PLOS ONE

RESEARCH ARTICLE

Validating metabarcoding-based biodiversity assessments with multi-species occupancy models: A case study using coastal marine eDNA

Beverly McClenaghan[1], **Zacchaeus G. Compson**[1], **Mehrdad Hajibabaei**[1,2,3]*

1 Centre for Environmental Genomics Applications, eDNAtec Inc., St. John's, NL, Canada, **2** Department of Integrative Biology, University of Guelph, Guelph, ON, Canada, **3** Centre for Biodiversity Genomics, University of Guelph, Guelph, ON, Canada

* hajibabaei@gmail.com

Citation: McClenaghan B, Compson ZG, Hajibabaei M (2020) Validating metabarcoding-based biodiversity assessments with multi-species occupancy models: A case study using coastal marine eDNA. PLoS ONE 15(3): e0224119. https://doi.org/10.1371/journal.pone.0224119

Editor: Hideyuki Doi, University of Hyogo, JAPAN

Received: October 4, 2019

Accepted: February 16, 2020

Published: March 19, 2020

Data Availability Statement: All sequencing data have been deposited into NCBI's Sequence Read Archive under accession number PRJNA574050.

Funding: This work was partly funded through a Petroleum R&D Grant from InnovateNL (contract number 5405.2121.101), an award from the Atlantic Canada Opportunities Agency's Atlantic Innovation Fund (project number 781-37749-207993), and a grant from Petroleum Research Newfoundland and Labrador. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the

Abstract

Environmental DNA (eDNA) metabarcoding is an increasingly popular method for rapid biodiversity assessment. As with any ecological survey, false negatives can arise during sampling and, if unaccounted for, lead to biased results and potentially misdiagnosed environmental assessments. We developed a multi-scale, multi-species occupancy model for the analysis of community biodiversity data resulting from eDNA metabarcoding; this model accounts for imperfect detection and additional sources of environmental and experimental variation. We present methods for model assessment and model comparison and demonstrate how these tools improve the inferential power of eDNA metabarcoding data using a case study in a coastal, marine environment. Using occupancy models to account for factors often overlooked in the analysis of eDNA metabarcoding data will dramatically improve ecological inference, sampling design, and methodologies, empowering practitioners with an approach to wield the high-resolution biodiversity data of next-generation sequencing platforms.

Introduction

Environmental DNA (eDNA) as a signal for diversity detection is rapidly advancing. In freshwater systems, in particular, eDNA is now used as a bioassessment tool in both single-species qPCR-based studies and in sequencing-based metabarcoding community assessments [1–3]. Approaches based on eDNA are also gaining traction in the marine environment [4,5]. Oceans are complex, highly diverse, and difficult to sample; therefore, identifying organisms from all trophic levels and taxonomic groups from a single survey method will greatly facilitate rapid, consistent biodiversity surveys [6]. eDNA metabarcoding provides a streamlined method of biodiversity assessment, generating high-resolution biodiversity data with time and effort savings during sample collection and analysis [7,8].

However, there are several levels of uncertainty associated with eDNA sampling for community assessments. The potential for false negatives during sampling, where a species present in the environment is not detected in surveys, can bias results [9]. False negatives can occur

authors and do not necessarily reflect the views of Petroleum Research or its members. B.M. and Z.G.C. are employees of eDNAtec Inc. and M.H. is the founder and Chief Scientific Officer of eDNAtec Inc. The compensations for authors B.M., Z.G.C. and M.H. were supported by the funder, but the funder did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Competing interests: B.M. and Z.G.C. are employees of eDNAtec Inc. and M.H. is the founder and Chief Scientific Officer of eDNAtec Inc. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

during field sampling and during lab processing. If imperfect detection is not accounted for, this could lead to biased estimates of species richness and individual species occupancy [10,11]. Accounting for false negatives will improve community-wide species occurrence estimates based on eDNA surveys and yield more robust ecological conclusions for making management decisions and informing sampling designs. Optimal sampling designs for eDNA metabarcoding studies are not well-established and differ from traditional ecological sampling methods in the cost and effort required for sample collection [12]. Additionally, there are several added variables that need to be accounted for in metabarcoding studies compared to traditional sampling approaches, such as sequencing depth and marker selection, which vary between studies and can affect metabarcoding results [5,13,14]. Sampling designs should be experimentally informed and optimized specifically for eDNA metabarcoding methods [15], yet this is seldom practiced, and these added sources of variation during sample processing are seldom considered in the same analysis as sampling design.

Occupancy modelling is a powerful tool to account for the additional sources of variation associated with next-generation biomonitoring approaches, and it has been used to assess imperfect detection in terrestrial bioassessment [16–18]. These models include 2-levels: the probability that a species occurs at a site (occupancy; ψ) and the probability of detecting a species at a site (probability of detection; p). Recently, occupancy models have been adapted for single-species eDNA studies by adding an additional stochastic level, the probability of capture (θ). Occupancy refers to the probability of a species' DNA occurring at a site, the probability of capture (θ) refers to the probability of capturing a species' eDNA in a field sample, and the probability of detection refers to the probability of detecting a species in a PCR replicate [19,20]. The use of occupancy models in single-species eDNA studies is not ubiquitous, but it is increasing [21].

Occupancy modelling can also be applied to whole communities through multi-species occupancy models, which are commonly applied to traditional surveys in terrestrial systems [22,23], yet seldom used in the context of DNA metabarcoding (S1 Table). In the same way that single-species models were adapted for eDNA studies through the inclusion of an additional stochastic level, multi-species models can be adapted for metabarcoding by including this additional level. Modeling communities together in a single multi-species model can improve the accuracy and predictive ability of occupancy models compared to single-species models [24]. The application of multi-species, multi-scale occupancy models to metabarcoding data is rare, focusing on small-scale lab manipulations [25], and no studies have implemented this modelling approach to improve sampling designs in natural systems (but see [26] for a single species example). Incorporating these models routinely in metabarcoding analysis will improve ecological inferences and species richness estimates, as well as facilitate the development of robust sampling designs for a relatively new technique where little thought has been dedicated to developing de novo sampling designs distinct from traditional sampling methods. The inclusion of covariates in occupancy models at each process level extends the application of the model, enabling discrimination between sources of variation in sampling effort and environmental factors. However, making conclusions based on models with covariates requires methods of model assessment and selection for multi-species, multi-scale models.

Here, we demonstrate how multi-species occupancy modelling can be used for the analysis of community biodiversity data resulting from eDNA metabarcoding and highlight the potential of these models for both improving methodologies and sound ecological inference. We present methods for model assessment and model comparison adapted for multi-scale, multi-species occupancy models. Finally, we demonstrate how these tools can improve inferential power from eDNA metabarcoding results using a case study in a coastal, marine environment.

Materials and methods

Model formulation

The multi-species, multi-scale occupancy model. We used a Bayesian modeling framework to develop a multi-species, hierarchical occupancy model with three stochastic levels: occupancy (ψ), probability of capture (θ), and probability of detection (p) (Fig 1). The occupancy process describes whether sampling sites are occupied or not by a given species' DNA. For eDNA sampling, there are often two levels of sampling replication within each site (e.g. [20,27]): biological replicates are samples collected from a single site in the field and technical replicates are repeated samples taken from a single biological replicate in the lab. The probability of capture refers to the probability that a species' DNA is collected in a sample, given that the species was present at the site. The probability of detection refers to the probability that a species was detected in a technical replicate, given that the species' DNA was collected in the sample. This model assumes no false positives occur in the data. While false positives may be a possibility in metabarcoding data [15], we used strict bioinformatic filtering to reduce this possibility (see *Bioinformatics* below). Further comments on false positives can be found in the *Discussion*.

This model can be fit to a dataset, y_{ijrk}, which is a binary indicator of whether a species k ($k = 1,2,\ldots K$) was detected (1) or not detected (0) in a technical replicate r ($r = 1,2,\ldots R$) from a given sample j ($j = 1,2,\ldots J$) at a given site i ($i = 1,2,\ldots I$). The model consists of three coupled Bernoulli trials to describe a four-dimensional array of data y_{ijrk}.

$$y_{ijrk}|w_{ijk} \sim Bernoulli(p_{ijrk}w_{ijk})$$

$$w_{ijk}|z_{ik} \sim Bernoulli(\theta_{ijk}z_{ik})$$

$$z_{ik} \sim Bernoulli(\psi_{ik})$$

We model our observations of detection ($y_{ijrk} = 1$) or non-detection ($y_{ijrk} = 0$) of species k in replicate r in sample j at site i as a random variable having parameter p_{ijrk}, which describes the probability of detection. The second random variable, w_{ijk}, describes the capture ($w_{ijk} = 1$) or non-capture ($w_{ijk} = 0$) of species k in sample j at site i having parameter θ_{ijk} (probability of capture). The third random variable, z_{ik}, describes the occurrence ($z_{ik} = 1$) or non-occurrence ($z_{ik} = 0$) of species k at site i having parameter ψ_{ik} (occupancy probability).

Model assessment and comparison. The goodness-of-fit of multi-species, multi-scale occupancy models can be assessed using Bayesian p-values of the deviance residuals. We adapted Bayesian p-value calculations from [28] for a multi-scale model (S1 File) to assess goodness-of-fit, where values close to 0.5 indicate a good fit and values >0.95 or <0.05 indicate a poor fit.

We also adapted model selection and cross-validation calculations from [28] for multi-scale, multi-species occupancy models to determine the best model. We calculated the Watanabe-Akaike information criterion (WAIC; [29]) and the conditional predictive ordinate criterion (CPO; [30]), and then evaluated the results of k-fold cross validation using the Brier score and the logarithmic score. Complete calculations for all model assessment and comparison methods can be found in S1 File.

Unknown species richness. An additional level can be added to the model described above for communities with unknown species richness [10]. This model uses data augmentation to estimate species richness for the sampling area through the inclusion of another

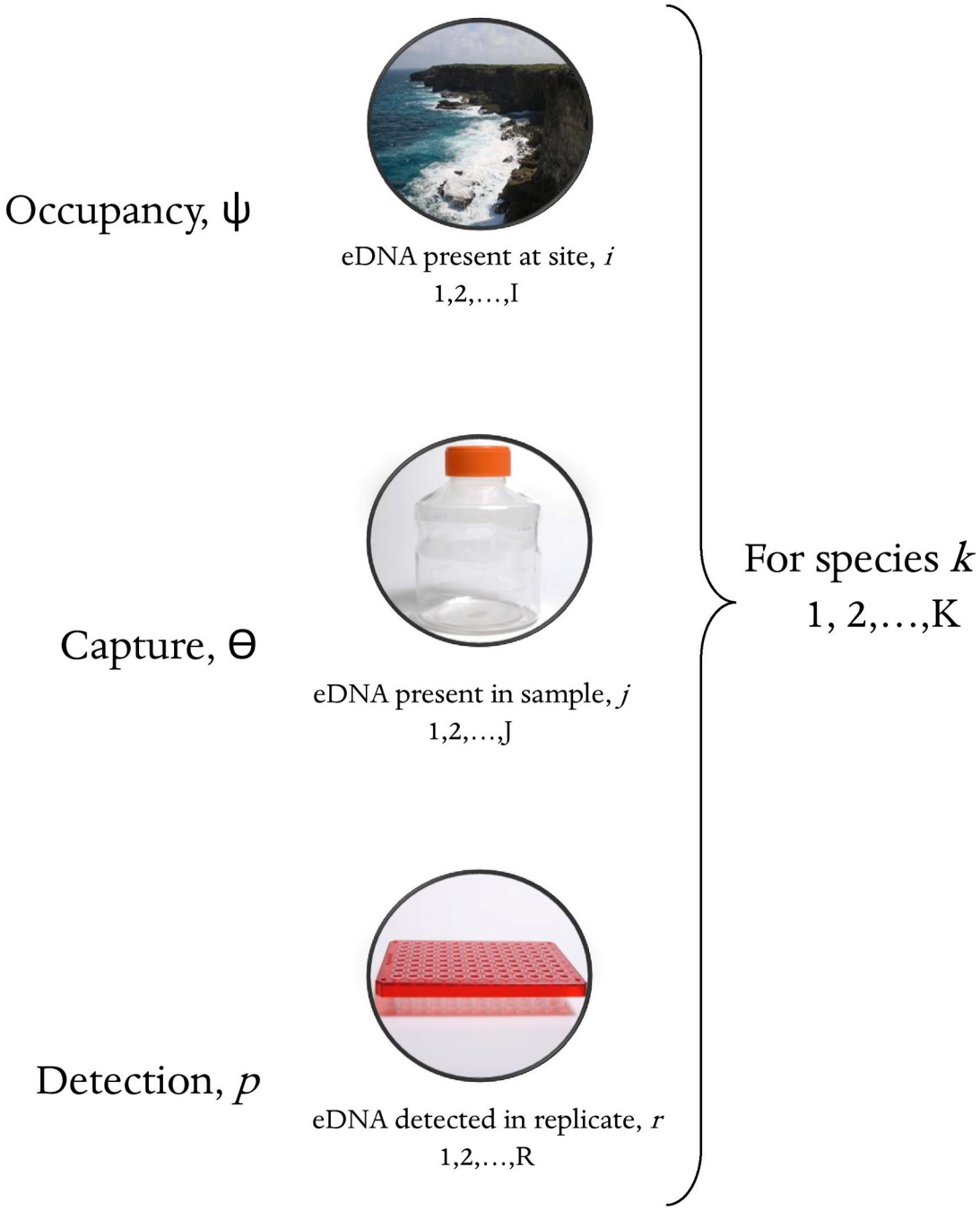

Fig 1. Schematic illustration of the three stochastic levels included in the multi-scale, multi-species occupancy model.

https://doi.org/10.1371/journal.pone.0224119.g001

Bernoulli variable:

$$a_k \sim Bernoulli(\Omega)$$

$$\Omega \sim Uniform(0, 1)$$

An upper limit to species richness (M) is specified a priori and considered large enough when the estimate of true species richness is sufficiently lower than M (i.e., the value of M is in the right tail of the posterior distribution of species richness; [31]). For species k (k = 1,2,. . .M), a_k = 1 if species

k is present in the metacommunity. Here, the random variable z_{ik} describes the occurrence (z_{ik} = 1) or non-occurrence (z_{ik} = 0) of species k at site i having parameter ψ_{ik}, which describes the occupancy probability for species present in the metacommunity.

$$z_{ik} \sim Bernoulli(\psi_{ik} a_k)$$

Species may be present in the metacommunity but be unobserved during surveys due to imperfect capture and detection. Summing a_k provides an estimate of species richness for the metacommunity, including observed and unobserved species.

Case study: Conception Bay, Newfoundland

Sample collection, processing and sequencing. Triplicate 250 mL water samples were collected from coastal surface water at eight sites along two transects in Conception Bay, Newfoundland and Labrador, Canada, on October 13–14, 2017 (see [5] for sampling details). No permits were required to collect samples since there are no regulations on collecting seawater. Water samples were filtered using 0.22 μm PVDF Sterivex filters (MilliporeSigma) and DNA was extracted from filter membranes using the DNeasy PowerWater Kit (Qiagen). Five target amplicons in the cytochrome *c* oxidase I (COI) region were amplified by PCR from each sample. Table 1 details the primer sets used to target these amplicons. Three PCR replicates were performed for each amplicon from each sample and then PCR replicates were pooled into a single PCR cleanup for each of the five amplicons with the QIAquick 96 PCR purification kit (Qiagen). Amplicons were then indexed using unique dual Nextera indexes (IDT). All amplicons were pooled into one library to normalize DNA concentration and the library was sequenced with a 300-cycle S4 kit on the NovaSeq 6000 following the NovaSeq XP workflow. Raw sequence reads are available in NCBI's sequence read archive under accession number PRJNA574050. Primers were trimmed from sequences and then DADA2 v1.8.015 [32] was used for quality filtering, joining paired end reads and denoising to produce exact sequence variants (ESVs). Taxonomy was assigned using NCBI's blastn tool v2.6.026 [33] to compare ESV sequences against the nt database. See [5] for detailed sampling, sequencing, and bioinformatic methodology.

Occupancy model implementation. Under the occupancy modelling framework described above, each collection site along each transect in Conception Bay was considered a different site in the occupancy model. Replicate bottles collected at a site were considered samples. Each of the five amplicons sequenced from each bottle was considered a technical replicate. While we conducted three replicate PCRs for each amplicon, PCR products were pooled for each amplicon prior to sequencing so we did not include PCR replicates for each amplicon separately in our models. However, PCR replicates can easily be accommodated in a multi-scale, multi-species occupancy modelling framework, such as the model described here.

Table 1. Primer pairs used to amplify five target amplicons in the COI region of the mitochondrial genome from water samples collected in Conception Bay, Newfoundland, Canada.

Marker	Target Length (bp)	Forward Primer	Reverse Primer	Reference
Fishe (Mini_SH-E)	226	5'-CACGACGTTGTAAAACGACACYAAICAYAAAGAYATIGGCAC-3'	5'-GGATAACAATTTCACACAGGCTTATRTTRTTTATICGIGGRAAIGC-3'	[34]
Fishc (Mini_SH-C)	127	5'-CACGACGTTGTAAAACGACACYAAICAYAAAGAYATIGGCAC-3'	5'-GGATAACAATTTCACACAGGGAARATCATAATGAAGGCATGIGC-3'	[34]
F230	235	5'-GGTCAACAAATCATAAAGATATTGG-3'	5'-CTTATRTTRTTTATNCGNGGRAANGC-3'	[35]
Leray	330	5'-GGWACWGGWTGAACWGTWTAYCCYCC-3'	5'-TAAACTTCAGGGTGACCAAAAAATCA-3'	[36]
BR5	310	5'-CCIGAYATRGCITTYCCICG-3'	5'-GTRATIGCICCIGCIARIACIGG-3'	[37]

https://doi.org/10.1371/journal.pone.0224119.t001

We included sequencing depth (number of reads per sample per amplicon) as a continuous covariate (α1) at the level of probability of detection. Additionally, we included amplicon identity as a categorical covariate (α2) at the level of probability of detection. We included water depth (m) as a continuous covariate (α3) at the level of occupancy. Covariates were included in the model as follows:

$$logit(p_{ijrk}) = lp_k + \beta 1_k * \alpha 1_{ijr} + \beta 2_{k\,\alpha 2_{ijr}}$$

$$logit(\psi_{ik}) = lpsi_k + \beta 3_k * \alpha 3_i$$

Continuous covariates were z-score standardized to have a mean of zero and a standard deviation of one to help with model convergence. We compared a null model (Model 1) with no covariates with seven models with different combinations of covariates (Table 2) using WAIC and CPO and using Brier and logarithmic scores for cross-validation. We assessed model fit using Bayesian p-values based on deviance residuals and by looking at diagnostic plots to examine model fit. We plotted deviance residuals for each species, site, and covariate. Species coefficients arise from additional community-level parameters:

$$lpsi_k \sim N(\mu_{lpsi}, \sigma_{lpsi})$$

$$ltheta_k \sim N(\mu_{ltheta}, \sigma_{ltheta})$$

$$lp_k \sim N(\mu_{lp}, \sigma_{lp})$$

$$\beta 1_k \sim N(\mu_{\beta 1}, \sigma_{\beta 1})$$

$$\beta 2_{k\,\alpha 2} \sim N(\mu_{\beta 2}, \sigma_{\beta 2})$$

$$\beta 3_k \sim N(\mu_{\beta 3}, \sigma_{\beta 3})$$

Community-level parameters were described by weakly informative hyperpriors [31]. All mean values for the above prior distributions were selected from a normal distribution and all standard deviations were selected from a uniform distribution.

$$\mu \sim N(0, 10)$$

$$\sigma \sim Uniform(0, 5)$$

Prior sensitivity was assessed by running the models with various prior parameterizations. Posterior distributions were similar across all priors.

All statistical analyses were conducted in R v3.5.1 [38]. MCMC sampling was achieved with JAGS [39], implemented using '*jagsUI*' v1.5.0 [40]. The model was written for JAGS in the BUGS language (see S2 File for BUGS model structure). We fit models using known species richness to conduct our model comparisons, and assessed models and model fit to determine the best model. MCMC sampling was run in three chains, each with 50,000 iterations, a burn in of 10,000, and a thinning rate of 10. Convergence was verified using the Gelman-Rubin diagnostic [41] and by evaluating trace plots. For all models, we report parameter estimates as the mean of the posterior distribution with the 95% highest posterior density interval (HDI; [42]) calculated using 'HDInterval' v0.2.0 [43]. We conducted a data augmented model with unknown species richness for the best model at varying levels of augmentation to determine

the minimal level of augmentation required, as described above in the *Unknown Species Richness* section. Significance of continuous covariates was assessed by determining if the 95% confidence intervals of parameter estimates overlapped with zero [31].

To investigate the effects of phylum on the probability of detection of each amplicon, we ran one additional model (Model 9), which included amplicon as a categorical covariate (α2) and group-level effects at the species-level following [31]. In this model, we only included metazoan phyla where at least 2 species were detected. Here, species coefficients arise from community-level parameters that vary by phylum:

$$lpsi_k \sim N(\mu_{lpsi[phylum]}, \sigma_{lpsi[phylum]})$$

$$ltheta_k \sim N(\mu_{ltheta[phylum]}, \sigma_{ltheta[phylum]})$$

$$lp_k \sim N(\mu_{lp[phylum]}, \sigma_{lp[phylum]})$$

$$\beta 1_k \sim N(\mu_{\beta 1[phylum]}, \sigma_{\beta 1[phylum]})$$

$$\beta 2_{k\,\alpha 2} \sim N(\mu_{\beta 2[phylum]}, \sigma_{\beta 2[phylum]})$$

$$\beta 3_k \sim N(\mu_{\beta 3[phylum]}, \sigma_{\beta 3[phylum]})$$

Results

We ran eight multi-species, multi-scale occupancy models with different combinations of covariates (i.e., water depth at the level of occupancy, sequencing depth and amplicon at the

Table 2. Model comparison between multi-scale, multi-species occupancy models using four methods (WAIC, CPO, Brier score and Log Score). The covariates (water depth at the sampling site, sequencing depth for each technical replicate, and amplicon sequenced for each technical replicate) included at each level of the model (occupancy: ψ, capture: θ, detection: p) are listed on the left. Bolded values indicate the best model for each method of model comparison.

MODELS	WAIC	CPO	Brier Score	Log Score
Model 1 $\psi(.)\ \theta(.)\ p(.)$	13,340	5,621,109	**98**	1842
Model 2 ψ(water depth) $\theta(.)$ p(sequencing depth, amplicon)	33,626	14,217,497	217	2034
Model 3 $\psi(.)\ \theta(.)\ p$(sequencing depth)	12,993	4,061,736	**98**	1834
Model 4 $\psi(.)\ \theta(.)\ p$(amplicon)	35,894	20,793,808	135	2823
Model 5 ψ(water depth) $\theta(.)$ $p(.)$	13,135	2,748,236	143	1759
Model 6 ψ(water depth) $\theta(.)$ p(amplicon)	35,649	17,897,824	224	2732
Model 7 ψ(water depth) $\theta(.)$ p(sequencing depth)	**12,869**	**1,697,452**	142	**1753**
Model 8 $\psi(.)\ \theta(.)\ p$(sequencing depth, amplicon)	33,499	16,555,815	133	2405

https://doi.org/10.1371/journal.pone.0224119.t002

level of detection probability) and assessed these models using model comparison and cross-validation methods adapted for this multi-scale approach (Table 2). Three of the model comparison methods (WAIC, CPO and one cross-validation score) were in agreement that Model 7 (ψ*(water depth)* θ*(.)* *p(sequencing depth)*) was the best model, while the Brier score from cross-validation suggested Model 1 (the null model) and Model 3 (ψ*(.)* θ*(.)* *p(sequencing depth)*) were the best models. We considered Model 7 our best model moving forward, given that most selection methods indicated this was the best model.

We assessed model fit using Bayesian *p*-values and diagnostic plots for all models but present the results for the best model only. We obtained a Bayesian *p*-value of 0.51, suggesting that Model 7 (ψ*(water depth)* θ*(.)* *p(sequencing depth)*) provided a good fit to our data overall; diagnostic plots, however, revealed higher deviance at sites with lower water depth, suggesting a poorer model fit at shallower sites (see S3 File). The community-wide estimate for occupancy was 0.29 (HDI: 0.22–0.36). Water depth had a significant effect on the community mean occupancy (Fig 2A and 2B), and we detected considerably more species at the shallowest sites compared to the other sites (274 species at two shallow water sites combined compared to 109 species across all six deep water sites). The community-wide probability of capture was 0.96 (HDI: 0.92–0.99) and the community-wide probability of detection was 0.14 (HDI: 0.12–0.17). Sequencing depth did not have a significant effect on the probability of detection for most species in this case study (Fig 2C and 2D). Species-specific estimates of occupancy, capture probability, and detection probability were also obtained from the model (S2 Table).

We estimated species richness for the survey area by running the best model (Model 7) with data augmentation. This model used the probabilities of capture and detection to estimate the number of species missed in sampling efforts. We detected 231 species overall, and the estimated species richness for the survey area was 286 (HDI: 264–306), indicating that 55 (HDI: 33–75) species were undetected during our surveys. In other words, our survey detected ~81% of the estimated species in our study area.

While it was not selected as our best model, we also ran Model 9 (ψ*(.)* θ*(.)* *p*(*amplicon*)) with a group-level effect (phylum) to investigate how the probability of each amplicon varied by phylum. Amplicons displayed relatively similar probabilities of detection at the community-level (Fig 3), however the probabilities of detection for each amplicon varied considerably when comparing between phyla, where some amplicons clearly failed to detect certain taxonomic groups (Fig 4).

Discussion

We applied a multi-species, multi-scale occupancy model to a DNA metabarcoding dataset generated from marine water samples and explored how the inclusion of categorial and continuous covariates at different levels improved model performance. The best model included sequencing depth as a covariate at the level of detection and water depth as a covariate at the level of occupancy, where we observed a higher species richness at shallower sites. One of the shallow water collection sites was within 1 km of a sewage outflow, which may have contributed to this result, although a high species richness was also observed at the second, shallow water site located >10 km from the sewage outflow. While sequencing depth was included in the best model, we did not observe a strong effect of sequencing depth. However, the samples were all sequenced on a NovaSeq instrument, which generates an unprecedented number of reads, yielding very high sequencing depths (mean number of filtered sequences per sample ± standard deviation: 8,519,055 ± 2,514,998) compared to many other barcoding studies (e.g. [44,45]). In studies where the mean sequencing depth is lower, differences in sequencing depth are likely to have greater effects [5,46]. In such cases, analyzing data using

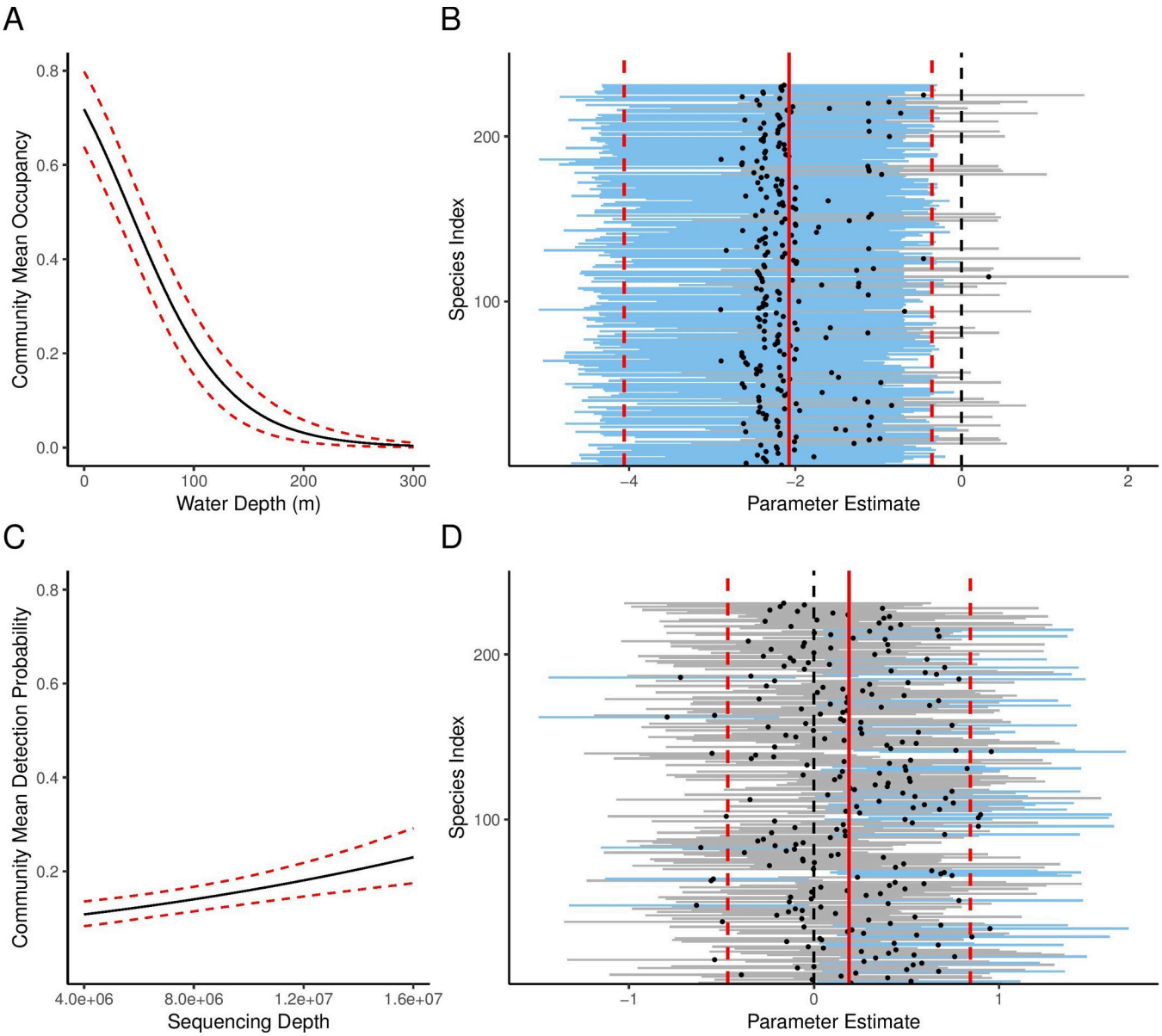

Fig 2. (A) Community mean occupancy by water depth (m) and (C) community mean probability of detection by sequencing depth predicted using a multi-species, multi-scale community occupancy model. The dashed lines represent the 95% confidence interval. Parameter estimate for each species for (B) the effect of water depth on occupancy and (D) the effect of sequencing depth on detection in a multi-species, multi-scale community occupancy model. Solid red line indicates the community mean and dashed red lines indicate the upper and lower limits of the 95% confidence intervals of the community mean parameter estimate. Blue lines indicate 95% confidence intervals of individual species parameter estimates that do not overlap with 0. Grey lines indicate 95% confidence intervals of individual species parameter estimates that do overlap with 0.

https://doi.org/10.1371/journal.pone.0224119.g002

occupancy models that include sequencing depth as a covariate will allow the variation in sequencing effort, which cannot always be controlled, to be accounted for when making ecological conclusions about biodiversity and occupancy.

The mean probability of capture estimate of 0.98 suggests a high probability of collecting a species' DNA in a given sample. However, the mean detection probability was relatively low at 0.15, likely because many species were not detected consistently by multiple amplicons, and a

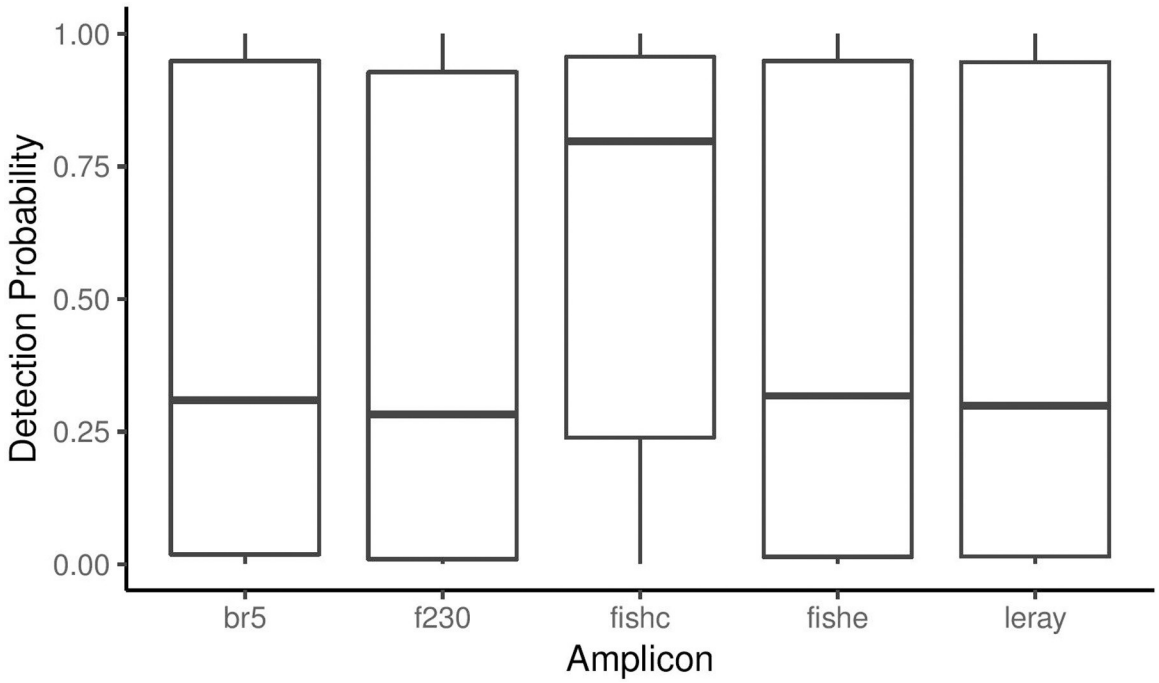

Fig 3. Mean detection probability estimated from occupancy model 9 (*ψ(.) θ(.) p(amplicon)*) for each species plotted by amplicon. The band in the middle of the box represents the median and the upper and lower edges of the box represent the upper and lower quartiles. The whiskers represent 1.5 times the inter-quartile range.

https://doi.org/10.1371/journal.pone.0224119.g003

low probability of detection can lead to overestimates for higher level parameters, including probability of capture [47]. Additionally, many species were only detected in the shallow water sites and were detected consistently across biological replicates at these sites, further contributing to a high estimate for the community mean probability of capture.

Species-specific probabilities of detection varied by amplicon and by phylum. Since we did not have replication within each amplicon for each sample, the effects of amplicon are confounded with the effects of occasion (i.e. PCR stochasticity). However, the effects of amplicon on species detectability were consistent across samples and it is very unlikely that this pattern would be observed due to PCR stochasticity alone, which is a random process. Therefore, we assume the effects we observed at the level of detectability were driven by the different amplicons used. Since the performance of each amplicon varies by taxonomic group (this study; [13]), including a variety of amplicons is important to detect species across the tree of life, and increasing the number of technical replicates using a single amplicon will not necessarily improve the community-wide probability of detection. Information such as this can be used to guide primer selection for future metabarcoding studies to maximize the probability of detection for target taxa or to ensure broad taxonomic coverage for holistic biodiversity surveys.

We used the occupancy modeling framework to estimate the species richness for the survey area and determined that 53 species or approximately 19% of the estimated number of species present were undetected during our surveys. Similar to many ecological studies, the case study presented here included a relatively low spatial coverage (n = 8 sites), but our occupancy modelling approach allowed us to assess false absences at two different sampling levels in our study and thereby understand what portion of biodiversity was missed, which is a significant improvement from most metabarcoding surveys [11]. Understanding how biological replication and technical replication affect biodiversity estimates can inform future sampling designs

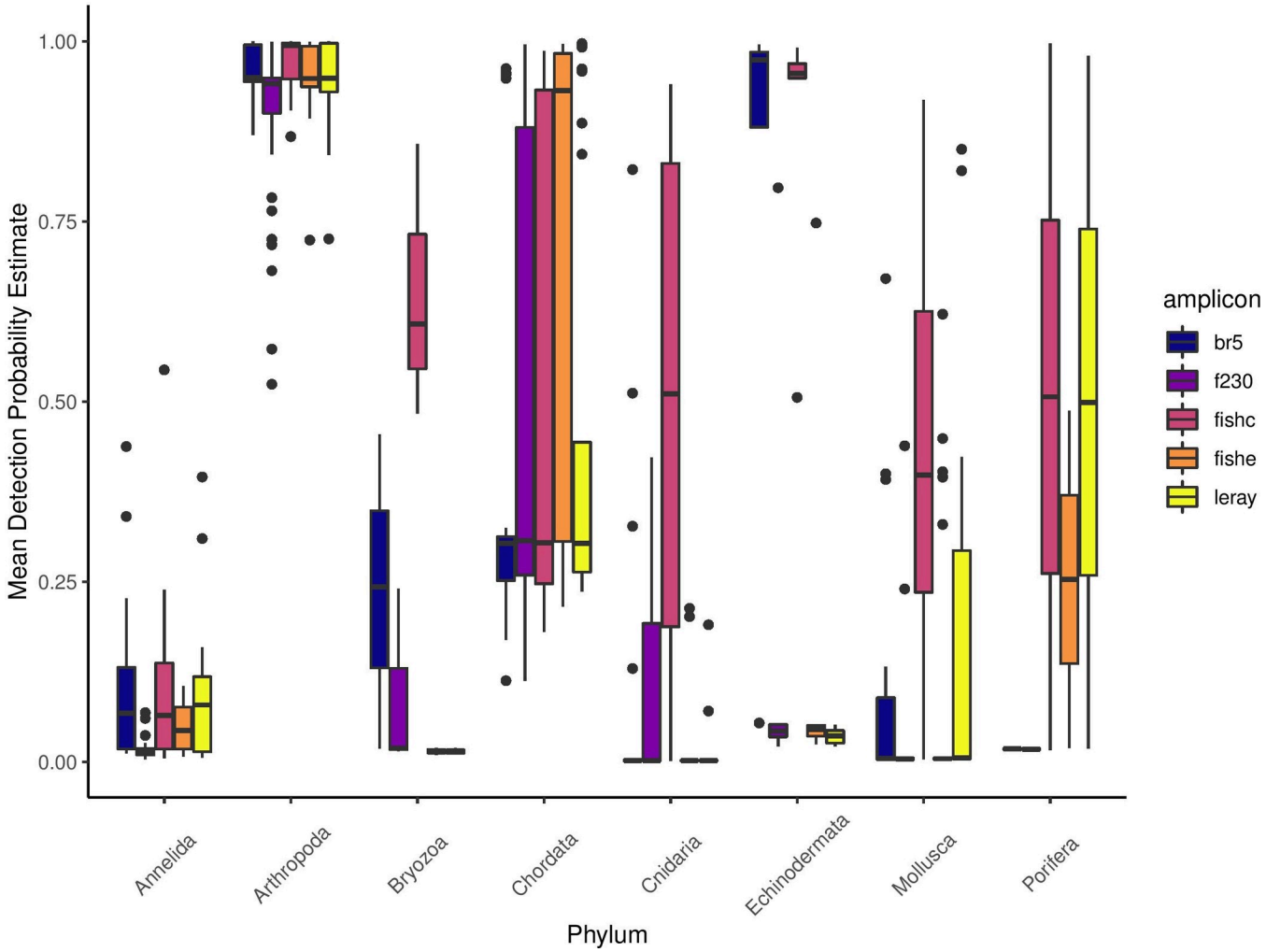

Fig 4. Mean detection probability estimated from occupancy model 9 ($\psi(.)\ \theta(.)\ p(amplicon)$) for each species plotted by amplicon and phylum for metazoan phyla only. The band in the middle of the box represents the median and the upper and lower edges of the box represent the upper and lower quartiles. The whiskers represent 1.5 times the inter-quartile range.

https://doi.org/10.1371/journal.pone.0224119.g004

to maximize biodiversity detection while minimizing cost and effort. In our case study, the sampling effort was limited, and thus there are several ways the proportion of species detected could be improved: (1) increasing sampling effort in the field by sampling more sites, (2) collecting more replicate biological samples at each site, and (3) including additional amplicons during laboratory processing. Given the limited extent and breadth of our sampling effort, the conclusions regarding the effect of covariates and the estimates of occupancy, capture, and detection probabilities for individual species should not be extrapolated to other systems. Further research should investigate the impacts of variation in sequencing depth and amplicons targeted on detection probability in metabarcoding studies, particularly in other ecosystems and across greater spatial scales.

Through the inclusion of environmental and experimental covariates, the multi-species occupancy framework can be applied for direct ecological assessment and to improve the methodology for next-generation biodiversity assessment. Metabarcoding results are often presented as taxonomic lists of species presence with alpha and beta-diversity estimates, and

sampling effort is often assessed using accumulation curves (e.g. [48]). However, these methods do not account for imperfect detection and cannot accommodate the many variables in the field and the lab that can impact these results. From an ecological perspective, environmental variables (e.g. temperature, salinity, turbidity) can be included at the level of occupancy to determine their effects on community diversity and the presence of individual species. From a methodological perspective, environmental and experimental variables (e.g. sample volume, sequencing depth) can be included at the level of field sampling and technical replication to understand how these factors affect metabarcoding results. Understanding the effects of these covariates facilitates the development of more robust experimental and survey designs. Furthermore, simulations using occupancy models can be used to optimize sampling effort, enabling practitioners to fine-tune the trade-off between field sampling and lab work [21]. The number of sites, biological samples, and technical replicates can all be optimized to maximize the species richness recovered from eDNA samples while minimizing effort. PCR level stochasticity, which is known to affect sequencing results [46,49], was not considered in our case study (i.e., PCR replicates were pooled before sequencing) but PCR replicates can easily be included as technical replicates in the model described here. PCR replicates are commonly included separately in single-species occupancy models for eDNA data [19,20,27]. By including PCR replicates as technical replicates, additional stochasticity in the sampling process can be accounted for, further improving inferences.

A key advantage of the occupancy modeling framework demonstrated here is its flexibility. Modifications to the model can allow several additional factors to be included, and a priori information can be used to guide model development. For example, multiple sampling periods have been included in dynamic, multi-season occupancy models to quantify temporal changes in community structure (e.g. [22]). Repeated eDNA sampling for metabarcoding could be modelled similarly to account for local extinction and colonization events between sampling periods. In addition to accounting for false negatives, several studies have developed methods for including false positives in occupancy models [50–52]. False positives may potentially arise from metabarcoding data through sequencing errors, PCR errors, and poor reference database coverage or quality [15,53,54]. Strict bioinformatic filtering helps to minimize the inclusion of these errors in resulting data sets; however, the possibility of false positives cannot be eliminated. Our model did not consider false positives, and, to our knowledge, these have yet to be incorporated into multi-species occupancy models. Following current protocols, abundance estimates from metabarcoding data are not reliable [55,56], but occupancy models can provide a means to estimate trends in abundance from the presence-absence data generated by metabarcoding based on documented relationships between occupancy and abundance [57–59]. The hierarchical modeling framework used in occupancy modeling can also be adapted to include or estimate taxa abundances [31]. As more research is done to understand the relationship between sequence read counts and species biomass, read counts could potentially be used as taxon abundances in a hierarchical modelling framework to estimate species biomass.

We demonstrate for the first time how a multi-scale, multi-species occupancy modelling framework can be used in a natural system to account for imperfect detection and allow for critical assessment of experimental and environmental factors influencing biodiversity data from eDNA metabarcoding. Despite the utility of these models for improving detection and targeting areas of variation in the pipeline from sample collection to sample processing, this approach has been underutilized in DNA metabarcoding studies (S1 Table; but see [25]). This multi-species occupancy modelling framework will be particularly useful for bioassessment studies using DNA metabarcoding because it will improve estimates of occupancy and species richness, aid in optimizing sampling efforts in the field and lab, and, using the model assessment methods described here, identify ecological and environmental factors affecting

occupancy, capture, and detection probabilities. Given the high stakes for documenting and understanding biodiversity that is under increasing anthropogenic threat [60] and declining [61] globally, new tools are imperative for rapid bioassessment [7,62,63]; yet, like any emergent technology, there is the potential to misuse these tools [64], which can have unforeseen consequences (e.g. [65]). In the case of DNA metabarcoding, neglecting to assess imperfect detection at key points along the sample collection and processing pipeline could lead to failure to detect species of interest, biased estimates of species richness, and miscalculations of species distributions, all of which have consequences for conservation and management [24,66,67]. We recommend incorporating multi-scale, multi-species occupancy modeling into the design and analysis of future metabarcoding studies.

Supporting information

S1 Table. Literature review of occupancy modeling for metabarcoding data. List of studies (n = 5) that have examined occupancy models in the context of DNA metabarcoding. To obtain this list of publications, we performed the following systematic, Boolean search using the Web of Science: "*DNA" AND "metabarcoding" AND "occupancy model*".
(DOCX)

S2 Table. Species-specific parameter estimates from multi-species, multi-scale occupancy model. Species-specific estimates of occupancy, capture and detection using mean covariates values (water depth (m) and sequencing depth) from a multi-scale, multi-species occupancy model for eDNA metabarcoding data from Conception Bay, Newfoundland.
(DOCX)

S1 File. Model assessment calculations for multi-species, multi-scale occupancy models. Calculations presented for likelihood, Bayesian p-value, WAIC, CPO, Brier score and log score for model assessment and comparison.
(DOCX)

S2 File. Model structure in BUGS language. BUGS model structure presented for the most complex model ψ(water depth) θ(.) p(sequencing depth, amplicon), the best model ψ(water depth) θ(.) p(sequencing depth) with data augmentation and the ψ(.) θ(.) p(amplicon) model with a group-level effect (phylum).
(DOCX)

S3 File. Diagnostic plots to asses fit of ψ(water depth) θ(.) p(sequencing depth) model. Plots of residual deviance for each species, site and covariate from a multi-species, multi-scale occupancy model based on eDNA metabarcoding data collected from Conception Bay, Newfoundland.
(DOCX)

Acknowledgments

We would like to thank Nicole Fahner, Joshua Barnes, Avery McCarthy, Greg Singer, Hoda Rajabi, and Emily Porter for their assistance in sample collection, processing and bioinformatics.

Author Contributions

Conceptualization: Beverly McClenaghan, Mehrdad Hajibabaei.

Formal analysis: Beverly McClenaghan.

Funding acquisition: Mehrdad Hajibabaei.

Methodology: Beverly McClenaghan.

Supervision: Mehrdad Hajibabaei.

Validation: Zacchaeus G. Compson.

Writing – original draft: Beverly McClenaghan.

Writing – review & editing: Zacchaeus G. Compson, Mehrdad Hajibabaei.

References

1. 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. https://doi.org/10.1111/mec.14350 PMID: 28921802
2. Thomsen PF, Willerslev E. Environmental DNA–An emerging tool in conservation for monitoring past and present biodiversity. Biol Conserv. 2015; 183: 4–18. https://doi.org/10.1016/j.biocon.2014.11.019
3. 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. https://doi.org/10.1016/j.tree.2014.04.003 PMID: 24821515
4. Jeunen G, Knapp M, Spencer HG, Lamare MD, Taylor HR, Stat M, et al. Environmental DNA (eDNA) metabarcoding reveals strong discrimination among diverse marine habitats connected by water movement. Mol Ecol Resour. 2019; 19: 426–438. https://doi.org/10.1111/1755-0998.12982 PMID: 30576077
5. Singer GAC, Fahner NA, Barnes JG, McCarthy A, Hajibabaei M. Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: a case study of eDNA metabarcoding seawater. Sci Rep. 2019; 9: 5991. https://doi.org/10.1038/s41598-019-42455-9 PMID: 30979963
6. Stat M, Huggett MJ, Bernasconi R, DiBattista JD, Berry TE, Newman SJ, et al. Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment. Sci Rep. 2017; 7: 12240. https://doi.org/10.1038/s41598-017-12501-5 PMID: 28947818
7. Baird DJ, Hajibabaei M. Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. Mol Ecol. 2012; 21: 2039–2044. https://doi.org/10.1111/j.1365-294x.2012.05519.x PMID: 22590728
8. 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. https://doi.org/10.1111/mec.13428 PMID: 26479867
9. Kéry M, Schmidt B. Imperfect detection and its consequences for monitoring for conservation. Community Ecol. 2008; 9: 207–216. https://doi.org/10.1556/ComEc.9.2008.2.10
10. Dorazio RM, Royle JA, Söderström B, Glimskär A. Estimating species richness and accumulation by modeling species occurence and detectability. Ecology. 2006; 87: 842–854. https://doi.org/10.1890/0012-9658(2006)87[842:esraab]2.0.co;2 PMID: 16676528
11. Guillera-Arroita G, Lahoz-Monfort JJ, MacKenzie DI, Wintle BA, McCarthy MA. Ignoring imperfect detection in biological surveys is dangerous: A response to 'Fitting and interpreting occupancy models".' White EP, editor. PLoS ONE. 2014; 9: e99571. https://doi.org/10.1371/journal.pone.0099571 PMID: 25075615
12. Evans NT, Shirey PD, Wieringa JG, Mahon AR, Lamberti GA. Comparative cost and effort of fish distribution detection via environmental DNA analysis and electrofishing. Fisheries. 2017; 42: 90–99. https://doi.org/10.1080/03632415.2017.1276329
13. Freeland J. The importance of molecular markers and primer design when characterizing biodiversity from environmental DNA (eDNA). Genome. 2017; 60: 358–374. https://doi.org/10.1139/gen-2016-0100 PMID: 28177833
14. Smith DP, Peay KG. Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing. Kellogg CA, editor. PLoS ONE. 2014; 9: e90234. https://doi.org/10.1371/journal.pone.0090234 PMID: 24587293
15. 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. https://doi.org/10.1111/1755-0998.12338 PMID: 25327646
16. Campos-Cerqueira M, Aide TM. Improving distribution data of threatened species by combining acoustic monitoring and occupancy modelling. Jones K, editor. Methods Ecol Evol. 2016; 7: 1340–1348. https://doi.org/10.1111/2041-210X.12599

17. Ramesh T, Downs CT. Impact of land use on occupancy and abundance of terrestrial mammals in the Drakensberg Midlands, South Africa. J Nat Conserv. 2015; 23: 9–18. https://doi.org/10.1016/j.jnc.2014.12.001

18. Steenweg R, Whittington J, Hebblewhite M, Forshner A, Johnston B, Petersen D, et al. Camera-based occupancy monitoring at large scales: Power to detect trends in grizzly bears across the Canadian Rockies. Biol Conserv. 2016; 201: 192–200. https://doi.org/10.1016/j.biocon.2016.06.020

19. Hunter ME, Oyler-McCance SJ, Dorazio RM, Fike JA, Smith BJ, Hunter CT, et al. Environmental DNA (eDNA) sampling improves occurrence and detection estimates of invasive Burmese pythons. Mahon AR, editor. PLoS ONE. 2015; 10: e0121655. https://doi.org/10.1371/journal.pone.0121655 PMID: 25874630

20. Schmidt BR, Kéry M, Ursenbacher S, Hyman OJ, Collins JP. Site occupancy models in the analysis of environmental DNA presence/absence surveys: a case study of an emerging amphibian pathogen. Yoccoz N, editor. Methods Ecol Evol. 2013; 4: 646–653. https://doi.org/10.1111/2041-210X.12052

21. Erickson RA, Merkes CM, Mize EL. Sampling designs for landscape-level eDNA monitoring programs. Integr Environ Assess Manag. 2019 [cited 15 May 2019]. https://doi.org/10.1002/ieam.4155 PMID: 30963692

22. Goijman AP, Conroy Michael J, Bernardos JN, Zaccagnini ME. Multi-season regional analysis of multi-species occupancy: Implications for bird conservation in agricultural lands in East-Central Argentina. Arlettaz R, editor. PLoS ONE. 2015; 10: e0130874. https://doi.org/10.1371/journal.pone.0130874 PMID: 26086250

23. Van der Weyde LK, Mbisana C, Klein R. Multi-species occupancy modelling of a carnivore guild in wildlife management areas in the Kalahari. Biol Conserv. 2018; 220: 21–28. https://doi.org/10.1016/j.biocon.2018.01.033

24. Guillera-Arroita G. Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. Ecography. 2017; 40: 281–295. https://doi.org/10.1111/ecog.02445

25. Doi H, Fukaya K, Oka S, 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: 1–8. https://doi.org/10.1038/s41598-018-37186-2

26. Lugg WH, Griffiths J, van Rooyen AR, Weeks AR, Tingley R. Optimal survey designs for environmental DNA sampling. Jarman S, editor. Methods Ecol Evol. 2018; 9: 1049–1059. https://doi.org/10.1111/2041-210X.12951

27. Strickland GJ, Roberts JH. Utility of eDNA and occupancy models for monitoring an endangered fish across diverse riverine habitats. Hydrobiologia. 2019; 826: 129–144. https://doi.org/10.1007/s10750-018-3723-8

28. Broms KM, Hooten MB, Fitzpatrick RM. Model selection and assessment for multi-species occupancy models. Ecology. 2016; 97: 1759–1770. https://doi.org/10.1890/15-1471.1 PMID: 27859174

29. Watanabe S. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. J Mach Learn Res. 2010; 11: 3571–3594.

30. Pettit LI. The conditional predictive ordinate for the normal distribution. J R Stat Soc Ser B Stat Methodol. 1990; 52: 175–184. https://doi.org/10.1111/j.2517-6161.1990.tb01780.x

31. Kéry M, Royle JA. Applied Hierarchical Modeling in Ecology. London: Academic press; 2016.

32. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016; 13: 581–583. https://doi.org/10.1038/nmeth.3869 PMID: 27214047

33. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215: 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712

34. Shokralla S, Hellberg RS, Handy SM, King I, Hajibabaei M. A DNA mini-barcoding system for authentication of processed fish products. Sci Rep. 2015;5. https://doi.org/10.1038/srep15894 PMID: 26516098

35. Gibson JF, Shokralla S, Curry C, Baird DJ, Monk WA, King I, et al. Large-scale biomonitoring of remote and threatened ecosystems via high-throughput sequencing. Fontaneto D, editor. PLoS ONE. 2015; 10: e0138432. https://doi.org/10.1371/journal.pone.0138432 PMID: 26488407

36. 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. https://doi.org/10.1186/1742-9994-10-34 PMID: 23767809

37. Shokralla S, Porter TM, Gibson JF, Dobosz R, Janzen DH, Hallwachs W, et al. Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. Sci Rep. 2015;5. https://doi.org/10.1038/srep09687 PMID: 25884109

38. R Core Team. R: A language and environment for statistical computing. R Found Stat Comput Vienna Austria. 2018. Available: https://www.R-project.org/

39. Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Proc 3rd Int Workshop Dsitributed Stat Comput. 2003. Available: https://www.ci.tuwien.ac.at/Conferences/DSC-2003/

40. Kellner K. jagUI: a wrapper around "rjags" to streamline "JAGS" analyses. 2018;R package version 1.5.0. Available: https://CRAN.R-project.org/package=jagsUI

41. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. J Comput Graph Stat. 1998; 7: 434–455. https://doi.org/10.2307/1390675

42. Kruschke JK. Doing Bayesian Data Analysis. 2nd ed. London: Academic Press; 2015.

43. Meredith M, Kruschke J. HDInterval: Highest (Posterior) Density Intervals. R Package Version 020. 2018. Available: https://CRAN.R-project.org/package=HDInterval

44. Leray M, Knowlton N. Censusing marine eukaryotic diversity in the twenty-first century. Philos Trans R Soc B Biol Sci. 2016; 371: 20150331. https://doi.org/10.1098/rstb.2015.0331 PMID: 27481783

45. Sigsgaard EE, Nielsen IB, Carl H, Krag MA, Knudsen SW, Xing Y, et al. Seawater environmental DNA reflects seasonality of a coastal fish community. Mar Biol. 2017;164. https://doi.org/10.1007/s00227-017-3147-4

46. Alberdi A, Aizpurua O, Gilbert MTP, Bohmann K. Scrutinizing key steps for reliable metabarcoding of environmental samples. Mahon A, editor. Methods Ecol Evol. 2018; 9: 134–147. https://doi.org/10.1111/2041-210X.12849

47. MacKenzie DI, Nichols JD, Lachman GB, Droege S, Royle JA, Langtimm CA. Estimating site occupancy rates when detection probabilities are less than one. Ecology. 2002; 83: 2248–2255. https://doi.org/10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2

48. Evans NT, Li Y, Renshaw MA, Olds BP, Deiner K, Turner CR, et al. Fish community assessment with eDNA metabarcoding: effects of sampling design and bioinformatic filtering. Can J Fish Aquat Sci. 2017; 74: 1362–1374. https://doi.org/10.1139/cjfas-2016-0306

49. Kebschull JM, Zador AM. Sources of PCR-induced distortions in high-throughput sequencing data sets. Nucleic Acids Res. 2015; 43: e143. https://doi.org/10.1093/nar/gkv717 PMID: 26187991

50. Royle JA, Link WA. Generalized site occupancy model allowing for false positive and false negative errors. Ecology. 2006; 87: 835–841. https://doi.org/10.1890/0012-9658(2006)87[835:gsomaf]2.0.co;2 PMID: 16676527

51. Lahoz-Monfort JJ, Guillera-Arroita G, Tingley R. Statistical approaches to account for false-positive errors in environmental DNA samples. Mol Ecol Resour. 2016; 16: 673–685. https://doi.org/10.1111/1755-0998.12486 PMID: 26558345

52. Guillera-Arroita G, Lahoz-Monfort JJ, van Rooyen AR, Weeks AR, Tingley R. Dealing with false-positive and false-negative errors about species occurrence at multiple levels. McCrea R, editor. Methods Ecol Evol. 2017; 8: 1081–1091. https://doi.org/10.1111/2041-210X.12743

53. Ficetola GF, Taberlet P, Coissac E. How to limit false positives in environmental DNA and metabarcoding? Mol Ecol Resour. 2016; 16: 604–607. https://doi.org/10.1111/1755-0998.12508 PMID: 27062589

54. Porter TM, Hajibabaei M. Automated high throughput animal CO1 metabarcode classification. Sci Rep. 2018; 8: 4226. https://doi.org/10.1038/s41598-018-22505-4 PMID: 29523803

55. Fonseca VG. Pitfalls in relative abundance estimation using eDNA metabarcoding. Mol Ecol Resour. 2018; 18: 923–926. https://doi.org/10.1111/1755-0998.12902

56. Lamb PD, Hunter E, Pinnegar JK, Creer S, Davies RG, Taylor MI. How quantitative is metabarcoding: A meta-analytical approach. Mol Ecol. 2019; 28: 420–430. https://doi.org/10.1111/mec.14920 PMID: 30408260

57. Hall K, MacLeod CD, Mandleberg L, Schweder-Goad CM, Bannon SM, Pierce GJ. Do abundance–occupancy relationships exist in cetaceans? J Mar Biol Assoc U K. 2010; 90: 1571–1581. https://doi.org/10.1017/S0025315410000263

58. Miranda LE, Killgore KJ. Abundance–occupancy patterns in a riverine fish assemblage. Freshw Biol. 2019; 64: 2221–2233. https://doi.org/10.1111/fwb.13408

59. Habel JC, Trusch R, Schmitt T, Ochse M, Ulrich W. Long-term large-scale decline in relative abundances of butterfly and burnet moth species across south-western Germany. Sci Rep. 2019;9. https://doi.org/10.1038/s41598-018-36956-2

60. Steffen W, Broadgate W, Deutsch L, Gaffney O, Ludwig C. The trajectory of the Anthropocene: The great acceleration. Anthr Rev. 2015; 2: 81–98. https://doi.org/10.1177/2053019614564785

61. IPBES. Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. Bonn, Germany: IPBES Secretariat; 2019.

62. Ji Y, Ashton L, Pedley SM, Edwards DP, Tang Y, Nakamura A, et al. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Holyoak M, editor. Ecol Lett. 2013; 16: 1245–1257. https://doi.org/10.1111/ele.12162 PMID: 23910579

63. Lacoursière-Roussel A, Howland K, Normandeau E, Grey EK, Archambault P, Deiner K, et al. eDNA metabarcoding as a new surveillance approach for coastal Arctic biodiversity. Ecol Evol. 2018; 8: 7763–7777. https://doi.org/10.1002/ece3.4213 PMID: 30250661

64. Cristescu ME, Hebert PDN. Uses and misuses of environmental DNA in biodiversity science and conservation. Annu Rev Ecol Evol Syst. 2018; 49: 209–230. https://doi.org/10.1146/annurev-ecolsys-110617-062306

65. Garcia M. Racist in the machine: The disturbing implications of algorithmic bias. World Policy J. 2016; 33: 111–117. https://doi.org/10.1215/07402775-3813015

66. Comte L, Grenouillet G. Species distribution modelling and imperfect detection: comparing occupancy versus consensus methods. Robertson M, editor. Divers Distrib. 2013; 19: 996–1007. https://doi.org/10.1111/ddi.12078

67. DeWan AA, Zipkin EF. An integrated sampling and analysis approach for improved biodiversity monitoring. Environ Manage. 2010; 45: 1223–1230. https://doi.org/10.1007/s00267-010-9457-7 PMID: 20237922