statistical model for calibration and computation of detection and quantification limits for low copy number

environmental DNA samples. Environmental DNA. 2021;3: 970–981. doi:10.1002/EDN3.220

48. Abbott C, Coulson M, Gagné N, Lacoursière□Roussel A, Parent GJ, Bajno R, et al. Guidance on the Use of Targeted Environmental DNA (eDNA) Analysis for the Management of Aquatic Invasive Species and Species at Risk. Canadian Science Advisory Secretariat. Canadian Science Advisory Secretariat (CSAS); 2021. Available: https://escholarship.org/content/qt4ts0c9g8/qt4ts0c9g8_noSplash_e37874707346976a26b1c2022238d2df.pdf%0Ahttps://westernregionalpanel.org/wp-content/uploads/2021/04/Canada_eDNAGuidanceDoc.pdf
49. Hatzenbuhler C, Kelly JR, Martinson J, Okum S, Pilgrim E. Sensitivity and accuracy of high-throughput metabarcoding methods for early detection of invasive fish species. Sci Rep. 2017;7: 1–10.
50. Taylor SC, Nadeau K, Abbasi M, Lachance C, Nguyen M, Fenrich J. The ultimate qPCR experiment: producing publication quality, reproducible data the first time. Trends Biotechnol. 2019;37: 761–774.
51. Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, et al. Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. Mol Ecol. 2017;26: 5872–5895. doi:10.1111/mec.14350
52. Hänfling B, Handley LL, Read DS, Hahn C, Li J, Nichols P, et al. Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. Mol Ecol. 2016;25: 3101–3119. doi:10.1111/mec.13660
53. Moushomi R, Wilgar G, Carvalho G, Creer S, Seymour M. Environmental DNA size sorting and degradation experiment indicates the state of Daphnia magna mitochondrial and nuclear eDNA is subcellular. Sci Rep. 2019;9: 1–9.
54. Piñol J, Mir G, Gomez□Polo P, Agustí N. Universal and blocking primer mismatches limit the use of high□throughput DNA sequencing for the quantitative metabarcoding of arthropods. Mol Ecol Resour. 2015;15: 819–830.
55. Sipos R, Székely AJ, Palatinszky M, Révész S, Márialigeti K, Nikolausz M. Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. FEMS Microbiol Ecol. 2007;60: 341–350.
56. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics. 2012;13: 1–11.
57. Riaz T, Shehzad W, Viari A, Pompanon F, Taberlet P, Coissac E. ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. Nucleic Acids Res. 2011;39: e145–e145.
58. Sidstedt M, Rådström P, Hedman J. PCR inhibition in qPCR, dPCR and MPS—mechanisms and solutions. Anal Bioanal Chem. 2020;412: 2009–2023.
59. Fonseca VG, Nichols B, Lallias D, Quince C, Carvalho GR, Power DM, et al. Sample richness and genetic diversity as drivers of chimera formation in nSSU metagenetic analyses. Nucleic Acids Res. 2012;40: e66–e66.
60. Miya M, Gotoh RO, Sado T. MiFish metabarcoding: a high-throughput approach for simultaneous detection of multiple fish species from environmental DNA and other samples. Fisheries Science. 2020;86: 939–970. doi:10.1007/s12562-020-01461-x
61. Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, et al. 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. Front Zool. 2013;10: 34. doi:10.1186/1742-9994-10-34

62. Leese F, Sander M, Buchner D, Elbrecht V, Haase P, Zizka VMA. Improved freshwater macroinvertebrate detection from environmental DNA through minimized nontarget amplification. Environmental DNA. 2021;3: 261–276.

63. Dorn-In S, Bassitta R, Schwaiger K, Bauer J, Hölzel CS. Specific amplification of bacterial DNA by optimized so-called universal bacterial primers in samples rich of plant DNA. J Microbiol Methods. 2015;113: 50–56.

64. Nichols R v, Vollmers C, Newsom LA, Wang Y, Heintzman PD, Leighton M, et al. Minimizing polymerase biases in metabarcoding. Mol Ecol Resour. 2018;18: 927–939.

65. Doi H, Fukaya K, Oka S ichiro, Sato K, Kondoh M, Miya M. Evaluation of detection probabilities at the water-filtering and initial PCR steps in environmental DNA metabarcoding using a multispecies site occupancy model. Sci Rep. 2019;9: 3581. doi:10.1038/s41598-019-40233-1

66. Mata VA, Rebelo H, Amorim F, McCracken GF, Jarman S, Beja P. How much is enough? Effects of technical and biological replication on metabarcoding dietary analysis. Mol Ecol. 2019;28: 165–175. doi:10.1111/mec.14779

67. Krehenwinkel H, Wolf M, Lim JY, Rominger AJ, Simison WB, Gillespie RG. Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. Scientific Reports 2017 7:1. 2017;7: 1–12. doi:10.1038/s41598-017-17333-x

68. Valentini A, Taberlet P, Miaud C, Civade R, Herder J, Thomsen PF, et al. Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. Mol Ecol. 2016;25: 929–942. doi:10.1111/mec.13428

69. Harrison JB, Sunday JM, Rogers SM. Predicting the fate of eDNA in the environment and implications for studying biodiversity. Proceedings of the Royal Society B: Biological Sciences. 2019;286: 20191409. doi:10.1098/rspb.2019.1409

Supplement 1: A note on sampling depth and the probability of observing zeros.

In the metabarcoding literature, it is often asserted that if more reads are sampled, it is more likely that rare variants will be observed. While this is true, the magnitude of the effect is likely less than one would like and only meaningfully changes the probability for a relatively narrow range of rare things. This supplement provides some simple examples for how to do that calculation.

Let us focus on only the last step of the metabarcoding process: multinomial sampling. As we are interested in isolating the contribution of sampling depth we can assert that everything prior to the sampling of DNA strands is identical; only the sampling depth changes. So for illustration, let's discuss a single taxon ("A") that comprises 0.0001% of the DNA (1 in 10,000 sequences) of the post-PCR product. We will assume that multinomial sampling is a decent approximation to the process (i.e., there are so many DNA copies floating around removing a few doesn't materially change the probability of observing a given taxon; for those who are uncomfortable with this assumption, the below can be reframed using the hypergeometric distribution in place of the multinomial). For a single taxon, the multinomial collapses to the binomial (i.e., we can think of many taxa collapsing to two groups: taxon A and not taxon A). The probability mass function for the binomial is:

$$p(k|\pi,n) = \frac{n!}{k!(n-k)!}\pi^k(1-\pi)^{n-k} \quad (1)$$

where k is the number of "successes" (observations of taxon A by the sequencer) and *n* is the number of sequences read. We are interested in a single value here: what is the probability of *k=0* (i.e. taxon A was not observed) as the number of sequences examined (*n*) increases. First, simplify for the case *k=0*

$$p(k=0|\pi,n) = (1-\pi)^n \quad (2)$$

then plug in $\pi = 0.0001$ a range of values for sampling depth(below I use 10 thousand, 100 thousand, and 1 million reads):

$$p(k = 0|\pi = 0.0001, n = 10{,}000) = (1 - 0.0001)^{10000} = 0.367861 \quad (3)$$

$$p(k = 0|\pi = 0.0001, n = 100{,}000) = (1 - 0.0001)^{100000} = 0.00045377$$

(4)

$$p(k = 0|\pi = 0.0001, n = 1{,}000{,}000) = (1 - 0.0001)^{1000000} = 3.7 \times 10^{-44}$$

(5)

So for a species that is rare we go from seeing 1 or greater sequences with probability of 0.64 (1-0.36) at 10,000 reads to seeing it with almost certainty at 100,000 or more reads. Let's do the calculation for a rarer sequence.

$$p(k = 0|\pi = 0.000001, n = 10{,}000) = (1 - 0.000001)^{10000} = 0.99 \quad (6)$$

$$p(k = 0|\pi = 0.000001, n = 100{,}000) = (1 - 0.000001)^{100000} = 0.90 \quad (7)$$

$$p(k = 0|\pi = 0.000001, n = 1{,}000{,}000) = (1 - 0.000001)^{1000000} = 0.367$$

(8)

So for a sequence that occurs at a rate of 1 in a million you go from observing 0 99% of the time with a sampling depth of 10,000 reads to 90% at 100 thousand reads to only 36% at 1 million reads.

Going the other direction, let's look at something that is more common, say 1 in 1,000:

$$p(k = 0|\pi = 0.001, n = 10{,}000) = (1 - 0.001)^{10000} = 0.00045377 \quad (9)$$

$$p(k = 0|\pi = 0.001, n = 100{,}000) = (1 - 0.001)^{100000} = 3.7 \times 10^{-44} \quad (10)$$

$$p(k = 0|\pi = 0.001, n = 1{,}000{,}000) = (1 - 0.001)^{1000000} = 0 \quad (11)$$

Thus, you are almost certain to see at least one copy at 10,000 or greater read depths.

So what does this mean in general? Basically, you will see rarer things at higher read depths but moving from a read depth of say 10,000 to 1 million will only meaningfully change non-detection of very rare sequence variants (sequences that make up somewhere between 1 in 10,000 and 1 in 1 million copies). If you think that there are a lot of taxa that you care about are in this very rare zone, it may make sense to do more sequencing. But things that are extremely rare (occur at a frequency of less than 1 in a million) still will not be detected. Note that things can be rare after PCR because they are rare in the sample or because they are poor amplifiers, or both. One caveat to the description above is it only includes the probability of observing exactly zero. Many researchers use a higher threshold to determine presence (say $k > 10$, for example). Calculating $k > K$ is not quite as easy as $p(k=0)$ in that more terms are involved, but it is certainly not a hard calculation and involves summing the probability of $k=0$ to $k=K$,

$$p(k \geq K|\pi, n) = 1 - \sum_{k=0}^{k=K} \frac{n!}{k!(n-k)!} \pi^{k} (1-\pi)^{n-k} \quad (12)$$

Supplemental 2 Alternate simulation results

Changing N_pcr1

It is important to understand how changing some of the parameters in the simulation affect the probability of non-detection. In Fig. S3.1 we used $N_{pcr1} = 20$ rather than $N_{pcr1} = 35$ presented in the main text. As N_{pcr1} declines, the probability of non-detects becomes more similar among species and only species with amplification efficiencies that are much lower than the average a (in this case $a_i < 0.4$) have increased non-detection probabilities (Fig. S3.1A).

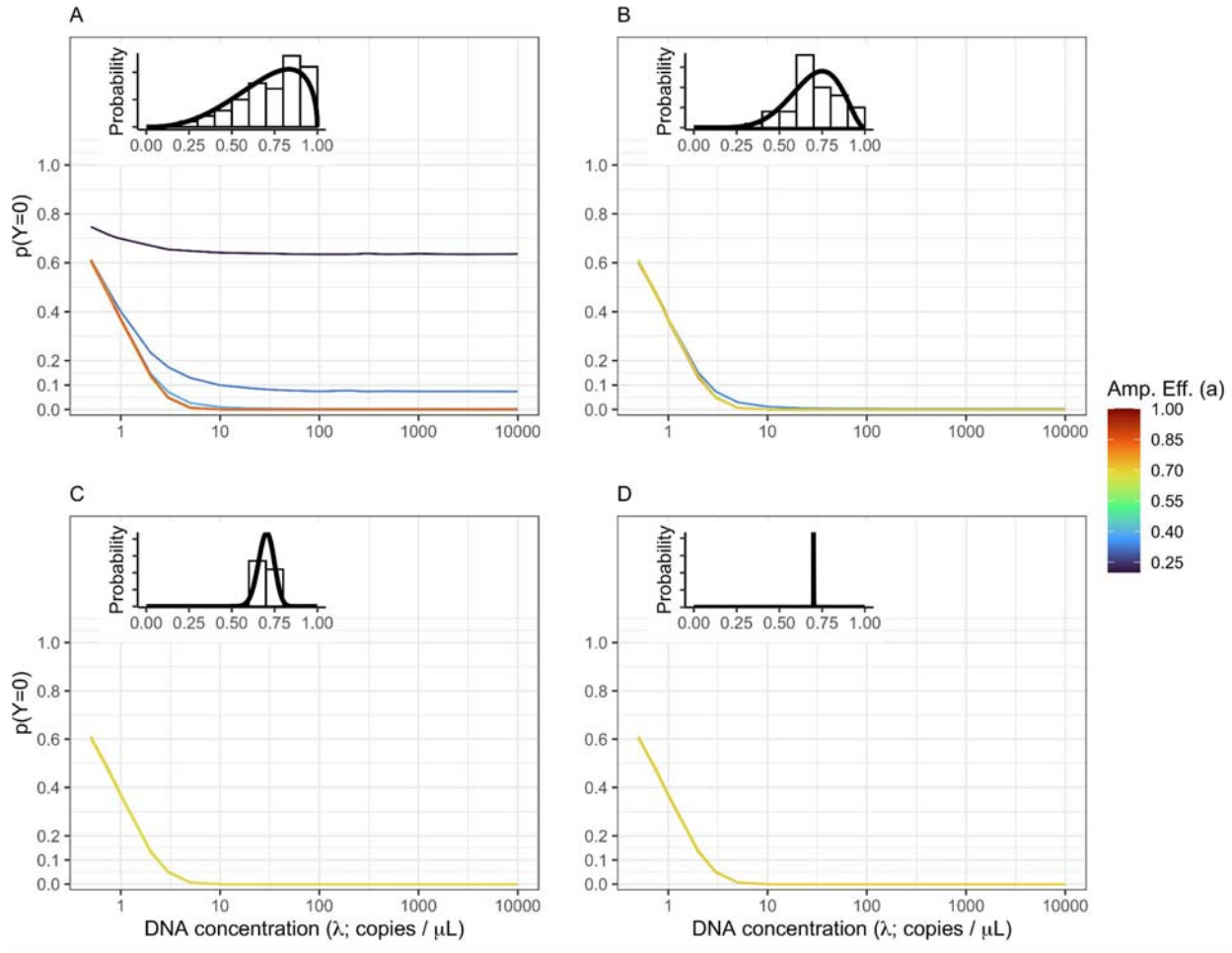

Figure S3.1. Non-detects Driven By Both DNA Concentration and Amplification Efficiency.

The probability of non-detection (*p(Y=0)*) is shown for a community of 50, equally abundant taxa with the amplification efficiency distribution shown inset in each panel. This simulation uses $N_{pcr1} = 20$ (see Fig. 1 for the same simulation but with $N_{pcr1} = 35$). The amount of among-taxa variation in amplification efficiency varies from highly variable *(A;* γ=5) to moderate variation (*B:* γ=10) to low variation (*C:* γ=100) to effectively no variation (*D:* γ=1,000,000). Both subsampling and amplification efficiencies influence the rate of non-detection. The probability of observing no DNA in a given technical replicate is highest at low DNA concentrations (<10 copies /µL). However, non-detects are possible for species with low amplification efficiencies and very likely (*p(Y=0)* > 0.5) for amplification well below average (in this case approximately $a_i < 0.3$ *)*.

Simulating uneven DNA concentrations

The base simulation presented in the main text assumes that the starting DNA concentration for each taxon is equivalent (i.e., for ten taxa, each comprises 10% of the DNA in a sample). While this assumption makes it easier to visualize the simulation results, it clearly does not represent natural communities which have skewed abundance distributions (some taxa are common while others are rare). To illustrate the consequences of a skewed abundance distribution we simulated a community of 20 taxa with 2 taxa each comprising 20% of the DNA, 8 taxa each with 5% of the DNA, and 10 taxa with 2% of the DNA. Otherwise, we followed the simulation parameters described in the main text. Figure S3.2 presents the patterns of non-detections for a single community of 20 taxa (Figure S3.2A) with large among-species variation in amplification efficiency ($\gamma = 5$) and for 20 communities of 20 taxa each overlaid on one figure (Fig. S3.2B).

838 Facets show the true starting proportion within each community (proportions of 0.02, 0.05, or
839 0.20).

840 As shown in the even community simulated in the main text, for all taxa non-detection
841 increases as DNA concentration declines and taxa with lower amplification rates show higher
842 probability of non-detection. But there is clearly an interaction between the community
843 proportion and amplification efficiency which affects the probability of non-detection.
844 Specifically, for two taxa with equivalent amplification efficiencies, the more abundant taxa
845 (community proportion of 0.20) have a much lower probability of non-detection than a relatively
846 rare species (community proportion of 0.02; Fig. S3.2B). Indeed, for taxa with a community
847 proportion of 0.02, at a constant DNA concentration, $p(Y = 0|\lambda{=}10) > 0.5$ when $a_i < 0.45$. In
848 contrast, for taxa with community proportions of 0.20, $p(Y = 0|\lambda{=}10) > 0.5$ only occurred for
849 one taxa in the 20 simulated communities with a very low amplificiation efficiency ($a_i = 0.19$).

850 Thus both community proportion and amplification efficiency affect the probability of
851 non-detection. In broad strokes, amplification efficiency will play a more important role in
852 determining non-detection when taxa are rare relative to other species in a sample. The
853 importance of amplification efficiency increases with PCR protocols that use a large number of
854 PCR cycles. Non-detection of relatively common taxa in a community will generally be less
855 influenced by relative amplification efficiency, but non-detection can still occur if amplification
856 efficiency is sufficiently low.

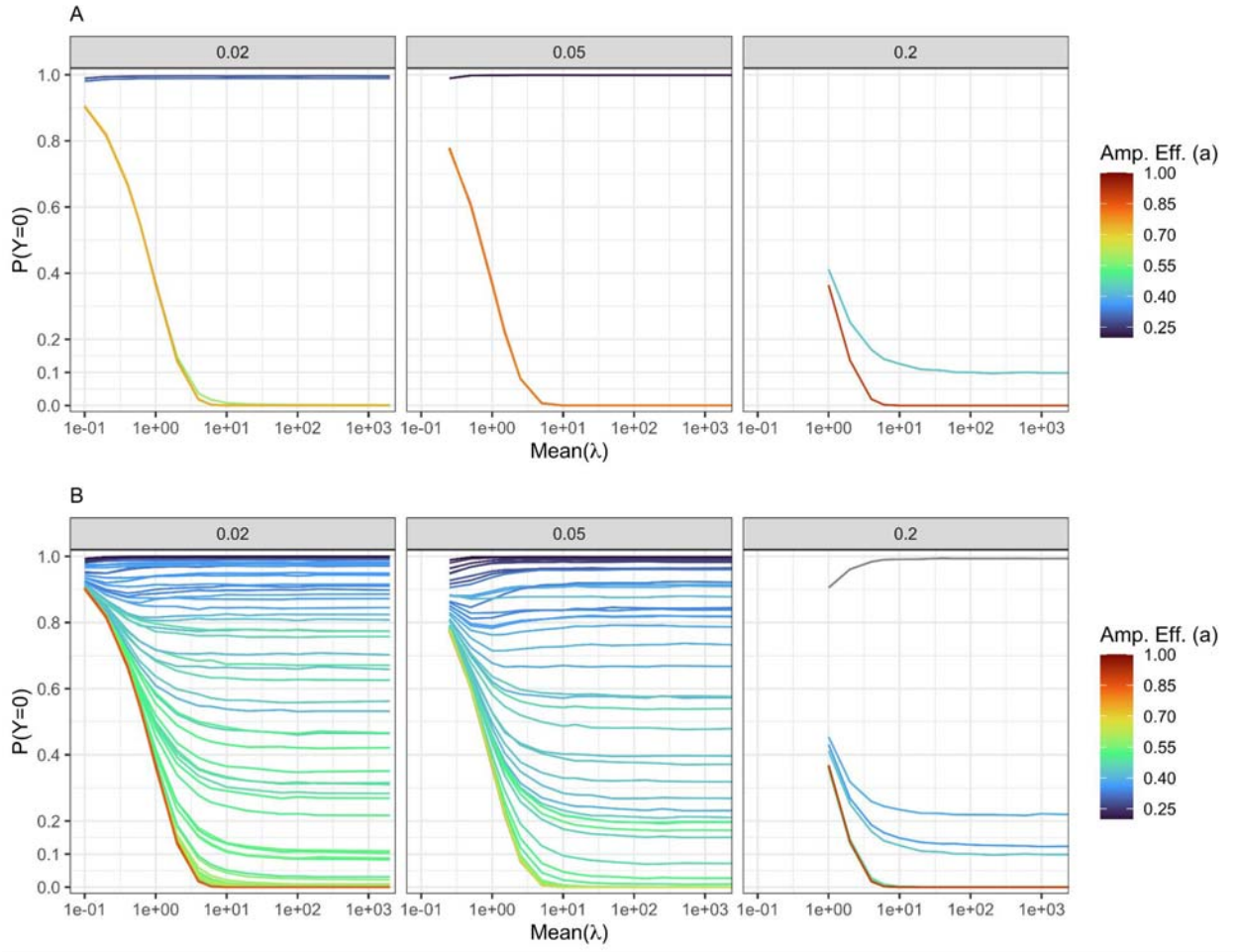

Figure S3.2. Non-detects Driven By Both DNA Concentration and Amplification Efficiency.

The probability of non-detection (*p(Y=0)*) is shown for a community of 20 taxa with 4 taxa comprising 0.20 of the initial DNA, 8 with 0.05 of the DNA, and 10 species comprising 2% of the DNA across a range of initial DNA concentrations. *A:* Presents results for a single 20 taxa community with facets representing the three abundance categories. *B* shows results for 20 communities of 20 taxa each to illustrate general patterns. For all simulations we use $N_{pcr1} = 35$ and a fixed amount of among-taxa variation in amplification efficiency ($\gamma = 5$). Clearly, relative abundance influence the rate

of non-detection with relatively rare taxa (those with 0.02 having larger probabilities of non-detection than common taxa (0.2) with equivalent amplification efficiencies (colors).