TITLE: NeMO: a flexible R package for nested multi-species occupancy modeling and eDNA study optimization

Bastien Macé[1,2], Stéphanie Manel[1,3], Alice Valentini[2], Mathieu Rocle[4], Nicolas Roset[5], Erwan Delrieu-Trottin[1]
[1]CEFE, Univ Montpellier, CNRS, EPHE-PSL University, IRD, Montpellier, France
[2]SPYGEN, Le-Bourget-du-Lac, France
[3]Institut Universitaire de France, Paris, France
[4]Compagnie Nationale du Rhône, Lyon, France
[5]Office Français de la Biodiversité, Lyon, France

Corresponding author: Bastien Macé

ABSTRACT

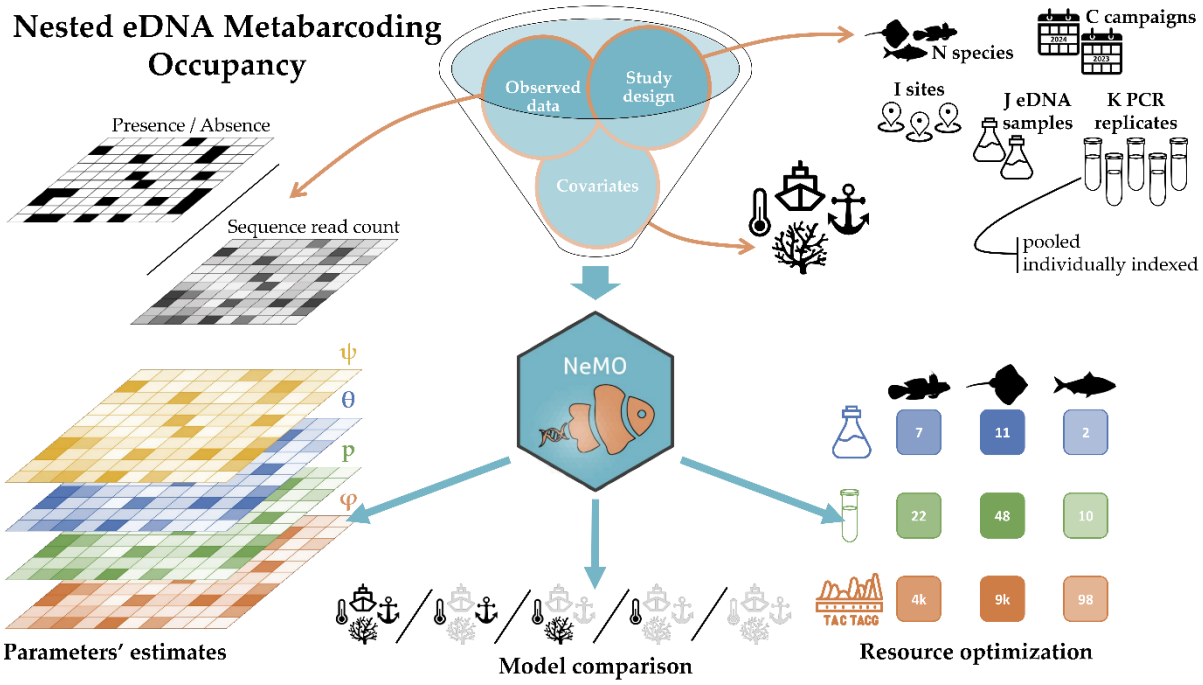
Biodiversity monitoring using environmental DNA (eDNA) metabarcoding has expanded rapidly, providing a non-invasive tool widely adopted by ecologists and stakeholders. However, eDNA surveys are prone to imperfect detection, and non-detections are often misinterpreted as true absences – a critical issue when monitoring rare or elusive species. Despite its implications for biodiversity assessments, detection uncertainty is rarely quantified in eDNA-based studies. Occupancy modeling offers a powerful solution to this limitation but remains underused, partly due to a lack of accessible and flexible tools. We developed NeMO (Nested eDNA Metabarcoding Occupancy), a user-friendly R package for fitting multi-species occupancy models in a Bayesian framework. NeMO explicitly accounts for the nested structure of eDNA metabarcoding workflows – typically involving multiple replication steps such as field samples and PCR replicates – while accommodating presence/absence or read-count data. The framework estimates species occupancy, eDNA collection probability, amplification probability, and expected read counts, and allows users to assess the influence of environmental or methodological covariates on each process. Crucially, NeMO helps rigorously assess detectability and optimize resource allocation in eDNA surveys. It estimates the minimum number of eDNA samples, PCR replicates, and sequencing depth required to reliably detect species when present, thereby guiding study design. We illustrate its utility using a fish biodiversity dataset from the Rhône River (France). NeMO integrates key modelling features into a single streamlined framework, providing researchers and practitioners with an accessible and effective tool to assess detectability and optimize resource allocation in eDNA metabarcoding surveys. Our results highlight the importance of quantifying detection uncertainty, which

has major implications for conservation monitoring and for designing cost-effective and reliable eDNA strategies.

KEY WORDS

Bayesian modeling, detection, elusive species, environmental DNA, false negatives, metabarcoding

GRAPHICAL ABSTRACT



`NeMO` provides a flexible framework for modeling multi-species site occupancy from eDNA data, estimating species presence and detectability despite imperfect detection. It supports diverse experimental designs, enables model comparison, and guides study design optimization by calculating minimal sampling effort to detect species, helping ecologists balance detection accuracy and resource allocation.

1. INTRODUCTION

Environmental DNA (eDNA) metabarcoding has emerged as a powerful tool for biodiversity monitoring (Blackman *et al.*, 2024). Unlike traditional survey methods, which often rely on direct observations or specimen collection, eDNA metabarcoding enables the detection of a wide range of species from environmental samples such as soil, water, or air (Taberlet *et al.*, 2012). This approach facilitates biodiversity assessments at the community level, providing a comprehensive view of ecosystem composition (Macé *et al.*, 2024; Stat *et al.*, 2017), including elusive species that conventional surveys may overlook (Beng & Corlett, 2020; Bohmann *et al.*, 2014).

Like any biodiversity assessment approach, eDNA metabarcoding can be prone to detection errors (Cristescu & Hebert, 2018; Darling & Mahon, 2011), with false positives (*i.e.*, species detected despite being absent from the sampled environment) and false negatives (*i.e.*, species not detected despite being present in the sampled environment). False positives can result from contaminations, PCR or sequencing errors (Ficetola *et al.*, 2016), or misidentifications due to incomplete or inaccurate reference databases (Polanco F. *et al.*, 2025). To mitigate this, stringent laboratory protocols (Ficetola *et al.*, 2016), including extraction blanks, PCR and sequencing controls, and bioinformatic filtering pipelines (Mathon *et al.*, 2021), are routinely implemented. Various factors contribute to false negatives, such as low DNA concentration (Barnes *et al.*, 2014), PCR primer biases during amplification (Alberdi *et al.*, 2018), or insufficient sampling effort (Ficetola *et al.*, 2015). Yet, unlike false positives, false negatives are often overlooked in eDNA studies. Absence data is frequently interpreted as true absences, without considering the possibility of undetected presence – an oversight that carries significant ecological implications (Biggs *et al.*, 2015; Pinfield *et al.*, 2019). Indeed, undetected species may lead to underestimated invasion dynamics, inadequate conservation efforts, or distorted community composition assessments.

Occupancy modeling offers a statistical framework to account for imperfect detection by estimating both occupancy and detection probabilities (MacKenzie *et al.*, 2006). Originally developed for traditional field surveys (MacKenzie *et al.*, 2002; Royle & Dorazio, 2008) relying on direct or indirect censuses, occupancy models have since been adapted for eDNA-based studies. These typically follow a hierarchical structure, incorporating multiple nested levels of replication (Takahashi *et al.*,

2023). Specifically, the spatial units under investigation (*i.e.*, sites) are generally sampled through different field replicates (*i.e.*, eDNA samples), with each of these analyzed using multiple PCR replicates. Recognizing this structure, Schmidt *et al.* (2013) introduced the first nested occupancy model for single-species eDNA studies, estimating occupancy probability alongside detection probabilities at both the eDNA sample and PCR replicate levels. Building on this foundation, Dorazio & Erickson (2018) refined the approach in their `eDNAoccupancy` R package. Although this package improves accessibility and flexibility, the routine application of occupancy models remains limited in eDNA research, especially in multi-species metabarcoding studies, where non-detections are still frequently interpreted as true absences.

To extend the capabilities of existing models, occupancy frameworks tailored to eDNA metabarcoding studies have recently been introduced (**Table 1**), adapting the multi-species model from Dorazio & Royle (2005). While most of these models were made publicly available as text files, they were not initially designed as user-friendly software, which limited their accessibility to the broader scientific community. A major step forward was made with the development of dedicated R packages for multi-species occupancy modeling in eDNA studies. Fukaya *et al.* (2022) and Diana *et al.* (2024) introduced the `occumb` and `eDNAPlus` packages, respectively, both of which also integrating an important methodological advancement, *i.e.*, the use of sequence read counts instead of solely presence/absence data. By incorporating quantitative sequencing information, these models refine detection probability estimates while keeping some limitations: `occumb` does not account for PCR replication while both are not allowing model comparisons. Building upon these advances, we introduce `NeMO` (Nested eDNA Metabarcoding Occupancy), a versatile R package designed to fully capture the complexity of eDNA metabarcoding datasets. `NeMO` accommodates a broad range of data structures, whether sequence read counts are available or only presence/absence data is provided. It also offers flexibility in handling different PCR replication schemes, whether replicates are individually indexed or pooled before sequencing. By integrating these features, `NeMO` enables users to choose the most suitable model based on the structure and requirements of their dataset. Like the previous occupancy modeling packages, `NeMO` allows the inclusion of covariates at multiple levels, enabling users to assess how

environmental and methodological factors influence occupancy and detection probabilities. Additionally, we implement a model comparison tool, a feature present in `eDNAoccupancy` but not yet incorporated into multi-species occupancy frameworks. One key application of occupancy models is optimizing resource allocation in eDNA studies by estimating the minimum number of eDNA samples and PCR replicates needed to achieve a high probability of detecting species when present, thus explicitly accounting for imperfect detection. `NeMO` facilitates these calculations using equations from McArdle (1990) and extends this functionality by incorporating the option to estimate the minimum sequencing depth required for reliable species detection.

To illustrate `NeMO` capabilities, we apply it to an empirical dataset from Pont *et al.* (2018), demonstrating its usefulness for ecological applications. Our goal is to provide an accessible, flexible, and powerful tool for scientists and stakeholders working with eDNA data.

2. MATERIAL & METHODS

2.1. Description of the nested eDNA metabarcoding occupancy model

eDNA-based surveys are typically conducted at multiple sites, with replication achieved by collecting several eDNA samples per site and running multiple PCRs for each sample. eDNA metabarcoding study designs therefore encompass three nested levels: site, eDNA sample, and PCR replicate (**Fig. 1a**). Additionally, high-throughput sequencing (HTS), which is commonly used to sequence eDNA barcodes (Taberlet *et al.*, 2018), allows generation of sequence read counts for each produced amplicon (*i.e.* DNA fragment from a specific target region that has been amplified and sequenced), and ultimately for each species detected, provided that sequences are successfully assigned. Building on this framework, we adapted models proposed in Dorazio & Erickson (2018) and Fukaya *et al.* (2022) to fit multi-species occupancy and estimate parameters related to i) the probability of species eDNA occurrence at a site, ii) the conditional probability of collecting eDNA from a species in a sample from a site, given that its eDNA is present at the site, iii) the conditional probability of amplifying eDNA from a species in a PCR replicate of a sample, given that its eDNA is collected in the sample, and iv) the expected sequence read count generated for the species. To ensure consistency with existing literature, we adhere as closely as

possible to the mathematical notations introduced by Schmidt *et al.* (2013) and Fukaya *et al.* (2022) for occupancy modeling in the context of the nested eDNA framework.

We consider a community of N entities ($N \in \mathbb{N}^*$), referred to as species for convenience, but which may also refer to taxa, operational taxonomic units (OTUs), amplicons, etc. Biodiversity monitoring of this community was repeated over time across C campaigns ($C \in \mathbb{N}^*$), covering I_c sites ($I_c \in \mathbb{N}^*$), where J_i samples ($J_{ic} \in \mathbb{N}^*$) were collected per site and divided into K_{ijc} PCR replicates ($K_{ijc} \in \mathbb{N}^*$). For each PCR replicate, HTS library preparation generated separate sequence reads. In this framework, the occupancy of species $n \in \{1, \ldots, N\}$ at site $i \in \{1, \ldots, I_c\}$ during campaign $c \in \{1, \ldots, C\}$ is represented by the binary latent variable Z_{nic} which indicates whether eDNA from the species is present ($Z_{nic} = 1$) or absent ($Z_{nic} = 0$). This follows a Bernoulli distribution:

$$Z_{nic} \sim Bernoulli(\psi_{nic}) \tag{1}$$

where $\psi_{nic} \in [0,1]$ is the probability that eDNA from species n being present at site i during campaign c. The detection of species n in sample $j \in \{1, \ldots, J_{ic}\}$ from site i during campaign c is represented by the binary latent variable A_{nijc} which indicates whether eDNA from the species is collected ($A_{nijc} = 1$) or not ($A_{nijc} = 0$). This also follows a Bernoulli distribution, conditional on occupancy:

$$A_{nijc} | Z_{nic} \sim Bernoulli\left(Z_{nic} \cdot \theta_{nijc}\right) \tag{2}$$

where $\theta_{nijc} \in [0,1]$ is the conditional probability of collecting eDNA from species n in sample j from site i during campaign c, given that species' eDNA is present at site i. The detection of species n in PCR replicate $k \in \{1, \ldots, K_{ijc}\}$ from sample j from site i during campaign c is represented by the binary latent variable W_{nijkc} which indicates whether eDNA from the species is amplified ($W_{nijkc} = 1$) or not ($W_{nijkc} = 0$). This also follows a Bernoulli distribution, conditional on eDNA collection:

$$W_{nijkc} | A_{nijc} \sim Bernoulli\left(A_{nijc} \cdot p_{nijkc}\right) \tag{3}$$

where $p_{nijkc} \in [0,1]$ is the conditional probability of amplifying eDNA from species n in PCR replicate k from sample j from site i during campaign c, given that species' eDNA was collected in sample j. The sequence read counts resulting from HTS are the observed variable Y_{nijkc}. These counts are

conditional on the per-PCR replicate total sequence read count $M_{ijkc} \in \mathbb{N}$ hereafter referred to as 'sequencing depth' for simplicity. The sequencing depth is defined as:

$$M_{ijkc} = \sum_{n=1}^{N} Y_{nijkc} \tag{4}$$

For i, j, k, c where $M_{ijkc} > 0$, the sequence read counts Y_{nijkc} follows a multinomial distribution, conditional on eDNA amplification:

$$Y_{nijkc} | W_{nijkc} \sim Multinomial(\pi_{nijkc}, M_{ijkc}) \tag{5}$$

where $\pi_{nijkc} \in [0,1]$ is the conditional probability of assigning a sequencing read to species n in PCR replicate k from sample j from site i during campaign c, given that species' eDNA was amplified in PCR replicate k. This parameter expresses the relative frequency of the sequence read count for species n under these conditions. It is defined as:

$$\pi_{nijkc} = \frac{W_{nijkc} \cdot S_{nijkc}}{max(1, \sum_{m=1}^{N} W_{mijkc} \cdot S_{mijkc})} \tag{6}$$

where the latent variable $S_{nijkc} \in \mathbb{N}$ is the sequence read count for species n in PCR replicate k from sample j from site i during campaign c, and W_{nijkc} is the amplification detection indicator for the same replicate. The denominator ensures normalization of probabilities and accounts for cases where the product $W_{mijkc} \cdot S_{mijkc}$ is null (*e.g.*, no amplification or zero read count for all species in the replicate). In such cases, the denominator defaults to 1. To accommodate the discrete and overdispersed nature of HTS data, we model the latent sequence read count S_{nijkc} using a negative binomial distribution:

$$S_{nijkc} \sim NB\left(\frac{\delta}{\delta + \varphi_{nijkc}}, \delta\right) \tag{7}$$

where $\varphi_{nijkc} \in \mathbb{N}$ is the expected sequence read count for species n in PCR replicate k from sample j from site i during campaign c, and δ is the user-defined dispersion parameter controlling the variance of the distribution (defaults to 1). The framework of this occupancy model is summarized in **Fig. 1b.**

The species-level parameters ψ_{nic}, θ_{nijc}, p_{nijkc}, and φ_{nijkc} are modeled as random effects. Each parameter includes a species-specific intercept α and species-specific slope coefficients β associated with covariates that can be incorporated into the model. The linear predictors are defined as:

$$logit(\psi_{nic}) = \alpha_{\psi_n} + \sum_{d=1}^{D} \beta_{\psi_d}^{(sp)} \cdot cov_{\psi_{n,d}}^{(sp)} + \sum_{h=1}^{H} \beta_{\psi_{n,h}} \cdot cov_{\psi_{ic,h}} \tag{8.1}$$

$$logit(\theta_{nijc}) = \alpha_{\theta_n} + \sum_{e=1}^{E} \beta_{\theta_e}^{(sp)} \cdot cov_{\theta_{n,e}}^{(sp)} + \sum_{l=1}^{L} \beta_{\theta_{n,l}} \cdot cov_{\theta_{ijc,l}} \tag{8.2}$$

$$logit(p_{nijkc}) = \alpha_{p_n} + \sum_{f=1}^{F} \beta_{p_f}^{(sp)} \cdot cov_{p_{n,f}}^{(sp)} + \sum_{r=1}^{R} \beta_{p_{n,r}} \cdot cov_{p_{ijkc,r}} \tag{8.3}$$

$$ln(\varphi_{nijkc}) = \alpha_{\varphi_n} + \sum_{g=1}^{G} \beta_{\varphi_g}^{(sp)} \cdot cov_{\varphi_{n,g}}^{(sp)} + \sum_{u=1}^{U} \beta_{\varphi_{n,u}} \cdot cov_{\varphi_{ijkc,u}} \tag{8.4}$$

Here:

- $cov_{\psi_{n,d}}^{(sp)}$, $cov_{\theta_{n,e}}^{(sp)}$, $cov_{p_{n,f}}^{(sp)}$, and $cov_{\varphi_{n,g}}^{(sp)}$ are Boolean or standardized species-level covariates associated with species n;
- $cov_{\psi_{ic,h}}$, $cov_{\theta_{ijc,l}}$, $cov_{p_{ijkc,r}}$, and $cov_{\varphi_{ijkc,u}}$ are Boolean or standardized spatial, methodological, or temporal covariate;
- $D, E, F, G \in \mathbb{N}$ are the numbers of species-level covariates;
- $H, L, R, U \in \mathbb{N}$ are the numbers of spatial, methodological, or temporal covariates.

Species-level slopes associated with species covariates follow zero-centered normal priors:

$$\begin{gathered} \beta_{\kappa_q}^{(sp)} \sim \mathcal{N}(0, \tau) \\ \kappa \in \{\psi, \theta, p, \varphi\}, \quad q = 1, \ldots, Q_\kappa^{(sp)}, \quad Q_\kappa^{(sp)} \in \{D, E, F, G\} \end{gathered} \tag{9}$$

where $\tau > 0$ is the user-defined hyperparameter controlling the precision (defaults to 1). Intercepts and slopes associated with spatial, methodological, or temporal covariates follow a multivariate normal community-level prior:

$$\begin{gathered} [\alpha_{\kappa_n} \ldots \beta_{\kappa_{n,q}}] \sim \mathcal{N}(\mu, \Omega) \\ \kappa \in \{\psi, \theta, p, \varphi\}, \quad q = 1, \ldots, Q_\kappa^{(sp)}, \quad Q_\kappa^{(sp)} \in \{D, E, F, G\} \end{gathered} \tag{10}$$

where μ is the community-mean vector and Ω the positive-definite community covariance matrix, capturing the variance of each parameter and the covariances between them. The community-mean vector μ follows a normal hyperprior:

$$\mu \sim \mathcal{N}(0, \tau) \tag{11}$$

where $\tau > 0$ is the user-defined hyperparameter controlling the precision (defaults to 1). The community covariance matrix Ω follows as Wishart hyperprior:

$$\Omega \sim \mathcal{W}(R, \upsilon) \tag{12}$$

where R is a symmetric scale matrix with elements $R_{q_1,q_2} = 1$ if $q_1 = q_2$ and $R_{q_1,q_2} = 0$ if $q_1 \neq q_2$, determining the shape of the Wishart distribution, and $\upsilon \in \mathbb{N}^*$ is the number of random effects.

Species are modeled as conditionally independent at the ecological levels (latent variables Z, A, W, S), but the associated parameters covary through a multivariate normal distribution (**Equation 10**), enabling borrowing of strength across the community. At the observation level (Y), the multinomial distribution creates negative correlations between species read counts due to the normalization and fixed sequencing depth. This compositional constraint reflects the technical reality of high-throughput sequencing, where species compete for a fixed pool of reads. Including species co-occurrence patterns in the modeling framework would probably provide a more detailed picture of species occupancy, but were not included here to avoid adding more ramifications to an already complex hierarchical model.

2.2. `NeMO` R package and associated functions

To analyze empirical data using our proposed model, we employed a Bayesian framework with Markov Chain Monte Carlo (MCMC) sampling, implemented with JAGS (Plummer, 2003). We developed a new R package, `NeMO` (Nested eDNA Metabarcoding Occupancy; available on GitHub: https://github.com/bastien-mace/NeMO), to facilitate this process. This package is designed to streamline the analysis of diverse eDNA metabarcoding experimental designs. Our proposed model comprehensively accounts for PCR replicates that are individually indexed for sequencing, a scenario that provides detailed information on presence/absence and sequence read count at the level of individual PCR replicates. However, pooling PCR replicates is a common strategy to reduce costs (Taberlet *et al.*, 2018), but it limits the available information to summarized data. Additionally, while incorporating sequence read counts into occupancy modeling offers a more detailed approach, it also increases model complexity, which may not always be necessary depending on study objectives. To accommodate these variations, `NeMO` allows fitting an occupancy model tailored to the selected modeling protocol with the `Nemodel()` function:

1. ‘PCR_rep_seq_read’: This model represents the most complex scenario, incorporating eDNA samples, individually indexed PCR replicates, and sequence read counts.

2. ‘seq_read’: This model simplifies the analysis by omitting PCR replicates indexing (k level) while still considering sequence read counts. It modifies the model described above by replacing W_{nijkc} with A_{nijc} in **Equations 5-6** and excluding **Equation 3**, as well as the calculation of p_{nijkc} and associated random effects.
3. ‘PCR_rep’: This model focuses on individually indexed PCR replicates without incorporating sequence read counts. Here, W_{nijkc} is replaced with Y_{nijkc} in **Equation 3**, and **Equations 4-7** are excluded, as well as the calculation of φ_{nijkc} and associated random effects.

In addition to fitting occupancy models with the `Nemodel()` function, `NeMO` offers auxiliary tools to facilitate model implementation. It enables the integration of environmental or methodological covariates at any level of the nested model structure. The `covarray()` function standardizes these covariates into structured arrays compatible with the model, ensuring flexibility across study designs. The `min_resources()` function calculates the minimum number of samples, PCR replicates, and sequencing depth needed to achieve a specified probability of detecting species when present, *i.e.*, to reliably detect species. The `WAIC()` function computes the Watanabe-Akaike Information Criterion (WAIC; Watanabe, 2010) for model comparison. These tools enhance `NeMO` by streamlining data preparation, guiding resource allocation, and aiding model selection. Detailed descriptions of these functions are available in **Appendix S1** and in the package documentation.

2.3. Application to empirical data

We modeled multi-species site-occupancy using a previously published dataset from an eDNA metabarcoding survey of fish conducted during a single campaign in 2016 along the Rhône River and its tributaries (Pont *et al.*, 2018). 80 sites were selected to ensure a representative coverage of the entire river with a standardized study design (**Fig. 2a**), involving 2 eDNA samples per site and 12 PCR replicates per sample. Details of the sampling and laboratory methods are available in the original publication. The dataset includes sequence read count data for 100 fish species at the selected sites, including false positives – species absent from the Rhône River but detected with eDNA metabarcoding, likely resulting from human consumption or sewage contamination. Model fitting was performed using

`NeMO` v1.0.0 with JAGS v4.3.2 (Plummer, 2003). To illustrate how covariates can be implemented, we included the distance to the sea as a site covariate at the occupancy level (ψ parameter), as fish diversity gradients in the Rhône basin are well documented and particularly pronounced (Tockner *et al.*, 2022). This dataset enabled us to run the `Nemodel()` function with the three different modeling protocols implemented, enabling comparison of model outputs. The raw sequence read count dataset was used for the 'PCR_rep_seq_read' protocol, while PCR replicates were pooled by sample for the 'seq_read' protocol, and sequence read counts were ignored and reduced to presence/absence data for the 'PCR_rep' protocol. Three MCMC chains were run in parallel, each with 150,000 iterations, a burn-in period of 100,000 and a thinning of 10. Eventually, we kept 5,000 iterations per chain (15,000 in total). Convergence was assessed using traceplots and the Gelman-Rubin statistic $\hat{R}$ (Brooks & Gelman, 1998). The Genotoul-Bioinfo facility (Toulouse, France) was used to run and parallelize the MCMC chains across individual cores on standard nodes (128 threads, 2 TiB of RAM) equipped with dual AMD EPYC 7713 processors. The area under the curve (AUC) of the receiver operating characteristic (ROC) curve was calculated for each modeling protocol to compare model performance (Fawcett, 2006). This analysis requires prior knowledge of whether a species is present or absent at a given site, which was obtained using an exhaustive list of Rhône fish species supplemented by long-term electrofishing data (2006-2017) described in Pont *et al.* (2018). This dataset was restricted to the 10 years preceding the eDNA sampling (2007-2016). ROC analyses were conducted by river section to account for dams (**Fig. 2a**), using only sections with available electrofishing data. Species were classified as true positives, false negatives, true negatives, or false positives based on fish species list, eDNA and electrofishing detections (see **Appendix S2** for more details). The `pROC` R package v1.18.5 (Robin *et al.*, 2011) was used to compute and plot ROC curves based on mean species occupancy probabilities from the latent Z array (**Equation 1**), providing a comprehensive assessment of model sensitivity and specificity.

Spearman's rank correlation coefficients were calculated between the parameters' link functions to examine whether the species eDNA occupancy probability (ψ) is linked to its probability of collection (θ), amplification (p), and to its expected sequence read count (φ). For species that were recorded in the electrofishing dataset, correlations were also calculated between these parameters and the mean number

of catches per species in the Rhône sections prospected over the ten-year survey period. This analysis aimed to investigate whether eDNA occupancy, collection, amplification and expected read count reflect quantitative abundance data. To determine the minimum resources required to reach a 95% detection probability (minimum number of samples J_{min}, minimum number of PCR replicates K_{min}, and minimum sequencing depth M_{min}), we used the `min_resources()` function from the `NeMO` package presented here.

Analyses were conducted using R v4.3.2 (R Core Team, 2023). Unless otherwise stated, values are reported as medians with their 95% highest density intervals (HDIs) across the 15,000 MCMC samples ran with the 'PCR_rep_seq_read' protocol.

3. RESULTS

3.1. Protocol comparison & Model outputs

Model performance was similar across the three modeling protocols, with ROC analysis yielding comparable AUC values, ranging from 0.797 for 'PCR_rep_seq_read' to 0.824 for 'seq_read' (**Fig. S1**). Estimates remained consistent across all species regardless of the protocol used. However, execution time varied substantially (**Fig. S1**). 'PCR_rep' was the fastest, completing in just 2 hours with the high-performance computing facility, whereas 'PCR_rep_seq_read', the most computationally intensive design, required nearly 10 days to run on the same system. Given the similarity in model outputs across protocols, we focused further analyses on 'PCR_rep_seq_read' results, as this approach provides information at every level of the nested design.

Posterior occupancy for each species at all sites was obtained from the model output by extracting the latent Z array (**Fig. 2b**). Similarly, posterior latent matrices for collection (A), amplification (W), and read count (S) can be derived, though their visualization is challenging, as it would require three or more dimensions. Mean occupancy was generally very high when the species was detected with eDNA (close to 1) and very low otherwise (close to 0). However, upstream-downstream occupancy gradients were evident for some species, with certain taxa showing higher occupancy probability downstream, even in sites where they were not detected (*e.g.*, *Barbus meridionalis*, *Cyprinidae – Complex 2*). These occupancy patterns align with the estimated effect of the

distance-to-sea covariate (β_ψ, **Fig. 2c**). Based on this estimate, some species can be clearly identified as restricted to upstream sites ($\beta_\psi > 0$: *e.g.*, *Coregonus lavaretus*, *Thymallus thymallus*) or downstream sites ($\beta_\psi < 0$: *e.g.*, *Alosa spp.*, *Chelon ramada*).

3.2. Correlations between parameters

Model outputs provided posterior probabilities for estimates related to eDNA occupancy (ψ), collection (θ), amplification (p), and expected read count (φ). The probability densities varied noticeably among species (**Fig. S2**). However, species with high eDNA occupancy probability also tended to exhibit higher detectability (θ, p, and φ). This trend is further supported by positive Spearman's rank correlation coefficients between all parameter pairs (**Fig. S3**). Correlations were particularly strong among detection-related parameters – eDNA collection (θ), amplification (p), and sequencing (φ) – ranging from 0.881 (HDI: 0.828 – 0.926) between $logit(\theta)$ and $log(\varphi)$ to 0.961 (HDI: 0.931 – 0.983) between $logit(\theta)$ and $logit(p)$. This indicates that species expected to yield high sequence read count also have eDNA more likely to be successfully amplified in PCR replicates and collected in samples.

Although occupancy and detection parameters were also positively correlated, these relationships were slightly weaker, with coefficients ranging from 0.573 (HDI: 0.482 – 0.655) between $logit(\psi)$ and $log(\varphi)$, to 0.64 (HDI: 0.541 – 0.72) between $logit(\psi)$ and $logit(p)$. As a result, species with high occupancy probability (*i.e.*, occurring in more sites) also exhibited higher detectability across samples, PCR replicates, and sequencing reads, and species with low occupancy probability also exhibited lower detectability. The majority of species, like *Rutilus rutilus* (high ψ, θ, p, and φ) and *Misgurnus fossilis* (low ψ, θ, p, and φ) closely followed the general pattern, while *Alosa spp.* and *Scardinius erythrophthalmus* diverged (**Fig. S3**), illustrating the distinction between rarity (low occupancy) and elusiveness (low detectability). *Alosa spp.* was rare throughout the river (low ψ) but highly detectable (high θ, p, and φ), reflecting its restriction to downstream sites (**Fig. 2b-c**) where its eDNA was consistently detected. In contrast, *Scardinius erythrophthalmus* was common (high ψ) but elusive (low θ, p, and φ). **Figure 3** further illustrates these differences, showing consistently high or

low parameter values for *Rutilus rutilus* and *Misgurnus fossilis*, while *Alosa spp.* and *Scardinius erythrophthalmus* exhibit the rare-conspicuous and common-elusive patterns, respectively.

Spearman's correlations were also positive when comparing the mean annual number of catches per species recorded during the ten-year electrofishing survey with parameters ψ, θ, p, and φ (**Table S1**). They ranged from 0.4 (HDI: 0.377 – 0.421) for the expected read count (φ) to 0.516 (HDI: 0.456 – 0.571) for occupancy probability (ψ). These results indicate a weak/moderate positive relationship between occupancy probability, detectability, and electrofishing-based abundance estimates.

3.3. Estimation of the minimum technical requirements (eDNA samples/PCR replicates/Sequencing depth) to reliably detect species

The minimum number of eDNA samples, PCR replicates, and minimum sequencing depth required to achieve a high probability of detecting species when present (95% detection probability) varied greatly among species (**Fig. 4**). Regarding eDNA sampling and PCR replication, our model shows that the effort reported in Pont *et al.* (2018) – *i.e.*, 2 eDNA samples and 12 PCR replicates – was sufficient to reliably detect most species (63%). In contrast, 18 species would have required additional sampling and PCR technical effort to reach the same detection probability (potential false negatives). Among these, five species required just one additional sample, while they would all have required more PCR replicates. For some extreme cases, such as *Leucaspius delineatus* and *Lota lota*, the additional PCR replicates required would have been prohibitively high. The minimum sequencing required to reliably detect species was calculated for each sample, as individually indexed PCR products were pooled by sample in equal volumes during library preparation, targeting a theoretical sequencing depth of 500,000 reads (Pont *et al.*, 2018). This estimated minimum sequencing depth was then compared to the observed sequencing depth, defined as the total number of reads per sample. Overall, sequencing depth was sufficient, as no samples fell below the required threshold for any species, except for *Leucaspius delineatus*, for which one sample at each of two locations did not reach the minimum required depth (**Fig. S4**). However, the second sample at each site still provided sufficient coverage to prevent detection failure.

4. DISCUSSION

We present NeMO, a Bayesian occupancy modeling framework tailored for metabarcoding-based eDNA studies with nested designs. This versatile R package accommodates diverse study designs, handling both presence/absence and sequence read count data, as well as individually indexed or pooled PCR replicates. NeMO accounts for multiple sources of uncertainty across hierarchical levels (collection, amplification, sequencing), and allows users to incorporate covariates influencing occupancy and detection probabilities. Beyond modeling, it enables formal model comparison to assess covariate effects and help optimize study design by evaluating whether sampling effort, PCR replication, and sequencing depth are sufficient to reliably detect species. By refining species occupancy estimates and guiding resource allocation, NeMO enhances the reliability of eDNA-based biodiversity assessments, particularly for elusive species.

We applied NeMO to a previously published eDNA metabarcoding dataset investigating fish diversity along the Rhône River (Pont *et al.*, 2018). Regardless of the modeling protocol used, the high AUC (~ 0.8) in ROC analysis indicated strong model performance in distinguishing species presence from absence. By accounting for imperfect detection, occupancy probabilities provided more informative insights than raw presence/absence or sequence read count data. Integrating distance to sea as a covariate further refined site-level estimates, successfully capturing well-documented biological gradients (Keith *et al.*, 2020) reflecting species distributions along the river system. Upstream-associated species had positive estimate values, including *Thymallus thymallus* and *Coregonus lavaretus*, the latter being restricted to alpine lakes. In contrast, downstream-associated species showed negative values, particularly diadromous species such as *Alosa spp.* (anadromous), *Anguilla anguilla* (catadromous), and also estuarine species such as *Chelon ramada*, reflecting their reliance on transitional or marine-influenced habitats.

Positive correlations between occupancy or detection parameter estimates with electrofishing data align with previous studies comparing eDNA with traditional metrics (Muri *et al.*, 2020; Pont *et al.*, 2018, 2023). These findings suggest that occupancy and detection probabilities could be meaningful proxies for species abundance in this river system. However, the correlations were moderate and these relationships should be interpreted cautiously when surveys differ in temporal and spatial scales. As

expected, strong positive correlations were observed among detection parameters (θ, p, and φ), probably due to high eDNA abundance increasing the likelihood of successful collection in a sample, amplification in PCR replicates, and of obtaining a high number of sequencing reads. Correlations between detection parameters and occupancy probability (ψ) were weaker, likely influenced by species-specific biological factors affecting DNA release into the environment, such as activity patterns, reproductive status, or body size (Stewart, 2019). Additionally, methodological factors like primer biases could affect DNA amplification, and sequencing efficiency despite species presence in the site surveyed (Elbrecht & Leese, 2015). Consequently, some species had low detectability and may remain undetected at sites they actually occupy. In this study, elusiveness (low detectability) is indeed particularly noticeable at the PCR level as p frequently displays low values and φ gets close to zero for many species (**Fig. S2**). While rare species are often elusive, exceptions exist: some common species remain difficult to detect (*Scardinius erythrophthalmus*), while certain rare species are conspicuous (*Alosa spp.*). This highlights the distinction between rarity and elusiveness, two concepts that should not be confused (Cerqueira *et al.*, 2013). It is worth noting that these concepts should always be contextualized, as a species may be globally rare but regionally common – like upstream or downstream species in the Rhône River – and elusiveness is method-dependent (*i.e.*, a species may be elusive because difficult to detect with one method but readily detectable with another).

`NeMO` enables retrospective evaluation of eDNA surveys by estimating the minimum number of eDNA samples, PCR replicates, and sequencing depth required to achieve a high probability of detecting species conditional on presence. While sequencing depth was consistently sufficient for the species surveyed by Pont *et al.* (2018), about one-third lacked enough samples or PCR replicates for reliable detection. This highlights the inevitable trade-off between resource allocation and species detectability. Although eDNA proved more effective than traditional methods for detecting elusive species – such as *Zingel asper*, *Mugil cephalus* and *Misgurnus fossilis*, which are elusive with eDNA metabarcoding but even more with electrofishing as they were never detected during the 10 years of survey – detectability must still be assessed before asserting its superiority. Our results suggest that increased replication effort could have allowed for the detection of these species at more sites. However, for some low-detectability species like *Leucaspius delineatus*, an ‘emerging’ species in the Rhône basin,

achieving a high probability of detection when the species is present would require an impractical number of PCR replicates per sample (> 1,000), highlighting the limitations of the selected eDNA study design for certain targets. To improve detection, adjustments to study design could include alternative primers or modified sampling protocols. For example, *Lota lota*, a deep-dwelling nocturnal species (Keith *et al.*, 2020), showed low detectability, likely because sampling was conducted in lentic systems at the surface and during daytime. Therefore, it is essential to cautiously interpret non-detections of such species, as it may not reflect true absence. We recommend integrating occupancy models into eDNA studies to validate non-detections and guide replication effort. This approach provides statistical support for filtering low-detection amplicons, as demonstrated by Pont *et al.* (2018), who excluded *Leucaspius delineatus* and *Lota lota* due to single-PCR detections – a decision now supported by `NeMO`. We further show that other species, such as *Salvelinus namaycush* or *Micropterus salmoides*, could have been filtered out as well. When eDNA sampling designs are insufficient for reliable species detection, practitioners face the challenge of deciding whether to increase the number of eDNA samples, PCR replicates, or sequencing depth. Ideally, minimum technical requirements should be met for all components. However, when budgetary or logistical constraints necessitate prioritization, we recommend the following approach: (i) identify which component(s) are most limiting for reliable detection based on study-specific objectives (*e.g.*, targeted species, community completeness…); (ii) prioritize improvement of the component with the largest shortfall; and (iii) consider practical trade-offs. In most cases, increasing PCR replication is more cost-effective than adding field samples, as field sampling involves the entire downstream processing chain. However, PCR replication is constrained by the amount of extracted DNA available, and additional extractions can only be performed if biological material remains. Regarding sequencing depth, it is not freely adjustable as it is determined by the number of samples multiplexed on a sequencing run. Therefore, increasing either eDNA samples or PCR replicates automatically reduces read depth per sample unless additional sequencing runs are performed, which substantially increases costs. These considerations highlight the importance of using pilot studies to optimize sampling design in a statistically-driven way before downstream analyses. Given the risks of misinterpreting species absences, we recommend comparing community patterns with

and without uncertain absences to assess their impact on α- and β-diversity, refining conclusions as necessary. Such practices can improve study efficiency and reliability of biodiversity assessments.

Among the three modeling protocols implemented in NeMO, the 'PCR_rep' protocol proved to be the fastest. When sequencing-related covariates are not included, models based on sequence read counts – such as the 'seq_read' and 'PCR_rep_seq_read' protocols in NeMO – could be avoided as they substantially increase computational time without improving performance. However, estimating the minimum sequencing depth required for reliable detection remains essential. For comparability, all models were run using the same number of MCMC iterations, which was chosen to ensure convergence of the most complex protocol, 'PCR_rep_seq_read'. In practice, this indicates that a smaller number of iterations would likely have been sufficient for the two other protocols, further reducing their computational time relative to 'PCR_rep_seq_read'. These analyses were conducted on a large empirical dataset, which contributes to the overall computation time. To facilitate testing and exploration of the workflow, the NeMO package includes a smaller example dataset derived from this dataset, allowing users to run the models and reproduce the main steps of the analysis within a few minutes. A dedicated vignette is also provided to guide users through the complete workflow using this example dataset.

While our framework accounts for false negatives and detection biases, it does not model false positives, which remain a common issue in eDNA studies. These are often mitigated through stringent bioinformatic pipelines or taxonomic filtering to exclude sequencing artifacts or unexpected detections arising from contamination or diet, which could also lead to false negatives if these filters are too stringent. Occupancy models that account for false positives do exist (Guillera-Arroita *et al.*, 2017; Lahoz-Monfort *et al.*, 2016), but they generally require additional data to distinguish true from false detections, such as an unambiguous detection method or calibration procedures applied at each stage of the multi-scale sampling hierarchy. These requirements add complexity to already intricate hierarchical models and are often impractical in large-scale metabarcoding studies. Additionally, in lotic systems, species detection may be spatially displaced due to downstream DNA transport, though this effect can be quantified (Perry *et al.*, 2024; Pont *et al.*, 2018). Future developments could thus integrate hydrological transport effects for both freshwater (Gimenez, 2024) and marine systems. More generally,

incorporating spatial autocorrelation (Chen & Ficetola, 2019) or species co-occurrence (Rota *et al.*, 2016) could improve inference, but at the cost of greater model complexity and computational demand.

DATA AVAILABILITY STATEMENT

No new data were generated for this study. We used a previously published dataset available on Dryad (DOI: 10.5061/dryad.t4n42rr). The R package developed as part of this research, `NeMO`, is available on GitHub (https://github.com/bastien-mace/NeMO). The package includes a vignette and an example dataset to guide users in its implementation.

ACKNOWLEDGEMENTS

We are grateful to the genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, https://doi.org/10.15454/1.5572369328961167E12) for providing computing and storage resources. We are grateful to Alexandre Granier for his insightful guidance on the structuring of our R package.

AUTHOR CONTRIBUTION

SM and EDT conceived the study. BM analyzed the data and drafted the first manuscript. MR and NR provided essential data for the study. All authors contributed to manuscript revisions and approved the final version.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interests.

TABLES

Table 1. Scientific papers reporting code or software for multi-species occupancy modeling applied to eDNA metabarcoding, among the 75 papers identified through a literature search using the following keywords: [("metabarcoding" OR ("eDNA" AND ("multispecies" OR "multi-species"))) AND "occupancy"] (Web of Science Core Collection on March 13, 2025). Full references are provided in **Appendix S3**.

Publication	Input data	PCR replication level	Covariates implementation	Model comparison	Resource optimisation	Additional features	Package
Doi *et al.* 2019	Presence/Absence	✓	✓	-	-	-	-
Bush *et al.* 2020	Presence/Absence	-	✓	✓	-	-	-
McClenaghan *et al.* 2020	Presence/Absence	✓	✓	✓	-	Species richness (Data augmentation)	-
Gold *et al.* 2021	Presence/Absence	-	-	-	-	-	-
McColl-Gausden *et al.* 2021	Presence/Absence	-	✓	-	-	-	-
Pukk *et al.* 2021	Presence/Absence	-	-	-	-	-	-
Fukaya *et al.* 2022	Sequence read count	-	✓	-	✓	-	occumb
Ji *et al.* 2022	Presence/Absence	-	✓	-	-	Species richness (Data augmentation)	-
Allen *et al.* 2023	Presence/Absence	-	✓	-	✓	-	-
Peixoto *et al.* 2023	Presence/Absence	-	✓	-	✓	-	-
Diana *et al.* 2024	Sequence read count	✓	✓	-	-	Spatial autocorrelation Species correlations	eDNAPlus
Rounds *et al.* 2024	Presence/Absence	-	✓	-	✓	-	-
This study	Presence/Absence OR Sequence read count	✓ individually indexed OR pooled	✓	✓	✓	-	NeMO

FIGURES

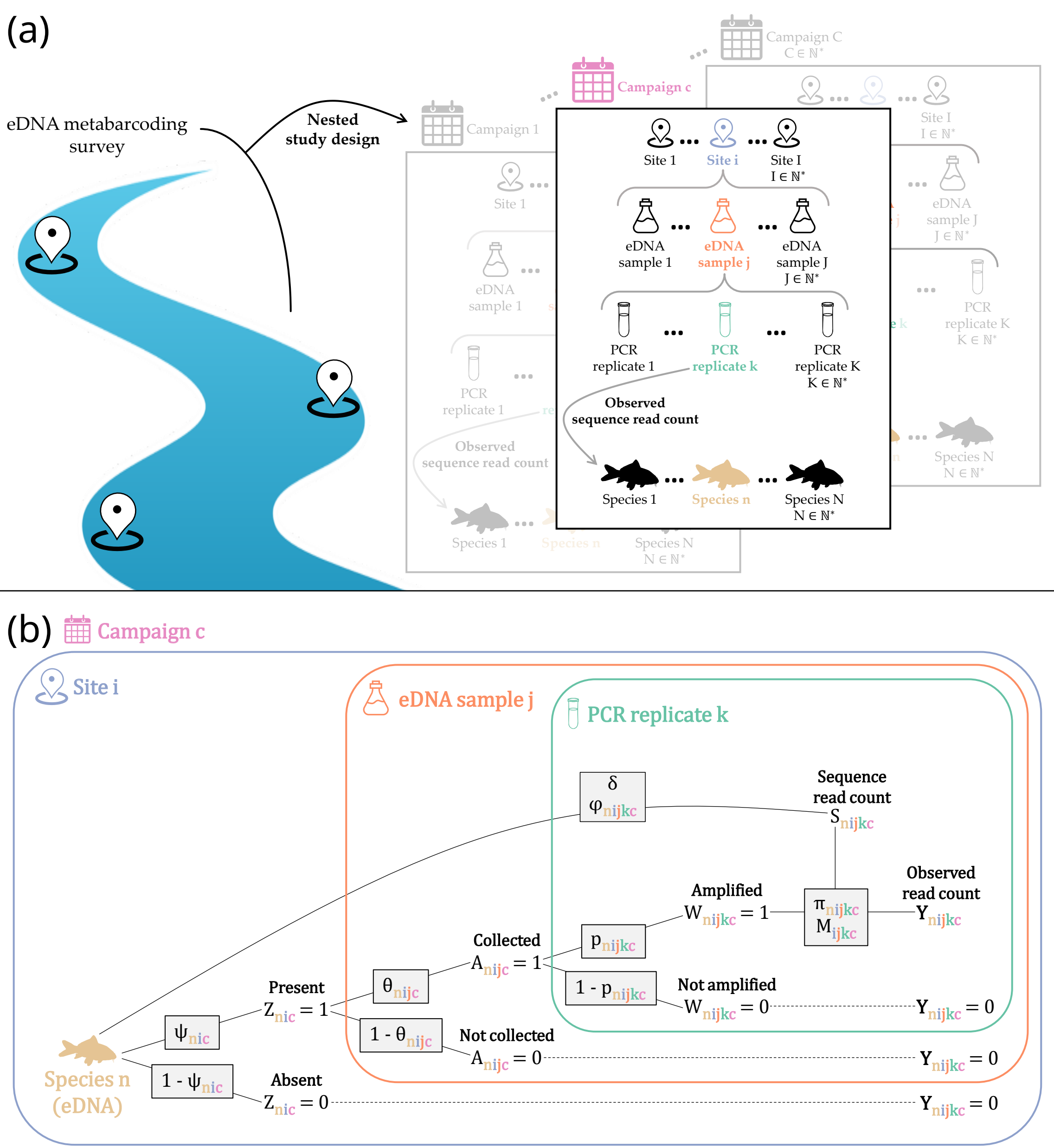
Figure 1. (a) Illustration of a nested study design for an eDNA metabarcoding survey in a river, monitoring N fish species over C campaigns, with replication across sites (I), samples (J), and PCR replicates (K). **(b)** Tree diagram illustrating the nested structure of the eDNA metabarcoding occupancy model. During a given campaign c, eDNA from species n may either be present at site i with occupancy probability ψ_{nic} (latent variable $Z_{nic} = 1$), or absent with probability $1 - \psi_{nic}$ ($Z_{nic} = 0$). If present, eDNA can be collected in an eDNA sample j with detection probability θ_{nijc} (latent variable $A_{nijc} = 1$), or remain uncollected with probability $1 - \theta_{nijc}$ ($A_{nijc} = 0$). If collected, eDNA can be amplified in a given PCR replicate k with detection probability p_{nijkc} (latent variable $W_{nijkc} = 1$), or remain unamplified with probability $1 - p_{nijkc}$ ($W_{nijkc} = 0$). If amplified, eDNA is sequenced with frequency π_{nijkc} relative to the total sequence read count M_{ijkc}, generating the observed sequence read count Y_{nijkc}. When eDNA is absent, not collected, or not amplified, the read count is $Y_{nijkc} = 0$. The observed read count depends on the sequence read count (latent variable S_{nijkc}) which follows a negative binomial distribution with mean φ_{nijkc} and dispersion parameter δ.

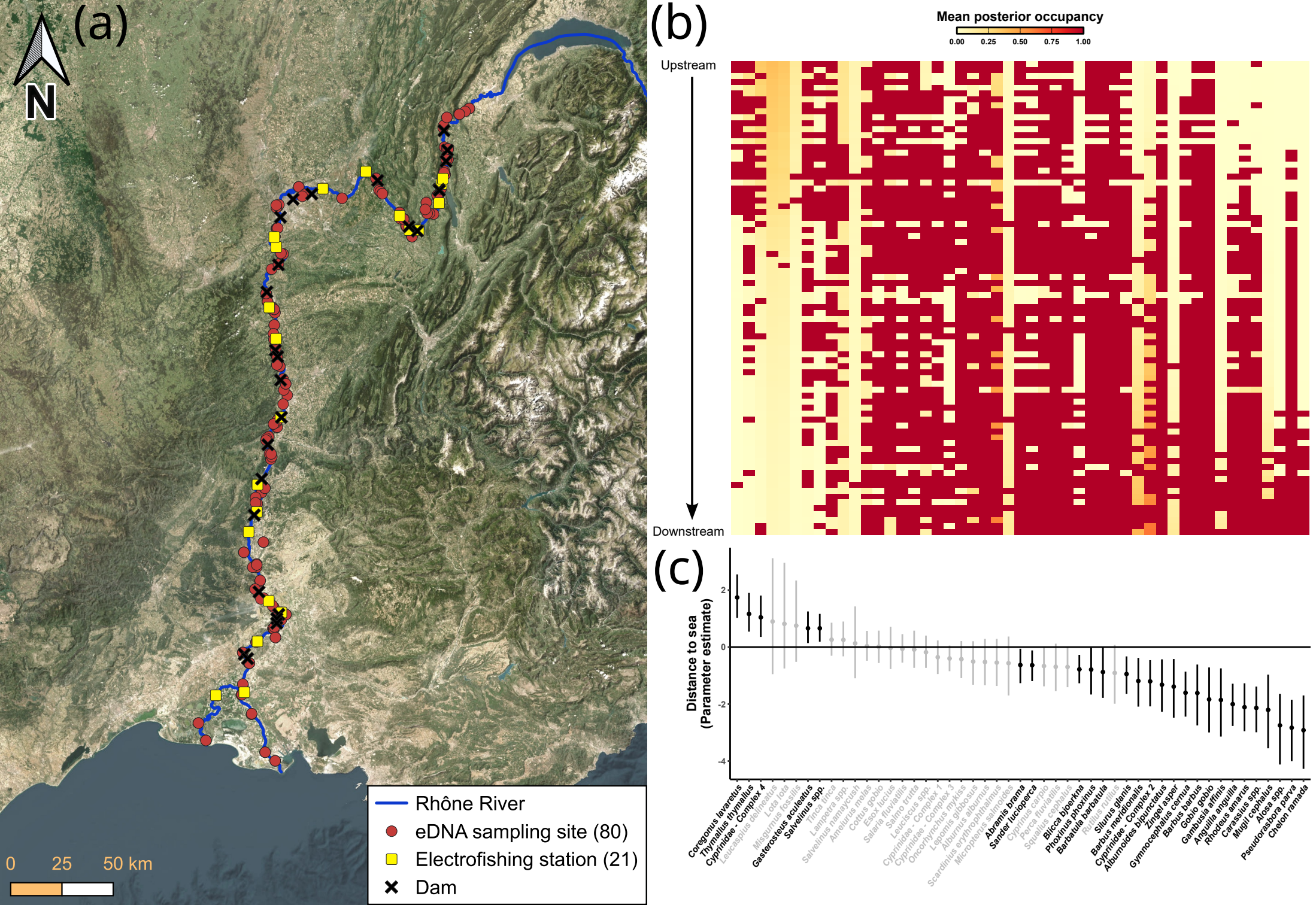
Figure 2. (a) Map of sampling locations along the Rhône River (France). The river thalweg is shown in blue. Red dots denote the 80 sites (including some located at the confluence of the Rhône and its tributaries) where two eDNA samples were collected per site. Yellow squares indicate stations where long-term electrofishing surveys were conducted, and black crosses mark the locations of dams. **(b)** Mean posterior occupancy (Z array) for species inhabiting the Rhône River at the 80 sampling sites. **(c)** Distance-to-sea occupancy parameter estimate (β_ψ) for each species inhabiting the Rhône River. Dots indicate the median estimates and their 95% HDI. Species with estimates whose HDI does not overlap zero are shown in black, others are shown in gray. Species identified at the genus level (*spp.*) or as part of complexes are detailed in **Table S2**.

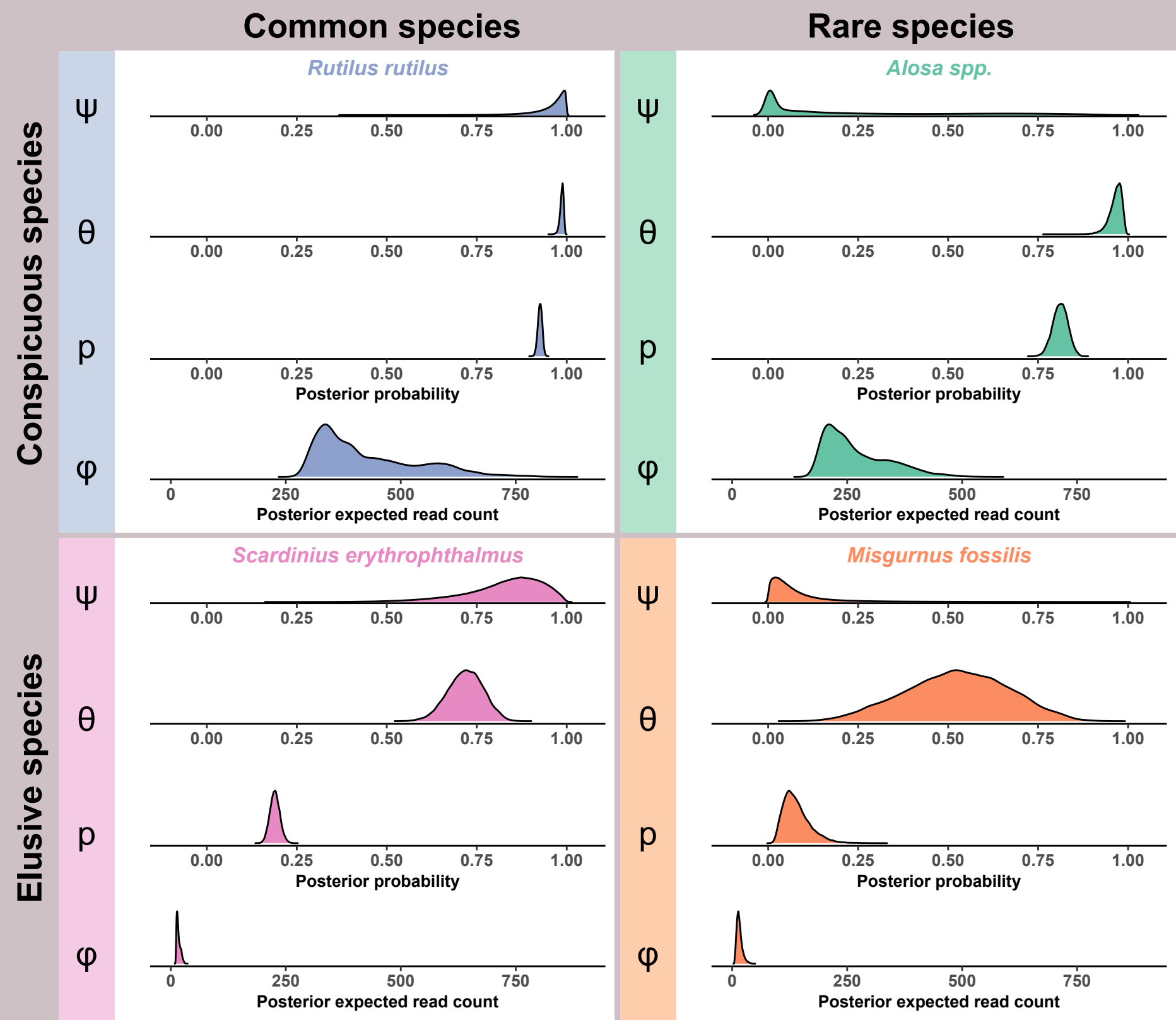
Figure 3. Ridge line plots depicting the densities of posterior probabilities for eDNA occupancy (ψ), collection (θ), amplification (p), and expected read count (φ) for four species inhabiting the Rhône

River. The highlighted species are *Rutilus rutilus* (blue), *Alosa spp.* (green), *Scardinius erythrophthalmus* (pink) and *Misgurnus fossilis* (orange). These species are categorized based on (1) their rarity – common species (*Rutilus rutilus* and *Scardinius erythrophthalmus*) have high occupancy probability; rare species (*Alosa spp.* and *Misgurnus fossilis*) have low occupancy probability – and (2) their detectability – conspicuous species (*Rutilus rutilus* and *Alosa spp.*) exhibit high eDNA collection and amplification probabilities, and have high expected read counts; elusive species (*Scardinius erythrophthalmus* and *Misgurnus fossilis*) are characterized by lower eDNA collection and amplification probabilities, as well as lower expected read count.

Figure 4. Tile plot illustrating the minimum number of eDNA samples (J_{min}; in blue), PCR replicates (K_{min}; in green), and sequencing depth (M_{min}; in red) required to reliably detect species with 95% confidence. Median values are shown, with their 95% HDI in brackets. For sequencing, values are calculated across all samples and sites, providing summary estimates. The sequencing depth used in the study was always sufficient to detect species at all sites (see **Fig. S4**). For samples and replicates, gray shading indicates cases where the resources used in the original study (2 eDNA samples and 12 PCR replicates) were sufficient across the entire HDI. Species names are underlined if they are non-indigenous in the Rhône River, and threatened species are color-coded according to their IUCN conservation status at the most local scale possible (yellow: Vulnerable; orange: Endangered; red: Critically Endangered). Species identified at the genus level (*spp.*) or as part of complexes are detailed in **Table S2**.


(a)
eDNA metabarcoding survey
Nested study design
Campaign 1
Campaign c
Campaign C
C ∈ ℕ*
Site 1
Site i
Site I
I ∈ ℕ*
eDNA sample 1
eDNA sample j
eDNA sample J
J ∈ ℕ*
PCR replicate 1
PCR replicate k
PCR replicate K
K ∈ ℕ*
Observed sequence read count
Species 1
Species n
Species N
N ∈ ℕ*
(b)
Campaign c
Site i
eDNA sample j
PCR replicate k
Species n (eDNA)
ψnic
1 - ψnic
Present
Znic = 1
Absent
Znic = 0
θnijc
1 - θnijc
Collected
Anijc = 1
Not collected
Anijc = 0
δ
φnijkc
pnijkc
1 - pnijkc
Amplified
Wnijkc = 1
Not amplified
Wnijkc = 0
πnijkc
Mijkc
Sequence read count
Snijkc
Observed read count
Ynijkc
Ynijkc = 0


(a)
N
Rhône River
eDNA sampling site (80)
Electrofishing station (21)
Dam
0
25
50 km
(b)
Mean posterior occupancy
0.00
0.25
0.50
0.75
1.00
Upstream
Downstream
(c)
Distance to sea
(Parameter estimate)
2
0
-2
-4
Coregonus lavaretus
Thymallus thymallus
Cyprinidae - Complex 4
Leucaspius delineatus
Lota lota
Misgurnus fossilis
Gasterosteus aculeatus
Salvelinus spp.
Tinca tinca
Lampetra spp.
Salvelinus namaycush
Ameiurus melas
Cottus gobio
Esox lucius
Salaria fluviatilis
Salmo trutta
Leuciscus spp.
Cyprinidae - Complex 1
Cyprinidae - Complex 3
Oncorhynchus mykiss
Lepomis gibbosus
Alburnus alburnus
Scardinius erythrophthalmus
Microptenus salmoides
Abramis brama
Sander lucioperca
Cyprinus carpio
Perca fluviatilis
Squalius cephalus
Blicca bjoerkna
Phoxinus phoxinus
Barbatula barbatula
Rutilus rutilus
Silurus glanis
Barbus meridionalis
Cyprinidae - Complex 2
Alburnoides bipunctatus
Zingel asper
Gymnocephalus cernua
Barbus barbus
Gobio gobio
Gambusia affinis
Anguilla anguilla
Rhodeus amarus
Carassius spp.
Mugil cephalus
Alosa spp.
Pseudorasbora parva
Chelon ramada


Common species
Rare species
Conspicuous species
Elusive species
Rutilus rutilus
Alosa spp.
Scardinius erythrophthalmus
Misgurnus fossilis
ψ
θ
p
φ
0.00
0.25
0.50
0.75
1.00
Posterior probability
0
250
500
750
Posterior expected read count

Minimum technical requirements

eDNA samples (J_{min}): ≤ 2; 2, 4, 6, 8, 10

PCR replicates (K_{min}): ≤ 12; 40, 80, 120, 160

Sequencing depth (M_{min}): 1000, 2000, 3000, 4000, 5000

	eDNA samples (J_{min})	PCR replicates (K_{min})	Sequencing depth (M_{min})
Abramis brama	1 [1; 2]	3 [2; 3]	36 [7; 13564]
Alburnoides bipunctatus	1 [1; 1]	2 [2; 2]	75 [13; 14804]
Alburnus alburnus	1 [1; 2]	2 [2; 2]	62 [11; 14432]
Alosa spp.	1 [1; 2]	2 [2; 3]	56 [8; 13300]
Barbatula barbatula	1 [1; 1]	2 [1; 2]	73 [13; 14771]
Barbus barbus	1 [1; 1]	1 [1; 1]	32 [7; 13594]
Barbus meridionalis	3 [2; 5]	31 [20; 51]	4207 [517; 15160]
Blicca bjoerkna	2 [1; 2]	4 [3; 4]	55 [10; 13920]
Chelon ramada	2 [1; 2]	4 [3; 4]	117 [10; 14190]
Cottus gobio	1 [1; 2]	2 [2; 3]	118 [20; 15238]
Cyprinidae - Complex 1	1 [1; 1]	2 [2; 2]	24 [5; 12276]
Cyprinidae - Complex 3	1 [1; 2]	3 [2; 3]	59 [11; 14286]
Cyprinidae - Complex 4	3 [2; 5]	22 [16; 31]	723 [99; 13929]
Gasterosteus aculeatus	2 [2; 3]	7 [6; 8]	1075 [151; 15751]
Gobio gobio	1 [1; 1]	2 [1; 2]	66 [12; 15062]
Gymnocephalus cernua	1 [1; 1]	2 [2; 2]	83 [14; 14891]
Lampetra spp.	3 [2; 5]	14 [11; 18]	909 [126; 15026]
Leucaspius delineatus	10 [5; 54]	162 [61; 1110]	5131 [1001; 12794]
Leuciscus spp.	2 [2; 2]	6 [5; 6]	337 [51; 15367]
Mugil cephalus	3 [2; 5]	31 [20; 54]	677 [32; 13812]
Perca fluviatilis	2 [1; 2]	3 [3; 4]	296 [46; 15801]
Phoxinus phoxinus	1 [1; 1]	2 [2; 2]	121 [20; 15316]
Rhodeus amarus	2 [1; 2]	4 [3; 4]	284 [44; 15801]
Rutilus rutilus	1 [1; 1]	2 [2; 2]	32 [6; 13519]
Salaria fluviatilis	1 [1; 2]	3 [3; 3]	212 [34; 15382]
Salmo trutta	1 [1; 2]	3 [3; 3]	126 [20; 15190]
Salvelinus spp.	2 [2; 3]	6 [5; 6]	254 [39; 14426]
Scardinius erythrophthalmus	3 [2; 4]	15 [13; 17]	860 [120; 15583]
Squalius cephalus	2 [1; 2]	3 [3; 3]	42 [8; 13824]
Tinca tinca	2 [2; 2]	8 [7; 8]	689 [100; 15873]
Ameiurus melas	2 [2; 3]	9 [8; 10]	316 [47; 15364]
Carassius spp.	2 [2; 2]	5 [5; 6]	1529 [208; 16098]
Cyprinidae - Complex 2	3 [2; 4]	30 [23; 41]	1091 [140; 14816]
Cyprinus carpio	2 [1; 2]	3 [3; 3]	331 [51; 15846]
Gambusia affinis	3 [2; 4]	18 [12; 30]	1073 [66; 14867]
Lepomis gibbosus	2 [2; 2]	6 [5; 6]	304 [47; 15538]
Micropterus salmoides	4 [3; 8]	45 [24; 104]	2565 [237; 13803]
Oncorhynchus mykiss	2 [1; 2]	2 [2; 3]	52 [9; 14214]
Pseudorasbora parva	2 [1; 2]	3 [3; 3]	384 [58; 16233]
Salvelinus namaycush	5 [3; 11]	45 [20; 192]	3923 [396; 13171]
Sander lucioperca	2 [1; 2]	3 [3; 3]	179 [28; 15451]
Silurus glanis	1 [1; 2]	3 [2; 3]	128 [21; 15238]
Coregonus lavaretus	2 [2; 3]	6 [5; 7]	217 [33; 14142]
Esox lucius	2 [2; 2]	4 [4; 5]	94 [16; 14612]
Lota lota	7 [4; 23]	96 [39; 406]	4717 [563; 12932]
Thymallus thymallus	2 [2; 3]	7 [6; 8]	257 [39; 14831]
Misgurnus fossilis	4 [2; 11]	40 [18; 136]	3297 [287; 13012]
Zingel asper	3 [2; 5]	20 [13; 34]	1622 [148; 14528]
Anguilla anguilla	2 [2; 2]	4 [4; 5]	450 [67; 16323]

REFERENCES

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. Methods in Ecology and Evolution, 9(1), 134–147. https://doi.org/10.1111/2041-210X.12849

Barnes, M. A., Turner, C. R., Jerde, C. L., Renshaw, M. A., Chadderton, W. L., & Lodge, D. M. (2014). Environmental Conditions Influence eDNA Persistence in Aquatic Systems. Environmental Science & Technology, 48(3), 1819–1827. https://doi.org/10.1021/es404734p

Beng, K. C., & Corlett, R. T. (2020). Applications of environmental DNA (eDNA) in ecology and conservation: Opportunities, challenges and prospects. Biodiversity and Conservation, 29(7), 2089–2121. https://doi.org/10.1007/s10531-020-01980-0

Biggs, J., Ewald, N., Valentini, A., Gaboriaud, C., Dejean, T., Griffiths, R. A., Foster, J., Wilkinson, J. W., Arnell, A., Brotherton, P., Williams, P., & Dunn, F. (2015). Using eDNA to develop a national citizen science-based monitoring programme for the great crested newt (Triturus cristatus). Biological Conservation, 183, 19–28. https://doi.org/10.1016/j.biocon.2014.11.029

Blackman, R., Couton, M., Keck, F., Kirschner, D., Carraro, L., Cereghetti, E., Perrelet, K., Bossart, R., Brantschen, J., Zhang, Y., & Altermatt, F. (2024). Environmental DNA: The next chapter. Molecular Ecology, 33(11), e17355. https://doi.org/10.1111/mec.17355

Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., Yu, D. W., & Bruyn, M. de. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. Trends in Ecology & Evolution, 29(6), 358–367. https://doi.org/10.1016/j.tree.2014.04.003

Brooks, S. P., & Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. Journal of Computational and Graphical Statistics, 7(4), 434–455. https://doi.org/10.1080/10618600.1998.10474787

Cerqueira, M. C., Cohn-Haft, M., Vargas, C. F., Nader, C. E., Andretti, C. B., Costa, T. V. V., Sberze, M., Hines, J. E., & Ferraz, G. (2013). Rare or elusive? A test of expert knowledge about rarity of Amazon forest birds. Diversity and Distributions, 19(7), 710–721. https://doi.org/10.1111/ddi.12033

Chen, W., & Ficetola, G. F. (2019). Conditionally autoregressive models improve occupancy analyses of autocorrelated data: An example with environmental DNA. Molecular Ecology Resources, 19(1), 163–175. https://doi.org/10.1111/1755-0998.12949

Cristescu, M. E., & Hebert, P. D. N. (2018). Uses and Misuses of Environmental DNA in Biodiversity Science and Conservation. Annual Review of Ecology, Evolution, and Systematics, 49, 209–230. https://doi.org/10.1146/annurev-ecolsys-110617-062306

Darling, J. A., & Mahon, A. R. (2011). From molecules to management: Adopting DNA-based methods for monitoring biological invasions in aquatic environments. Environmental Research, 111(7), 978–988. https://doi.org/10.1016/j.envres.2011.02.001

Diana, A., Matechou, E., Griffin, J., Yu, D. W., Luo, M., Tosa, M., Bush, A., & Griffiths, R. A. (2024). eDNAPlus: A Unifying Modeling Framework for DNA-based Biodiversity Monitoring. Journal of the American Statistical Association, 120(549), 120–134. https://doi.org/10.1080/01621459.2024.2412362

Dorazio, R. M., & Erickson, R. A. (2018). ednaoccupancy: An r package for multiscale occupancy modelling of environmental DNA data. Molecular Ecology Resources, 18(2), 368–380. https://doi.org/10.1111/1755-0998.12735

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(470), 389–398. https://doi.org/10.1198/016214505000000015

Elbrecht, V., & Leese, F. (2015). Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass—Sequence Relationships with an Innovative Metabarcoding Protocol. PLOS ONE, 10(7), e0130324. https://doi.org/10.1371/journal.pone.0130324

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., Gielly, L., Lopes, C. M., Boyer, F., Pompanon, F., Rayé, G., & Taberlet, P. (2015). Replication levels, false

presences and the estimation of the presence/absence from eDNA metabarcoding data. Molecular Ecology Resources, 15(3), 543–556. https://doi.org/10.1111/1755-0998.12338

Ficetola, G. F., Taberlet, P., & Coissac, E. (2016). How to limit false positives in environmental DNA and metabarcoding? Molecular Ecology Resources, 16(3), 604–607. https://doi.org/10.1111/1755-0998.12508

Fukaya, K., Kondo, N. I., Matsuzaki, S. S., & Kadoya, T. (2022). Multispecies site occupancy modelling and study design for spatially replicated environmental DNA metabarcoding. Methods in Ecology and Evolution, 13(1), 183–193. https://doi.org/10.1111/2041-210X.13732

Gimenez, O. (2024). Spatial Occupancy Models for Data Collected on Stream Networks. Aquatic Conservation: Marine and Freshwater Ecosystems, 34(11), e70013. https://doi.org/10.1002/aqc.70013

Guillera-Arroita, G., Lahoz-Monfort, J. J., van Rooyen, A. R., Weeks, A. R., & Tingley, R. (2017). Dealing with false-positive and false-negative errors about species occurrence at multiple levels. Methods in Ecology and Evolution, 8(9), 1081–1091. https://doi.org/10.1111/2041-210X.12743

Keith, P., Poulet, N., Denys, G., Changeux, T., Feunteun, E., & Persat, H. (2020). Les poissons d'eau douce de France. Biotope ; MNHN.

Lahoz-Monfort, J. J., Guillera-Arroita, G., & Tingley, R. (2016). Statistical approaches to account for false-positive errors in environmental DNA samples. Molecular Ecology Resources, 16(3), 673–685. https://doi.org/10.1111/1755-0998.12486

Macé, B., Mouillot, D., Dalongeville, A., Bruno, M., Deter, J., Varenne, A., Gudefin, A., Boissery, P., & Manel, S. (2024). The Tree of Life eDNA metabarcoding reveals a similar taxonomic richness but dissimilar evolutionary lineages between seaports and marine reserves. Molecular Ecology, 33(12), e17373. https://doi.org/10.1111/mec.17373

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J., & Langtimm, C. A. (2002). Estimating Site Occupancy Rates When Detection Probabilities Are Less Than One. Ecology, 83(8), 2248–2255. https://doi.org/10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L., & Hines, J. E. (2006). Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence. Elsevier/Academic Press.

Mathon, L., Valentini, A., Guérin, P.-E., Normandeau, E., Noel, C., Lionnet, C., Boulanger, E., Thuiller, W., Bernatchez, L., Mouillot, D., Dejean, T., & Manel, S. (2021). Benchmarking bioinformatic tools for fast and accurate eDNA metabarcoding species identification. Molecular Ecology Resources, 21(7), 2565–2579. https://doi.org/10.1111/1755-0998.13430

McArdle, B. H. (1990). When Are Rare Species Not There? Oikos, 57(2), 276–277. https://doi.org/10.2307/3565950

Muri, C. D., Handley, L. L., Bean, C. W., Li, J., Peirson, G., Sellers, G. S., Walsh, K., Watson, H. V., Winfield, I. J., & Hänfling, B. (2020). Read counts from environmental DNA (eDNA) metabarcoding reflect fish abundance and biomass in drained ponds. Metabarcoding and Metagenomics, 4, e56959. https://doi.org/10.3897/mbmg.4.56959

Perry, W. B., Seymour, M., Orsini, L., Jâms, I. B., Milner, N., Edwards, F., Harvey, R., de Bruyn, M., Bista, I., Walsh, K., Emmett, B., Blackman, R., Altermatt, F., Lawson Handley, L., Mächler, E., Deiner, K., Bik, H. M., Carvalho, G., Colbourne, J., … Creer, S. (2024). An integrated spatio-temporal view of riverine biodiversity using environmental DNA metabarcoding. Nature Communications, 15(1), 4372. https://doi.org/10.1038/s41467-024-48640-3

Pinfield, R., Dillane, E., Runge, A. K. W., Evans, A., Mirimin, L., Niemann, J., Reed, T. E., Reid, D. G., Rogan, E., Samarra, F. I. P., Sigsgaard, E. E., & Foote, A. D. (2019). False-negative detections from environmental DNA collected in the presence of large numbers of killer whales (Orcinus orca). Environmental DNA, 1(4), 316–328. https://doi.org/10.1002/edn3.32

Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling. 3rd International Workshop on Distributed Statistical Computing (DSC 2003); Vienna, Austria, 124, 1–10.

Polanco F., A., Rozanski, R., Marques, V., Helmkampf, M., Mouillot, D., Manel, S., Albouy, C., Puebla, O., & Pellissier, L. (2025). A Confidence Scoring Procedure for eDNA

Metabarcoding Records and Its Application to a Global Marine Fish Dataset. Environmental DNA, 7(2), e70077. https://doi.org/10.1002/edn3.70077

Pont, D., Meulenbroek, P., Bammer, V., Dejean, T., Erős, T., Jean, P., Lenhardt, M., Nagel, C., Pekarik, L., Schabuss, M., Stoeckle, B. C., Stoica, E., Zornig, H., Weigand, A., & Valentini, A. (2023). Quantitative monitoring of diverse fish communities on a large scale combining eDNA metabarcoding and qPCR. Molecular Ecology Resources, 23(2), 396–409. https://doi.org/10.1111/1755-0998.13715

Pont, D., Rocle, M., Valentini, A., Civade, R., Jean, P., Maire, A., Roset, N., Schabuss, M., Zornig, H., & Dejean, T. (2018). Environmental DNA reveals quantitative patterns of fish biodiversity in large rivers despite its downstream transportation. Scientific Reports, 8(1), Article 1. https://doi.org/10.1038/s41598-018-28424-8

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12(1), 77. https://doi.org/10.1186/1471-2105-12-77

Rounds, C. I., Arnold, T. W., Chun, C. L., Dumke, J., Totsch, A., Keppers, A., Edblad, K., García, S. M., Larson, E. R., Nelson, J. K. R., & Hansen, G. J. A. (2024). Aquatic invasive species exhibit contrasting seasonal detectability patterns based on environmental DNA: Implications for monitoring. Freshwater Biology, 69(10), 1479–1493. https://doi.org/10.1111/fwb.14320

Royle, J. A., & Dorazio, R. M. (2008). Hierarchical modeling and inference in ecology: The analysis of data from populations, metapopulations and communities. Elsevier/Academic Press.

Schmidt, B. R., Kéry, M., Ursenbacher, S., Hyman, O. J., & Collins, J. P. (2013). Site occupancy models in the analysis of environmental DNA presence/absence surveys: A case study of an emerging amphibian pathogen. Methods in Ecology and Evolution, 4(7), 646–653. https://doi.org/10.1111/2041-210X.12052

Stat, M., Huggett, M. J., Bernasconi, R., DiBattista, J. D., Berry, T. E., Newman, S. J., Harvey, E. S., & Bunce, M. (2017). Ecosystem biomonitoring with eDNA: Metabarcoding across the tree of

life in a tropical marine environment. Scientific Reports, 7(1), Article 1. https://doi.org/10.1038/s41598-017-12501-5

Stewart, K. A. (2019). Understanding the effects of biotic and abiotic factors on sources of aquatic environmental DNA. Biodiversity and Conservation, 28(5), 983–1001. https://doi.org/10.1007/s10531-019-01709-8

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). Environmental DNA: For Biodiversity Research and Monitoring. Oxford University Press.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular Ecology, 21(8), 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Takahashi, M., Saccò, M., Kestel, J. H., Nester, G., Campbell, M. A., van der Heyde, M., Heydenrych, M. J., Juszkiewicz, D. J., Nevill, P., Dawkins, K. L., Bessey, C., Fernandes, K., Miller, H., Power, M., Mousavi-Derazmahalleh, M., Newton, J. P., White, N. E., Richards, Z. T., & Allentoft, M. E. (2023). Aquatic environmental DNA: A review of the macro-organismal biomonitoring revolution. Science of The Total Environment, 873, 162322. https://doi.org/10.1016/j.scitotenv.2023.162322

Tockner, K., Zarfl, C., Robinson, C., Aamodt, K., Skuja, A., Agrita, B., Akbulut, N., Akbulut, A., Alexevnina, M., Andersen, H., Arnaut, N., Ayrault, S., Baklanov, M., Barthélémy, C., Bäthe, J., Baumgartner, C., Bayari, S., Behrendt, H., Beisel, J.-N., & Zogaris, S. (2022). Rivers of Europe (Second Edition). Elsevier

Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. Journal of Machine Learning Research, 11(116), 3571–3594.