Multispecies site occupancy modeling and study design for spatially replicated environmental DNA metabarcoding

Keiichi Fukaya,* Natsuko Ito Kondo, Shin-Ichiro S. Matsuzaki, and Taku Kadoya

National Institute for Environmental Studies, Tsukuba, Ibaraki 305-8506, Japan

*fukaya.keiichi@nies.go.jp

February 15, 2021

Abstract

1. Environmental DNA (eDNA) metabarcoding has become widely applied to gauge biodiversity in a noninvasive and cost-efficient manner. The detection of species using eDNA metabarcoding is, however, imperfect owing to various factors that can cause false negatives in the inherent multi-stage workflow.

2. Imperfect detection in the multi-stage workflow of eDNA metabarcoding also raises an issue of study design: namely, how available resources should be allocated among the different stages to optimize survey efficiency.

3. Here, we propose a variant of the multispecies site occupancy model for eDNA metabarcoding studies where samples are collected at multiple sites within a region of interest. This model describes the variation in sequence reads, the unique output of the high-throughput sequencers, in terms of the hierarchical workflow of eDNA metabarcoding and interspecific heterogeneity, allowing the decomposition of the sources of variation in the detectability of species throughout the different stages of the workflow. We also introduced a Bayesian decision analysis framework to identify the study design that optimizes the efficiency of species detection with a limited budget.

4. An application of the model to freshwater fish communities in the Lake Kasumigaura watershed, in Japan, highlighted a remarkable inhomogeneity in the detectability of species, indicating a potential risk of the biased detection of specific species. Species with lower site occupancy probabilities tended to be difficult to detect as they had lower capture probabilities and lower dominance of the sequences. The expected abundance of sequence reads was predicted to vary by up to 23.5 times between species.

5. An analysis of the study design suggested that ensuring multiple within-site replications of the environmental samples is preferred in order to achieve better species detection efficiency, provided that a throughput of tens of thousands of sequence reads was secured.

6. The proposed framework makes the application of eDNA metabarcoding more error-tolerant, allowing ecologists to monitor ecological communities more efficiently.

Key words: Bayesian decision analysis, environmental DNA (eDNA) metabarcoding, false negatives, freshwater fish communities, hierarchical modeling, imperfect detection, multi-species site occupancy models, study design

1 Introduction

High-throughput sequencing (HTS) technology has achieved a breakthrough in biology, particularly in the fields of ecology and evolution (Shokralla *et al.* 2012). In ecology, HTS can be applied alongside targeted PCR amplification to obtain the sequence data of target genetic markers (such as COI and rRNA gene regions that act as “DNA barcodes”) directly and effectively from environmental samples. This approach, called environmental DNA (eDNA) metabarcoding (Taberlet *et al.* 2012), enables the simultaneous detection of a number of species from environmental samples. Originally applied to profile microbial communities, the use of eDNA metabarcoding for macro-organisms is increasing and becoming a useful tool in ecology and conservation biology to gauge biodiversity in a noninvasive and cost-efficient manner (Taberlet *et al.* 2012, Deiner *et al.* 2017).

The workflow of eDNA metabarcoding is multi-staged, requiring different procedures in the field and laboratory and on the computer (Deiner *et al.* 2017). First, in the field, researchers collect environmental samples from target sites to assess ecological communities. Then, the collected environmental samples are transferred to the laboratory where DNA molecules are extracted and amplified for the subsequent HTS of the target marker gene. Next, HTS yields a high number of raw sequence reads that are processed in a computer using bioinformatic tools, including quality filtering, read clustering, and assigning sequences to species (or, more generally, taxa) using a reference database. These eDNA metabarcoding procedures will eventually result in a species-by-sample table of sequence read counts, which can then be used to characterize communities in terms of the presence-absence (or even “relative abundance”) of species and measures of α - and β -diversity.

The detection of species using eDNA may be more sensitive than that of traditional survey methods (Smart *et al.* 2015, Yamamoto *et al.* 2017, McColl-Gausden *et al.* 2020). Nevertheless, varying detection probability can still lead to the imperfect detection of species, especially in community-wide eDNA metabarcoding, compared to that of targeted species-specific approaches such as quantitative PCR (Harper *et al.* 2018, Bylemans *et al.* 2019, Doi *et al.* 2019). False-negative detection errors are caused by a number of factors within the multi-stage eDNA metabarcoding procedure. For example, sampling processes, including both sample collection in the field and repeated subsampling for library preparation in the laboratory, can result in random losses of rare sequences (Zhan *et al.* 2014a, Shelton *et al.* 2016). Primer-template mismatches can have a huge impact on the efficiency of PCR amplification and reduce the detection probability of species with low primer specificity (Piñol *et al.* 2015, Shelton *et al.* 2016). A smaller throughput (i.e., the total number of sequence reads) can lead to failures in the detection of rare sequences in the sample (Shelton *et al.* 2016). Sequence filtering procedures, used to remove artifactual sequences, may erroneously filter out actual target sequences (Zhan *et al.* 2014b). Some species can never be detected when the reference database is incomplete. Hence, for the reliable assessment of ecological communities using eDNA metabarcoding, it is crucial to adequately account for such detection errors that occur throughout the multi-stage procedure.

Quantitative knowledge of imperfect detection in eDNA metabarcoding can have important implications from the perspective of study design. To ensure the accurate assessment of species diversity using eDNA metabarcoding, researchers may choose more general sampling strategies, including increasing the number of sampling sites, increasing the number of replicates per site, or increasing the throughput. Given that budget is often limited, decisions must be made to allocate the available resources among these options. A key consideration when identifying optimal resource allocation is to evaluate the detectability of species at different stages of the eDNA metabarcoding workflow. This approach allows the prediction of the efficiency of species detection among possible resource allocations under a given budget, which can then be used to determine the preferred study design.

To address these issues, we developed a novel extension of the multispecies site occupancy modeling framework (Dorazio & Royle 2005, Dorazio *et al.* 2006). Multispecies site occupancy models account for the imperfect detection of species in a metacommunity and have recently been applied in eDNA metabarcoding (Doi *et al.* 2019, Bush *et al.* 2020, McClenaghan *et al.* 2020, McColl-Gausden *et al.* 2020). However, to apply these models to eDNA metabarcoding data, the original sequence read counts must be aggregated into binary detection and non-detection indicators. We propose the adoption of an observation submodel based on a multinomial distribution, instead of the traditional Bernoulli observation submodel, to apply the multispecies site occupancy modeling framework directly to eDNA metabarcoding data. The objective of this approach is to allow the explicit decomposition of the sources of variation in the detectability of species in different stages of the process (Fig. 1). Furthermore, we introduce a Bayesian framework for optimizing resource allocation in a multi-stage workflow with a limited budget, which can be derived naturally based on the proposed model. We apply the proposed methods to an eDNA metabarcoding study of freshwater fish communities in the Lake Kasumigaura watershed, in Japan.

2 Hierarchical modeling for eDNA metabarcoding

2.1 Formulation

We considered studies of eDNA metabarcoding with spatially-replicated sampling designs (Fig. 1). Specifically, we assumed that the occurrence of I focal species was monitored at J sites sampled from an area of interest. We simply referred to taxonomic groups that emerged in an application of eDNA metabarcoding as “species”. At site j , K_j replicates of environmental samples were collected. For each replicate, a unique sequence library was prepared to obtain separate sequence reads for each replicate. We denoted the resulting sequence read counts of species for replicate k ($= 1, \dots, K_j$) at site j ($= 1, \dots, J$), obtained using HTS and subsequent bioinformatic processing, as $\mathbf{y}_{jk} = (y_{1jk}, \dots, y_{Ijk})$. We denoted the per-sample total sequence reads, which we called “throughput”, as $N_{jk} = \sum_i y_{ijk}$. Furthermore, we described a basic model structure with the least number of parameters; this model can be readily extended to account for additional heterogeneity, as described in Section 2.3.

We assumed that, conditional on the throughput N_{jk} , sequence reads $\mathbf{y}_{jk}$ followed a

multinomial distribution (Shelton *et al.* 2016, Harrison *et al.* 2020). For j and k satisfying $N_{jk} > 0$, the observation model for $\mathbf{y}_{jk}$ is expressed as

$$\mathbf{y}_{jk} \sim \text{Multinomial}(N_{jk}, \boldsymbol{\pi}_{jk}), \quad (1)$$

where $\boldsymbol{\pi}_{jk} = (\pi_{1jk}, \dots, \pi_{Ijk})$ are the multinomial cell probabilities expressing the relative frequency of the sequence of each species in the library of replicate k at site j . This observation model defines the per-sample marginal probability of species detection as $1 - (1 - \pi_{ijk})^{N_{jk}}$, implying that species detection through HTS depends on both the relative frequency of the species sequence and the throughput. We model π_{ijk} as follows:

$$\pi_{ijk} = \frac{u_{ijk} r_{ijk}}{\sum_{m=1}^I u_{mjk} r_{mjk}}, \quad (2)$$

where u_{ijk} is a latent indicator variable representing the inclusion of the sequence of species i in the library of replicate k at site j and r_{ijk} is a latent variable that is proportional to the relative frequency of the sequence of species i , conditional on its presence in the library of replicate k at site j . Therefore, we decomposed the relative frequency of the sequence of species into binary components defining the presence or absence of the sequence in the library (u) and quantitative components defining the degree of dominance of the sequence (r). The expected frequency of species i in the sequence reads of replicate k at site j was assumed to be proportional to r_{ijk} when the library included the sequence of species ($u_{ijk} = 1$), whereas it was assumed to be zero when the library did not contain the sequence of species ($u_{ijk} = 0$). We noted that Equation 2 requires $\sum_{m=1}^I u_{mjk} > 0$; this is assured by defining the observation model (Equation 1) as conditional on a positive throughput ($N_{jk} > 0$), which implies that u_{ijk} is positive for one or more species.

The inclusion of the sequence of species can be modeled with a Bernoulli distribution as follows:

$$u_{ijk} \sim \text{Bernoulli}(z_{ij}\theta_i), \quad (3)$$

where z_{ij} is a latent indicator variable representing the occupancy of species i at site j and θ_i is the probability of per-replicate sequence capture that is conditional on the site occupancy of species i . Equation 3 implies that a replicate library includes the sequence of species i with probability θ_i when the species occupies site j ($z_{ij} = 1$), whereas replicate libraries of unoccupied sites ($z_{ij} = 0$) never contain the sequence of the species. In practice, θ_i could depend on the concentration of sequence copies in the field and the volume of collected medium.

The latent variable r_{ijk} is assumed to follow a gamma distribution as follows:

$$r_{ijk} \sim \text{Gamma}(\phi_i, 1), \quad (4)$$

where we denote a gamma distribution with shape a and rate b by $\text{Gamma}(a, b)$. Then, $\boldsymbol{\pi}_{jk}$ follows a zero-inflated version of the Dirichlet distribution (Tang & Chen 2019), with a vector

of expected values conditional on $(u_{1jk}, \dots, u_{Ijk})$ of $(u_{1jk}\phi_1/A_{jk}, \dots, u_{Ijk}\phi_I/A_{jk})$, where $A_{jk} = \sum_m u_{mjk}\phi_m$. Therefore, ϕ_i is a species-specific parameter that governs the relative dominance of the sequence. Species that tend to yield more sequence reads, when they are present, are characterized by larger values of ϕ_i . The relative dominance of sequences may vary depending on the relative abundance of species and the relative efficiency of PCR amplification.

We assume that the latent indicator of site occupancy z_{ij} follows a Bernoulli distribution as follows:

$$z_{ij} \sim \text{Bernoulli}(\psi_i), \quad (5)$$

where ψ_i represents the site occupancy probability of species i . We noted that, in the context of eDNA-based monitoring, the site occupancy of species should be defined by the presence of DNA sequences rather than the presence of individuals in the site, as the presence of species can only be confirmed by the detection of sequences. In a lotic environment, such as rivers, long-distance transport of eDNA may result in the presence of sequences at sites where no individuals exist in the vicinity (Nukazawa *et al.* 2018). Consequently, general care should be taken when interpreting site occupancy when it is inferred from eDNA.

The model is characterized by a set of species-level parameters $(\phi_i, \theta_i, \psi_i)$. We can regard these parameters as random effects and assume that they follow a prior community-level distribution as follows (Dorazio & Royle 2005, Dorazio *et al.* 2006):

$$(\log \phi_i, \text{logit } \theta_i, \text{logit } \psi_i) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (6)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the community mean vector and the covariance matrix of $(\phi_i, \theta_i, \psi_i)$ on the appropriate link scales. This formula allows information to be shared between species, thereby facilitating parameter inferences for the entire community, including for rare species (Iknayan *et al.* 2014).

A key assumption of the model is the absence of false positives; however, they can still occur in eDNA metabarcoding. We discuss a general consideration for the issue of false positives in Section 5.

2.2 Inference

With a fully Bayesian approach using Markov chain Monte Carlo (MCMC) algorithms, we can fit the model and conduct flexible posterior analyses. For this approach, we needed to specify prior distributions for the model parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in order to obtain samples from the posterior of all unknown quantities $(r, u, z, \phi, \theta, \psi, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Specifically, for the mean vector $\boldsymbol{\mu}$, we can specify a vague normal prior for each element. Furthermore, for the covariance matrix $\boldsymbol{\Sigma}$, we can specify a vague prior for each of the scale and correlation elements (Dorazio & Royle 2005, Dorazio *et al.* 2006). Posterior samples can be readily obtained using standard general-purpose MCMC software, for example JAGS (Plummer 2003; see the following application).

The process of species detection using eDNA metabarcoding can be characterized based

on the resulting posterior of parameters. Specifically, the characteristics of individual species detection at different stages (Fig. 1) are represented by the posterior distribution of the species-level parameters $p(\phi_i | \mathbf{y})$, $p(\theta_i | \mathbf{y})$, and $p(\psi_i | \mathbf{y})$. In contrast, community-level characteristics can be inferred based on the posterior of the parameters in the community-level priors $p(\boldsymbol{\mu} | \mathbf{y})$ and $p(\boldsymbol{\Sigma} | \mathbf{y})$, which represent the community mean and interspecific variation in the species-level parameters, respectively.

The occupancy of species at particular sites can be inferred based on the posterior probability $\Pr(z_{ij} = 1 | \mathbf{y})$. Under the assumption that there are no false positives, the obvious solution $\Pr(z_{ij} = 1 | \mathbf{y}) = 1$ is obtained in cases where the sequence of species i is detected at site j . In cases of non-detection, the posterior probability $\Pr(z_{ij} = 1 | \mathbf{y})$ is less than one and can be evaluated based on the posterior samples of z_{ij} . We can also infer the number of species that occur at a particular site, which can be expressed as $\sum_i z_{ij}$ (Dorazio & Royle 2005) as a derived quantity.

2.3 Covariates and random effects for species-level parameters

Covariates and random effects can be modeled as the linear predictors of species-level parameters on appropriate link scales. The addition of such effects allows us to account for the heterogeneities in species-level parameters in a flexible manner. Covariates and random effects can be species-specific, site-specific, or replicate-specific. Possible covariates include site-specific environmental variables for site occupancy probability ψ , replicate-specific volume of filtered water for sequence capture probability θ , and a species-specific degree of primer-template mismatches for the relative dominance of sequences ϕ ; see Section 4 for an illustration. Note that for models with such additional terms, the prior specification described in Section 2.2 will no longer apply. Depending on the parameters included, appropriate priors need to be specified.

3 Analysis for optimal study design

3.1 An overview

The problem of identifying the optimal study design for site occupancy studies was formulated by the criterion of maximizing the accuracy of the occupancy estimator under a given amount of survey budget or effort (or minimizing the amount of survey budget to achieve the desired estimator accuracy). This approach may be readily applicable to the simplest single-species site occupancy models with a large sample size, for which an analytical expression for the variance of the estimator is available (MacKenzie & Royle 2005, Guillera-Aroita *et al.* 2010, 2014, Lugg *et al.* 2018). A numerical simulation approach can be used for other classes of site occupancy models (Bailey *et al.* 2007); however, this may require extensive computation resources, especially for multispecies models (Sanderlin *et al.* 2014).

Here, as an alternative approach that is less computationally expensive, but still useful, we introduced a Bayesian decision analysis framework to determine the optimal study design

under uncertainty (Chaloner & Verdinelli 1995, Dorazio & Johnson 2003, Guillera-Arroita *et al.* 2014). In this framework, the optimal study design is identified based on a profile of values of some arbitrary utility function averaged over the posterior (or even prior) distribution of unknown quantities. As a utility function relevant to eDNA metabarcoding, we considered the efficiency of species detection which can be derived from the multispecies site occupancy model described in Section 2.1.

We considered two different situations for optimizing study design under a given budget: to maximize survey efficiency for the assessment of (1) local species diversity and (2) regional species diversity. In the former, we infer the situation of balancing the number of replicates and the amount of throughput for given sites, to maximize the opportunity of detecting the occupying species at each site. In the latter, we infer the situation of balancing the number of sites, number of replicates, and the amount of throughput to maximize the number of species in the region expected to be detected. In the following, we describe a method for the former situation in detail. The method can be extended to apply to the latter situation, which we describe in detail in Appendix S1.

3.2 Study design for local species diversity

To model local species diversity, we assume that the occurrence of focal species I on sites J is modeled as described in Section 2.1. We used the result of model fitting, namely the posterior distribution, to obtain a profile of the efficiency of species detection among different resource allocations between the number of replicates per site and per throughput. The posterior distribution can be replaced by a prior distribution that reflects some guesses about the value of unknown quantities when no data was available at the beginning of the study (see Section 5). We briefly denoted the posterior distribution as $p(\xi | \mathbf{y})$ by expressing all unknowns collectively as ξ .

First, a cost function is introduced. As the number of sites is fixed, we need to consider only the costs associated with the procedures required after the collection of the environmental samples at each site. Specifically, we assumed the costs of HTS and library preparation (including filtration, DNA extraction, and PCR). We denoted the cost per sequence read for HTS as λ_1 and the cost per replicate for library preparation as λ_2 . If we assume the same number of replicates per site and a constant throughput, the total cost for a study design with J sites, K replicates, and throughput N is expressed as $\lambda_1 JKN + \lambda_2 JK$. The total cost must be less than or equal to a given amount of budget B . The budget constraint can be expressed as:

$$\lambda_1 JKN + \lambda_2 JK \leq B. \quad (7)$$

We assumed that we can obtain the maximum number of sequence reads per replicate, $N = (B - \lambda_2 JK) / \lambda_1 JK$, for $K = 1, 2, \dots, \lfloor B / \lambda_2 J \rfloor$, where $\lfloor \cdot \rfloor$ represents the floor function.

Next, we defined utility as the expected number of detected species per site. This

criterion is legitimate as it can be used to maximize the opportunity to verify the site occupancy of species at the level of local communities. Under the model described in Section 2.1, the expected number of detected species per site conditional on $\mathbf{r} = \{r_{ijk}\}$ and $\mathbf{u} = \{u_{ijk}\}$, denoted as $U(K | \mathbf{r}, \mathbf{u})$, is expressed as:

$$U(K | \mathbf{r}, \mathbf{u}) = \frac{1}{J} \sum_{j=1}^J \sum_{i=1}^I \left\{ 1 - \prod_{k=1}^K \left(1 - \frac{u_{ijk} r_{ijk}}{\sum_{m=1}^I u_{mjk} r_{mjk}} \right)^{\frac{B - \lambda_2 J K}{\lambda_1 J K}} \right\}. \quad (8)$$

for $K = 1, 2, \dots, \lfloor B/\lambda_2 J \rfloor$. By taking the average of $U(K | \mathbf{r}, \mathbf{u})$ over the posterior predictive distribution of $\mathbf{r}$ and $\mathbf{u}$, we obtain the expected utility $U(K)$ as follows:

$$U(K) = \int \int \sum_{\mathbf{u}} U(K | \mathbf{r}, \mathbf{u}) p(\mathbf{r}, \mathbf{u} | \boldsymbol{\xi}) p(\boldsymbol{\xi} | \mathbf{y}) d\mathbf{r} d\boldsymbol{\xi}, \quad (9)$$

where $p(\mathbf{r}, \mathbf{u} | \boldsymbol{\xi})$ represents the joint distribution of $\mathbf{r}$ and $\mathbf{u}$, conditional on $\boldsymbol{\xi}$, as specified by Equations 3 and 4. Note that, as is evident from taking the average over the posterior distribution, $U(K)$ fully accounts for uncertainty in the unknowns of the model. When a large number of posterior samples are available, Equation 9 can be evaluated using the Monte Carlo integration by plugging the posterior predictive samples of u and r into Equation 8 and taking the sample average.

The preferred value of K (and the corresponding value of N) can be identified based on the profile of $U(K)$. Note, however, that in practical applications, the budget amount itself is often considered. In this case, we could obtain a profile of $U(K)$ for each possible budget amount and compare it between the budget levels to guide decision making (see Section 4).

4 Applications in fish eDNA metabarcoding

4.1 Materials and Methods

We applied the proposed model to a study of freshwater fish communities conducted in the Lake Kasumigaura watershed, Japan. The details of data acquisition and model fitting are fully described in Appendix S2. In August 2017, we collected three environmental samples (i.e., replicates) at 50 sites across the watershed (Matsuzaki *et al.* 2019). We used the universal primer set MiFish-U (Miya *et al.* 2015) to amplify a short fragment of fish DNA from the samples and sequenced the amplified DNA fragments using an MiSeq sequencer (Illumina, San Diego, CA, USA). We detected 50 freshwater fish taxa (henceforth referred to as “species”) which are listed in Table S1. In the following analysis, we omitted six of the 150 samples as no sequence reads were obtained. The statistical distribution of the throughputs of the remaining samples was right-skewed (Fig. S1), with a mean of 77,910 and a standard deviation of 98,035.

We analyzed the dataset using the proposed multispecies site occupancy model, with the covariate effects of the degree of primer-template mismatches on ϕ and that of lack of

285 vegetation on ψ . The covariate effects were modeled as follows:

$$\log \phi_i = \alpha_{0i} + \alpha_1 x_i \quad (10)$$

$$\text{logit } \psi_{ij} = \gamma_{0i} + \gamma_{1i} v_j, \quad (11)$$

286 where x_i represents the degree of primer-template mismatches (total number of mismatched
287 bases in the priming region of the forward and reverse primers; Table S1) of species i , and v_j
288 indicates whether the riverbank at site j lacks aquatic and riparian vegetation. We treated α_{0i} ,
289 γ_{0i} , and γ_{1i} as random species effects. The community-level prior distribution for species-level
290 parameters was then specified, as follows, instead of Equation 6:

$$(\alpha_{0i}, \text{logit } \theta_i, \gamma_{0i}, \gamma_{1i}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (12)$$

291 The model was fitted with JAGS software (Plummer 2003), in which we specified a vague prior
292 for each parameter.

293 With the posterior samples, we examined profiles of the efficiency of species detection,
294 as described in Section 3.2. We set $\lambda_1 = 0.01$ JPY and $\lambda_2 = 5,000$ JPY based on
295 approximative estimates of the cost of HTS and library preparation as of January 2019,
296 respectively. We set 875,000 JPY as the baseline budget level, which approximated the actual
297 budget for our study. Given that the maximum number of replicates for this budget level was
298 three, we examined profiles of the efficiency of species detection under two hypothetical
299 scenarios. First, we considered a situation where decisions were made within the range of
300 budget levels under which the maximum number of replicates was fixed at three (scenario 1).
301 Specifically, we compared profiles of the expected number of species detected over a range of
302 budget levels around the baseline, $B = \{775 \times 10^3, 825 \times 10^3, 875 \times 10^3, 925 \times 10^3, 975 \times 10^3\}$.
303 In the second scenario (scenario 2), we considered a situation where decisions were made
304 within a larger range of budget levels with a different maximum number of replicates. We
305 compared profiles of the expected number of species detected over a range of budget levels,
306 $B = \{437.5 \times 10^3, 875 \times 10^3, 1,750 \times 10^3, 2,625 \times 10^3, 3,500 \times 10^3\}$, which corresponds to 0.5,
307 1, 2, 3, and 4 times the baseline budget, respectively. The throughput allocated under each
308 budget level is shown in Table 1. Note that under these budget levels, tens of thousands of
309 sequence reads were secured for all conditions (minimum throughput: 16,667 reads).

310 4.2 Results

311 4.2.1 Species-level parameters

312 Site occupancy probabilities (ψ) varied considerably among species and showed a bimodal
313 distribution, with more species tending to show occupancy probabilities close to one and
314 others close to zero (Fig. 2A). Overall, the occupancy probability of species was reduced in the
315 absence of vegetation (Fig. 2A, orange diamonds). A negative effect of lack of vegetation was
316 identified at the community-level; the posterior median and the 95% highest posterior density

interval (HPDI) of the community mean of γ_1 were -1.039 and $(-1.360, -0.748)$, respectively.

Sequence capture probabilities (θ) also varied among species and tended to be uniformly distributed in the range of about 0.5 to 1 (Fig. 2B).

Sequence relative dominance (ϕ) showed a right-skewed distribution, with a small number of species exhibiting noticeably high values, whereas a large number of species showed low values (Fig. 2C). Species with a higher degree of primer-template mismatches tended to show lower sequence relative dominance (Fig. 2C, bluish circles). The posterior median of α_1 was -0.195 , whereas the 95% HPDI $(-0.468, 0.115)$ overlapped with 0. The species ratio of sequence relative dominance predicted that, on average, and conditional on their presence, the species with the highest sequence dominance (*Rhinogobius* spp.) had 23.5 (95% HPDI; 5.7–54.3) times more sequence reads than the species with the lowest sequence dominance (*Salangichthys microdon*).

Estimates of the correlation coefficients between the species random effects are shown in Table 2. Positive correlations between the intercept of sequence relative dominance (α_{0i}), the sequence capture probability (θ_i), and the intercept of site occupancy probability (γ_{0i}), were evident on the link scales. In contrast, there was no apparent correlation between the slope of site occupancy probability (γ_{1i}) and other random effects.

4.2.2 Species site occupancy

The posterior probability of the site occupancy, $\Pr(z_{ij} = 1 | \mathbf{y})$ of undetected species was lower than 0.15 in most cases (Fig. 3, bluish tiles), indicating that a reasonable effort had been made in species detection. Nevertheless, for some sites and species, we still found higher posterior probabilities of the site occupancy of undetected species (up to ~ 0.3 ; Fig. 3, orangish tiles). The posterior probabilities of site occupancy tended to be higher in species with higher site occupancy probabilities (ψ) and lower sequence relative dominance (ϕ), such as *Channa argus* and *Oryzias latipes*.

4.2.3 Study design analysis

In scenario 1 (Fig. 4A), the efficiency of species detection increased with the number of replicates, reaching a maximum at three replicates for all budget levels considered. At each number of replicates, the efficiency of species detection increased with the amount of budget, indicating that increased throughput (Table 1) also improves the efficiency of species detection. Nevertheless, under the range of budgets considered, the gain from increasing throughput was limited relative to increasing replicates.

In scenario 2 (Fig. 4B), the efficiency of species detection attainable varied significantly between budget levels. When the budget level was half of the baseline, only a single replicate could be taken and the efficiency of species detection was notably limited. When the budget level was several times greater than that of the baseline, it was possible to further improve the efficiency of species detection, which could be achieved by obtaining additional replicates. However, the magnitude of the gain in the efficiency of species detection rapidly decreased as

the number of replicates increased. Thus, there are few additional benefits to spending more than three times the baseline budget.

5 Discussion

In this study, we proposed a novel variant of the multispecies site occupancy model (Dorazio & Royle 2005, Dorazio *et al.* 2006) for spatially-replicated eDNA metabarcoding that is directly applicable to sequence read counts, the unique output of recent HTS technology. With a series of submodels describing ecological and observation processes, the model describes the multi-stage process of species detection using eDNA metabarcoding and allows for the decomposition of the variability of sequence reads, in addition to the inference of site occupancy by species (Fig. 1). Furthermore, we introduced a Bayesian decision analysis framework of study designs for eDNA metabarcoding. These approaches are expected to make eDNA metabarcoding more efficient and robust for the imperfect detection of species.

In previous studies of eDNA metabarcoding, multispecies site occupancy models have been applied by aggregating sequence reads into binary detection/non-detection data (Doi *et al.* 2019, Bush *et al.* 2020, McClenaghan *et al.* 2020, McColl-Gausden *et al.* 2020). In such an approach, sequence read counts are not explicitly modeled, making it difficult to adequately decompose the combined effects of throughput and relative frequency of species sequences on the probability of species detection. By adopting a multinomial observation component, the proposed model overcomes this issue and thus provides a detailed description of the process of species detection using eDNA metabarcoding. The inference of the model is feasible with general-purpose software such as JAGS (Plummer 2003), as with the traditional multispecies site occupancy model (Dorazio & Royle 2005, Dorazio *et al.* 2006). The development of an R package is planned to make the methods introduced in this study even more convenient for researchers and practitioners to use.

An application to freshwater fish communities in the Lake Kasumigaura watershed highlighted the heterogeneity of the species-level parameters (Fig. 2). Covariate modeling revealed a tendency for species with a higher degree of primer-template mismatches to exhibit lower sequence relative dominance. Although the 95% HPDI of the coefficient for the effect of primer-template mismatch (α_1) included zero, the observed tendency was consistent with previous results showing that primer-template mismatches can have a strong negative effect on the relative abundance of sequence reads owing to biased PCR amplification (Piñol *et al.* 2015, Shelton *et al.* 2016). The estimates of sequence relative dominance predict that there is a large inequality in the allocation of sequence reads among species; given that the majority of sequence reads are assigned to a small number of species, a false-negative detection error of a disproportionately large number of less dominant species is expected to occur at the stage of HTS when throughput is limited. In addition, the correlations of species random effects (Table 2) imply that rare species with lower site occupancy probability tend to be difficult to detect when they are present, compared to the detection of common species with higher occupancy probability. These results, revealed by the use of the proposed model, underscore the large

inhomogeneity in eDNA metabarcoding, which can result in the biased detection of specific species.

Analysis of the study design suggested that multiple replicates can result in better efficiency in species detection. In the eDNA metabarcoding of freshwater fish communities in the Lake Kasumigaura watershed, the number of species detected per site was expected to increase with an increase in the number of replicates (Fig. 4), as long as a throughput of tens of thousands of sequence reads was secured (Table 1). In reality, obtaining a small number of replicates can also be an efficient strategy for assessing regional species diversity (Fig. S2). We note, however, that this finding may be specific to the application of the model to the freshwater fish communities in the Lake Kasumigaura watershed. It is likely that the optimal balance between throughput and the number of replicates (and even the number of sites) can vary depending on, for example, the distribution of species-level parameters, the costs required for different stages of the workflow, or the amount of research budget available. Therefore, a further examination into different systems and conditions is required.

In the Bayesian decision analysis framework for the study design, we used the posterior distribution $p(\boldsymbol{\xi} | \mathbf{y})$ to obtain the expected utility $U(K)$ (Equation 9). This approach allowed for rational decisions based on the dataset obtained. In contrast, when a dataset is not available, a prior distribution $p(\boldsymbol{\xi})$, instead of a posterior distribution $p(\boldsymbol{\xi} | \mathbf{y})$, can be used to obtain the expected utility (Chaloner & Verdinelli 1995, Guillera-Arroita *et al.* 2014). Such a treatment could be used to design preliminary research and develop a sequential design approach as addressed earlier (Guillera-Arroita *et al.* 2014).

In eDNA metabarcoding, false positives can occur because of various technical factors, including contamination, amplification or sequencing errors, or tag switching (Zhan & MacIsaac 2015, Furlan *et al.* 2020). As unmodeled false positives can cause a significant bias in the inference of site occupancy (Royle & Link 2006), adequate precautions must be taken in applying the proposed model to an eDNA metabarcoding study. We recommend that efforts be made to reduce as many false positives as possible by, for example, performing careful experimental manipulations, utilizing negative controls, conducting conservative quality filtering, and using clustering and de-noising approaches. Generally, reducing false positives can lead to an increase in false negatives (Zhan *et al.* 2014b, Zhan & MacIsaac 2015). It would be reasonable, however, to make more efforts to avoid generating false positives than false negatives during the process of data acquisition as the proposed model formally accounts for false negatives under the assumption of the absence of false positives. It should be noted that the model never corrects false negatives caused by incomplete reference databases. As a complementary approach, further research is warranted to extend the model to account for false positives in addition to false negatives.

Acknowledgements

We are grateful to the following people for their support in the fieldwork and laboratory experiments: Ayato Kohzu, Mirai Watanabe, Kazuhiro Komatsu, Haruyo Yamaguchi, Haruko

Ando, Koichi Shimotori, Sayuri Nakamura, Megumi Nakagawa, Megumi Yoshida, Junko Hiramata, Tomomi Shindo, Junko Machizawa, and Nobuyoshi Nakajima. We thank I. K. Shimatani for providing valuable feedback on the earlier version of this manuscript. Funding was provided by the Environmental Restoration and Conservation Agency (ERTDF, No. 4–1705) and the Japan Society for the Promotion of Science (KAKENHI, Nos 20H03010 and 20K06102). This research was supported by allocation of computing resources of the HPE SGI 8600 supercomputer from the Institute of Statistical Mathematics. The authors have no conflicts of interest to declare.

Author Contributions

KF and TK conceived the study and designed the methodology; SSM designed and organized the fieldwork; NK conducted eDNA metabarcoding and organized the dataset; KF analyzed the dataset; KF and NK drafted the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Data Availability Statement

The raw sequence reads will be archived in the DDBJ Sequence Read Archive. The dataset and codes will be archived in the Dryad Digital Repository upon acceptance.

References

- Bailey, L.L., Hines, J.E., Nichols, J.D. & MacKenzie, D.I. (2007) Sampling design trade-offs in occupancy studies with imperfect detection: examples and software. *Ecological Applications*, **17**, 281–290.
- Bush, A., Monk, W.A., Compson, Z.G., Peters, D.L., Porter, T.M., Shokralla, S., Wright, M.T.G., Hajibabaei, M. & Baird, D.J. (2020) DNA metabarcoding reveals metacommunity dynamics in a threatened boreal wetland wilderness. *Proceedings of the National Academy of Science*, **117**, 8539–8545.
- Bylemans, J., Gleeson, D.M., Duncan, R.P., Hardy, C.M. & Furlan, E.M. (2019) A performance evaluation of targeted eDNA and eDNA metabarcoding analyses for freshwater fishes. *Environmental DNA*, **1**, 402–414.
- Chaloner, K. & Verdinelli, I. (1995) Bayesian experimental design: a review. *Statistical Science*, **10**, 273–304.
- Deiner, K., Bik, H.M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D.M., de Vere, N., Pfrender, M.E. & Bernatchez, L. (2017) Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Molecular Ecology*, **26**, 5872–5895.

- Doi, H., Fukaya, K., Oka, S., Sato, K., Kondoh, M. & Miya, M. (2019) Evaluation of detection probabilities at the water-filtering and initial PCR steps in environmental DNA metabarcoding using a multispecies site occupancy model. *Scientific Reports*, **9**, 3581.
- Dorazio, R.M. & Johnson, F.A. (2003) Bayesian inference and decision theory — a framework for decision making in natural resource management. *Ecological Applications*, **13**, 556–563.
- Dorazio, R.M. & Royle, J.A. (2005) Estimating size and composition of biological communities by modeling the occurrence of species. *Journal of the American Statistical Association*, **100**, 389–398.
- Dorazio, R.M., Royle, J.A., Söderström, B. & Glimskär, A. (2006) Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology*, **87**, 842–854.
- Furlan, E.M., Davis, J. & Duncan, R.P. (2020) Identifying error and accurately interpreting environmental DNA metabarcoding results: a case study to detect vertebrates at arid zone waterholes. *Molecular Ecology Resources*, **20**, 1259–1276.
- Guillera-Aroita, G., Ridout, M.S. & Morgan, B.J. (2010) Design of occupancy studies with imperfect detection. *Methods in Ecology and Evolution*, **1**, 131–139.
- Guillera-Aroita, G., Ridout, M.S. & Morgan, B.J. (2014) Two-stage Bayesian study design for species occupancy estimation. *Journal of Agricultural, Biological, and Environmental Statistics*, **19**, 278–291.
- Harper, L.R., Lawson Handley, L., Hahn, C., Boonham, N., Rees, H.C., Gough, K.C., Lewis, E., Adams, I.P., Brotherton, P., Phillips, S. & Hänfling, B. (2018) Needle in a haystack? A comparison of eDNA metabarcoding and targeted qPCR for detection of the great crested newt (*Triturus cristatus*). *Ecology and Evolution*, **8**, 6330–6341.
- Harrison, J.G., Calder, W.J., Shastri, V. & Buerkle, C.A. (2020) Dirichlet-multinomial modelling outperforms alternatives for analysis of microbiome and other ecological count data. *Molecular Ecology Resources*, **20**, 481–497.
- Iknayan, K.J., Tingley, M.W., Furnas, B.J. & Beissinger, S.R. (2014) Detecting diversity: emerging methods to estimate species diversity. *Trends in Ecology and Evolution*, **29**, 97–106.
- Lugg, W.H., Griffiths, J., van Rooyen, A.R., Weeks, A.R. & Tingley, R. (2018) Optimal survey designs for environmental DNA sampling. *Methods in Ecology and Evolution*, **9**, 1049–1059.
- MacKenzie, D.I. & Royle, J.A. (2005) Designing occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology*, **42**, 1105–1114.
- Matsuzaki, S.S., Kohzu, A., Kadoya, T., Watanabe, M., Osawa, T., Fukaya, K., Komatsu, K., Kondo, N., Yamaguchi, H., Ando, H., Shimotori, K., Nakagawa, M., Kizuka, T., Yoshioka, A., Sasai, T., Saigusa, N., Matsushita, B. & Takamura, N. (2019) Role of wetlands in

mitigating the trade-off between crop production and water quality in agricultural
landscapes. *Ecosphere*, **10**, e02918.

McClenaghan, B., Compson, Z.G. & Hajibabaei, M. (2020) Validating metabarcoding-based
biodiversity assessments with multi-species occupancy models: a case study using coastal
marine eDNA. *PLoS ONE*, **15**, e0224119.

McColl-Gausden, E.F., Weeks, A.R., Coleman, R.A., Robinson, K.L., Song, S., Raadik, T.A. &
Tingley, R. (2020) Multi-species models reveal that eDNA metabarcoding is more sensitive
than backpack electrofishing for conducting fish surveys in freshwater streams. *Molecular
Ecology (in press)*.

Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J.Y., Sato, K., Minamoto, T.,
Yamamoto, S., Yamanaka, H., Araki, H., Kondoh, M. & Iwasaki, W. (2015) MiFish, a set of
universal PCR primers for metabarcoding environmental DNA from fishes: detection of
more than 230 subtropical marine species. *Royal Society Open Science*, **2**, 150088.

Nukazawa, K., Hamasuna, Y. & Suzuki, Y. (2018) Simulating the advection and degradation
of the environmental DNA of common carp along a river. *Environmental Science and
Technology*, **52**, 10562–10570.

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

Plummer, M. (2003) JAGS: a program for analysis of Bayesian graphical models using Gibbs
sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical
Computing*, volume 124, pp. 1–10. Technische Universit at Wien, Austria.

Royle, J.A. & Link, W.A. (2006) Generalized site occupancy models allowing for false positive
and false negative errors. *Ecology*, **87**, 835–841.

Sanderlin, J.S., Block, W.M. & Ganey, J.L. (2014) Optimizing study design for multi-species
avian monitoring programmes. *Journal of Applied Ecology*, **51**, 860–870.

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

Shokralla, S., Spall, J.L., Gibson, J.F. & Hajibabaei, M. (2012) Next-generation sequencing
technologies for environmental DNA research. *Molecular Ecology*, **21**, 1794–1805.

Smart, A.S., Tingley, R., Weeks, A.R., van Rooyen, A.R. & McCarthy, M.A. (2015)
Environmental DNA sampling is more sensitive than a traditional survey technique for
detecting an aquatic invader. *Ecological Applications*, **25**, 1944–1952.

- 535 Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. (2012) Towards
536 next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, **21**,
537 2045–2050.
- 538 Tang, Z.Z. & Chen, G. (2019) Zero-inflated generalized Dirichlet multinomial regression model
539 for microbiome compositional data analysis. *Biostatistics*, **20**, 698–713.
- 540 Yamamoto, S., Masuda, R., Sato, Y., Sado, T., Araki, H., Kondoh, M., Minamoto, T. & Miya,
541 M. (2017) Environmental DNA metabarcoding reveals local fish communities in a
542 species-rich coastal sea. *Scientific Reports*, **7**, 40368.
- 543 Zhan, A., He, S., Brown, E.A., Chain, F.J.J., Therriault, T.W., Abbott, C.L., Heath, D.D.,
544 Cristescu, M.E. & MacIsaac, H.J. (2014a) Reproducibility of pyrosequencing data for
545 biodiversity assessment in complex communities. *Methods in Ecology and Evolution*, **5**,
546 881–890.
- 547 Zhan, A. & MacIsaac, H.J. (2015) Rare biosphere exploration using high-throughput
548 sequencing: research progress and perspectives. *Conservation Genetics*, **16**, 513–522.
- 549 Zhan, A., Xiong, W., He, S. & MacIsaac, H.J. (2014b) Influence of artifact removal on rare
550 species recovery in natural complex communities using high-throughput sequencing. *PLoS*
551 *ONE*, **9**, e96928.

552 **Tables and figures**

Table 1. Values of throughput under two scenarios for the study design analysis of eDNA metabarcoding on freshwater fish communities in the Lake Kasumigaura watershed. The throughput is obtained as $N = (B - \lambda_2 JK) / \lambda_1 JK$ for $K = 1, 2, \dots, \lfloor B / \lambda_2 J \rfloor$ where we set $J = 50$ sites, $\lambda_1 = 0.01$ JPY, and $\lambda_2 = 5,000$ JPY; see Sections 3.2 and 4.1. The unit of budget is 10^3 JPY.

Number of replicates (K)	Budget (B)												
	Scenario 1						Scenario 2						
	775	825	875	925	975	437.5	875	1,750	2,625	3,500			
1	1,050,000	1,150,000	1,250,000	1,350,000	1,450,000	375,000	1,250,000	3,000,000	4,750,000	6,500,000			
2	275,000	325,000	375,000	425,000	475,000	—	375,000	1,250,000	2,125,000	3,000,000			
3	16,667	50,000	83,333	116,667	150,000	—	83,333	666,667	1,250,000	1,833,333			
4	—	—	—	—	—	—	—	375,000	812,500	1,250,000			
5	—	—	—	—	—	—	—	200,000	550,000	900,000			
6	—	—	—	—	—	—	—	83,333	375,000	666,667			
7	—	—	—	—	—	—	—	—	250,000	500,000			
8	—	—	—	—	—	—	—	—	156,250	375,000			
9	—	—	—	—	—	—	—	—	83,333	277,778			
10	—	—	—	—	—	—	—	—	25,000	200,000			
11	—	—	—	—	—	—	—	—	—	136,364			
12	—	—	—	—	—	—	—	—	—	83,333			
13	—	—	—	—	—	—	—	—	—	38,462			

Table 2. Posterior summary of correlation coefficients between the four species random effects. The posterior median and the 95 % HPDI (in brackets) are shown. Bold type indicates that the 95 % HPDI does not include zero.

	α_{0i}	logit θ_i	γ_{0i}	γ_{1i}
α_{0i}	—	—	—	—
logit θ_i	0.62 (0.17, 0.93)	—	—	—
γ_{0i}	0.82 (0.65, 0.95)	0.78 (0.49, 0.97)	—	—
γ_{1i}	-0.47 (-0.94, 0.14)	-0.21 (-0.78, 0.47)	-0.51 (-0.91, 0.16)	—

Fig. 1. A diagram of the multispecies site occupancy model for spatially-replicated eDNA metabarcoding. A hierarchical dependence structure, characterized by a set of species-level parameters (occupancy probability ψ , sequence capture probability θ , and relative dominance of sequence ϕ), is assumed between the latent variables z (site occupancy of species), u (presence or absence of DNA sequences of species in the replicate samples), and data y (sequence read counts). The sequence reads are assumed to follow a multinomial distribution that is conditional on the presence or absence (u) and relative dominance (r) of the sequences of all species, in addition to throughput. Species-level parameters are assumed to follow a community-level normal prior distribution with a mean vector μ and a covariance matrix Σ .

Fig. 2. Estimates of species-level parameters: (A) site occupancy probabilities (ψ), (B) sequence capture probabilities (θ), and (C) sequence relative dominance (ϕ) for freshwater fishes in the Lake Kasumigaura watershed. In the upper panels, the posterior median and 95% HPDI for each species are represented by circles and lines, respectively. For site occupancy probability (A), presence ($v_j = 0$) and absence ($v_j = 1$) of vegetation at the site are distinguished by different symbols and colors. For the latter, we display only the posterior medians for visualization. For sequence relative dominance (C), the degree of primer-template mismatches is indicated by different colors. Species are listed in order of site occupancy probabilities. Lower panels show histograms of the posterior medians. For site occupancy probability (A), the histogram indicates the statistical distribution of posterior medians for vegetated sites.

Fig. 3. Tile plot of the posterior probabilities of species site occupancy. Gray tiles indicate that the species is detected; thus, the posterior probability is one.

Fig. 4. Profiles of the efficiency of species detection under specified amounts of research budgets for the eDNA metabarcoding of freshwater fish communities in the Lake Kasumigaura watershed. The efficiency of species detection is expressed by the expected number of species detected per site (see Section 3.2). The baseline profile is indicated by the black line. See Table 1 for the values of throughput for each condition.