Received: 26 December 2024 | Accepted: 1 May 2025

DOI: 10.1111/2041-210X.70064

Methods in Ecology and Evolution

REVIEW

From anarchy to clarity, data pre-processing and statistical choices influence quantitative environmental DNA (eDNA) analyses

Jonas Bylemans[1] | Teun Everts[2,3] | Rein Brys[2] | Richard P. Duncan[4]

[1]University of Savoie Mont Blanc, INRAE, CARRTEL, Thonon-les-Bains, France

[2]Genetic Diversity, Research Institute for Nature and Forest (INBO), Geraardsbergen, Belgium

[3]Plant Conservation and Population Biology, KU Leuven, Leuven, Belgium

[4]Centre for Conservation Ecology and Genomics, Institute for Applied Ecology, University of Canberra, Bruce, Australian Capital Territory, Australia

Correspondence
Jonas Bylemans
Email: jonas.bylemans@inrae.fr

Funding information
AnaEE France and the Pôle R&D Écla; Research Foundation Flanders, Grant/Award Number: S01822N

Handling Editor: Chloe Robinson

Abstract

1. Environmental DNA (eDNA) analyses hold great potential for increasing species detection sensitivities and estimating species abundances. The rapidly growing user base, continuous method development and optimisation have led to diverse approaches for capturing and analysing eDNA. While significant efforts have been made to standardise field and laboratory protocols, a notable gap remains in understanding the consequences of data pre-processing and statistical choices on the final results obtained, particularly in quantitative eDNA analyses. These insights are crucial for developing best-practice guidelines that can harmonise analytical workflows.
2. To address this gap, we conducted an extensive literature review focusing on quantitative species-specific eDNA studies. We assessed the diversity of data pre-processing and statistical choices made to evaluate the correlation between eDNA concentrations and species' abundance or biomass, and collected the raw datasets when available. We then applied commonly used data analysis strategies to the datasets to formulate general recommendations for improving the reliability and reproducibility of quantitative eDNA analyses.
3. Our results indicate that, within the available literature, statistical methods are not always clearly described and raw data are rarely made publicly available. Furthermore, the choice of data pre-processing strategies and statistical tests used to assess quantitative correlations can significantly influence the likelihood of detecting positive correlations and the effect sizes.
4. Overall, we recommend the following: (i) increase transparency in method descriptions and data availability; (ii) assess correlations using mixed-effect models that can account for data characteristics; (iii) avoid pre-processing quantitative eDNA data, especially when combined with sub-optimal statistical tests. Implementing these guidelines should enhance the accessibility and transparency of quantitative eDNA data and ultimately their use for managers and policy makers.

Jonas Bylemans and Teun Everts shared first authors.

KEYWORDS
abundance, biomass, digital PCR (dPCR), environment DNA, qPCR, statistical guidelines

1 | INTRODUCTION

Analysing DNA present in environmental samples, commonly referred to as environmental DNA or eDNA, has emerged as a powerful, cost-efficient, and non-invasive tool for species detection (Ficetola et al., 2008; Pilliod et al., 2013; Takahara et al., 2012) and community monitoring (Bista et al., 2017; Creer et al., 2016; Hänfling et al., 2016). The explosive growth in the user base, continuous method development, and widespread adoption of eDNA analyses in recent years have led to a diverse array of methods for capturing and analysing eDNA (Hakimzadeh et al., 2023; Tsuji et al., 2019). This rapid expansion has prompted the need for clear guidelines and standardised protocols (Goldberg et al., 2016; Loeza-Quintana et al., 2020; Mathon et al., 2021). Although considerable attention has been given to field and laboratory methods, a similar level of emphasis has not been placed on the impact of data pre-processing and the possible statistical choices on the final conclusions.

Accurately modelling eDNA data requires careful consideration of the inherent characteristics of the data. First, environmental DNA survey data are typically hierarchically structured, with often multiple samples collected from numerous sites, and multiple PCR replicates for each sample (Buxton et al., 2021; Ficetola et al., 2015; Furlan et al., 2016). Analyses need to accommodate this hierarchical data structure as samples from the same sites and PCR replicates from the same samples will not be independent of each other. Second, surveys generate data in the form of counts (i.e. copies L^{-1} (C/L) or read numbers) which theoretically can only take positive integer values (e.g. one litre of sampled water cannot contain half a copy of DNA). A common approach for analysing such data is to assume the counts are Poisson distributed. However, the variance of count data is often significantly larger than the mean, resulting in overdispersion (Bliss & Fisher, 1953) which needs to be accommodated in the analysis. Third, zero values are common and may represent both false (e.g. related to errors in sampling or laboratory protocols) or true zeros (i.e. a species is not detected because it is not present at a sampling location). A higher than expected frequency of zero counts will cause zero-inflation (Blasco-Moreno et al., 2019; Heilbron, 1994), which also needs to be accommodated in the analysis. Consequently, accounting for the data structure, overdispersion, and zero-inflation is crucial to drawing sound statistical conclusions but remains a challenging endeavour (Arnqvist, 2020; Blasco-Moreno et al., 2019; Ives, 2015; O'Hara & Kotze, 2010; St-Pierre et al., 2018; Warton et al., 2016).

Environmental DNA surveys are widely used for species detection, but can be used to estimate species' abundances or biomass as it can be expected that, under ideal conditions, estimates of DNA copy numbers obtained from quantitative species-specific eDNA surveys (i.e. using quantitative real-time PCR (qPCR) or digital PCR (dPCR)) will be positively correlated with species abundance (A) or biomass (B). A wide range of data pre-processing strategies and statistical methods have been used to estimate these relationships, raising the question of whether different approaches could lead to varying outcomes and conclusions. In some cases, the hierarchical structure of the data (i.e. sampling multiple sites, mesocosms or experimental tanks; collecting multiple samples per site; performing multiple PCR replicates) is included in the statistical model specifications (Eichmiller et al., 2016; Hinlo et al., 2018; Lacoursière-Roussel et al., 2016). In other cases, the hierarchical structure is overlooked or the data are simplified by averaging over sample and/or PCR replicates (Doi et al., 2017; Kutti et al., 2020; Skinner et al., 2020; Thalinger et al., 2019). Many studies have dealt with the issue of over dispersed counts by log transforming eDNA concentration data, which often stabilises the variance (Doi et al., 2015; Dougherty et al., 2016; Thomsen et al., 2012). However, the impact of data transformations on statistical analyses remains contentious (Ives, 2015; O'Hara & Kotze, 2010). To address the issue of zero-inflation in eDNA data, one approach is to establish a lower bound for eDNA quantification. This can entail filtering out zeros or values below the limit of detection (LOD) or quantification (LOQ; Dunn et al., 2017; Takahara et al., 2012; Takahashi et al., 2020), although this approach ignores true zeros (Blasco-Moreno et al., 2019; Klymus et al., 2020). Another common approach to reduce zero-inflation is to average over replicates (i.e. taking mean values per sample or site) prior to analyses (Dougherty et al., 2016; Spear et al., 2021). However, averaging removes variation inherent to the data generating processes. Additionally, a wide range of statistical tests have been applied to assess the correlation between eDNA concentrations and abundance or biomass (A/B) including Pearson or Spearman correlations (Jo, 2023; Plough et al., 2018), linear or generalised linear models (Eichmiller et al., 2016; Pilliod et al., 2013) and Bayesian models (Erickson et al., 2016). These different statistical choices may strongly influence the resulting conclusions (Arnqvist, 2020; Gould et al., 2025).

To our knowledge, two studies have assessed the impact of different data pre-processing strategies (Jo, 2023) or statistical tests (Chambert et al., 2018) on quantitative species-specific eDNA data. However, a thorough evaluation of the combined impacts of the most common data pre-processing strategies and statistical tests is lacking. To bridge this gap, we performed a systematic literature search for quantitative species-specific eDNA studies that assessed the correlation between eDNA concentrations and species A/B. We evaluated the variety of data pre-processing strategies and statistical choices in the current literature and checked for the availability of raw data (i.e. where measurements are available for all levels of replication). We subsequently identified common pre-processing strategies and statistical tests and applied them to re-analyse the available empirical

datasets. Specifically, we assessed (i) the common analytical choices made during the analyses of quantitative eDNA data and the extent of variability in these approaches and (ii) how these analytical choices influence the ability to detect eDNA-abundance correlations and their effect size. Finally, we used our results to identify key considerations and best-practice guidelines that can improve the reliability and reproducibility of quantitative eDNA analyses.

2 | LITERATURE REVIEW AND DATA COLLECTION

We conducted a literature search on the 16th of March 2023 using Publish or Perish v 8.8.4275 (Harzing, 2007) to identify studies examining the relationship between quantitative and species-specific eDNA measurements and species A/B (Figure 1). A subset of 500 published studies was retrieved from Google Scholar, using search criteria that included "environmental DNA" or "eDNA" in the title, along with additional keywords such as {"biomass" OR "abundance"} AND "quantitative" AND {"qPCR" OR "ddPCR"}. We also considered studies referenced in a recent review (Rourke et al., 2021) and a meta-analysis (Carvalho et al., 2022) on the quantitative nature of species-specific eDNA surveys, as well as studies encountered prior to finalising the dataset (i.e. June 2023).

We performed a first evaluation of the retained publications by assessing if they reported correlation analyses between eDNA concentrations and species A/B, and at least three A/B measurements were documented (Figure 1). For all retained publications, we extracted details of the study system and focal species as well as choices made throughout the analytical workflow. These choices included the use of site, sample and PCR replicates during sampling and molecular analyses; methods for capturing and quantifying eDNA concentrations; the approaches used for summarising and/or transforming the raw data; statistical tests applied for correlation analyses and the availability of the associated data. We obtained quantitative estimates of potential observer biases through an independent assessment of a random subset of twenty-two published studies by two of the authors. Publications with publicly available raw data (i.e. eDNA concentrations reported for all replication levels) were compiled into a standard data format by a single author (see author contributions statement). Data compilation consisted of recording, for each publication (or individual regression if multiple correlations were performed within a

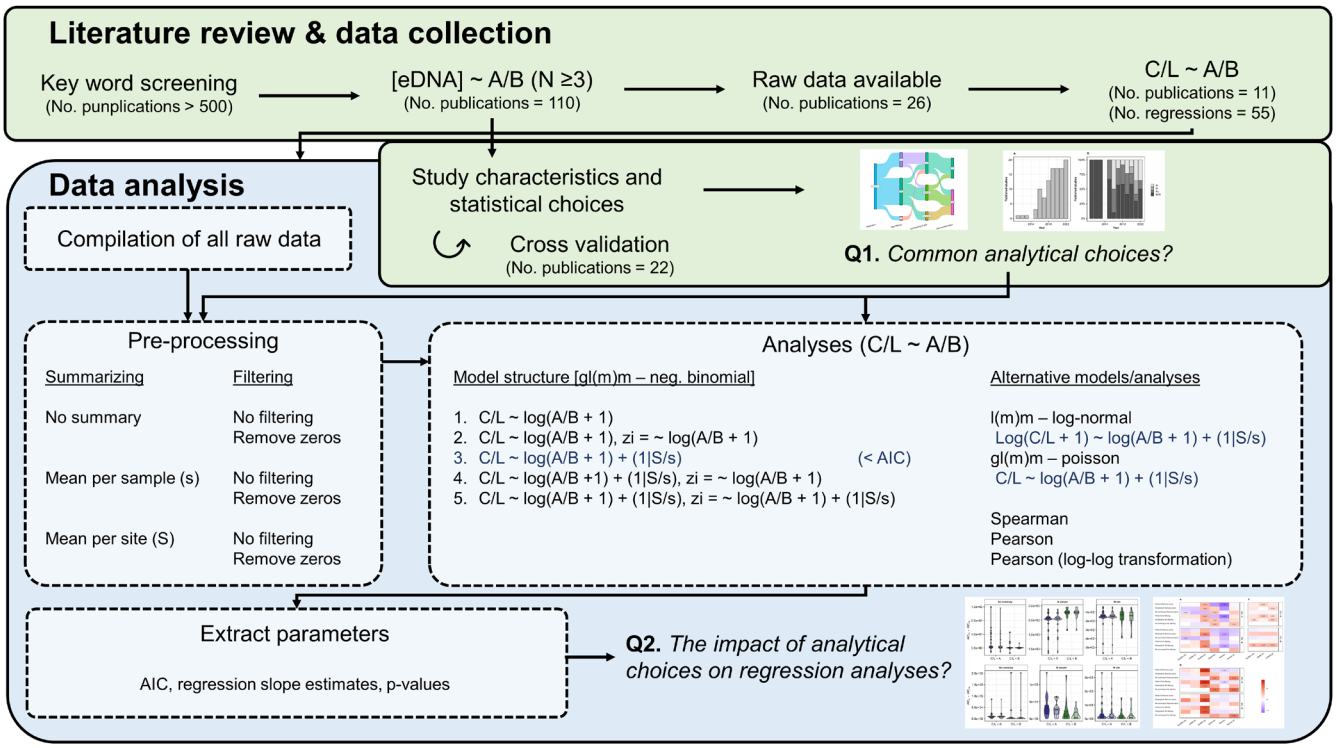

FIGURE 1 Schematic overview of the data collection and data analysis process. Research questions used to formulate guidelines on quantitative eDNA analyses are given (Q1 and Q2). Briefly, a literature review was used to identify studies reporting correlation analysis between environmental DNA concentrations ([eDNA]) and species abundance (A) or biomass (B). Relevant studies were used to assess study trends and the variety of analytical choices used during the analyses of quantitative eDNA data. The available raw data from these studies was subsequently compiled and pre-processed using strategies that were commonly found in literature. Generalized linear (mixed-effect) models (gl(m)m) with a negative binomial distribution, and including a zero-inflation model where relevant, were used to determine the optimal model structure (e.g. models highlighted in blue for illustrative purposes) based on Akaike Information Criterion (AIC) values. Subsequently, models with alternative distributions were fitted and alternative statistical tests were used to assess correlations and the impact of the various choices on regression outcomes. (additional legend: C/L = eDNA copies per litre, A/B = species abundance or biomass, zi = zero-inflation model, s = sample replicate, S = site replicate, l(m)m = linear (mixed-effect) model).

study), characteristics of the molecular assay used; limits of detection (LOD) and/or quantification (LOQ) if available; unique site, sample and/or PCR replicates; measures of eDNA concentrations in a single PCR; standardised eDNA concentration measurement (i.e. C/L or ng/L); and measurements of species A and B (Bylemans et al., 2025). In some cases, temporal surveys were conducted in one or a limited number of sites, rendering sampling events within a single site no longer independent of one another. Temporal sampling (i.e. sampling one site at different time points as species A/B change), however, also results in a hierarchical data structure similar to spatially explicit eDNA surveys. Therefore, we considered both spatial and temporal sampling surveys when compiling the data. Site level replicates were thus considered to be (i) spatially distinct locations, (ii) temporally distinct sampling of one or a few locations and (iii) experimental units (i.e. ponds, mesocosms or aquaria) sharing the same species A and/or B.

Visualisation of the data was conducted in R version 4.2.3 (R Development Core Team, 2010) using the R packages tidyverse (Wickham et al., 2019) and ggsankey (R package by David Sjoberg). To quantify the degree of observer bias during data extraction, we determined the overall congruence between observers using the percentage similarity for seven study characteristics (i.e. eDNA capture and quantification method, replication levels, type of data filtering and transformation, approach used to summarise the data, and the statistical test used) from the twenty-two independently assessed studies. We visualised trends in the number of publications and the data availability using bar plots while the variety of methodological choices were visualised using Sankey diagrams.

A total of 110 peer-reviewed publications published between 2011 and 2023 reported correlation analyses between some measure of eDNA concentration and species A/B. The overall congruence between independent observers carrying out the study assessments was 80.5% with the similarity for all seven study characteristics ranging from 65.9% to 90.9% (Figure S1) suggesting that the effects of observer biases on data collection is likely limited. The number of studies reporting correlation analyses between eDNA concentrations and species abundance/biomass increased over time (Figure 2). While the percentage of studies making their data publicly available increased from 2015 onwards, only a limited number of studies published the full raw data (i.e. 26 out of 106 published studies or 24.5% of studies between 2011 and 2022; Figure 2). Most studies to date were conducted in natural freshwater ecosystems with lentic and lotic systems being approximately equally represented, and marine and terrestrial ecosystems being underrepresented (Figure S2). Most studies focussed on fishes and used filtration and a quantitative real-time PCR approach with fluorescent probes to capture and quantify eDNA concentrations, respectively (Figure S2).

Across the selected studies, key choices throughout their analytical workflow were visualised in Figure 3. Most studies use replication throughout the eDNA analysis workflow (ca. 67%) (i.e. multiple sites sampled with multiple samples taken per site and multiple PCR replicates per sample). While most studies do not apply any data filtering prior to statistical analyses (ca. 74%), commonly encountered pre-processing strategies involve setting a lower bound to eDNA quantification results (i.e. filtering out zeros or values below the assay's LOD or LOQ) (ca. 19%) (Figure 3). Summarising eDNA quantification data prior to analysis is a commonly used strategy, with about one third of all published studies calculating the mean eDNA concentrations per sample (ca. 36%) or per site (ca. 32%) prior to

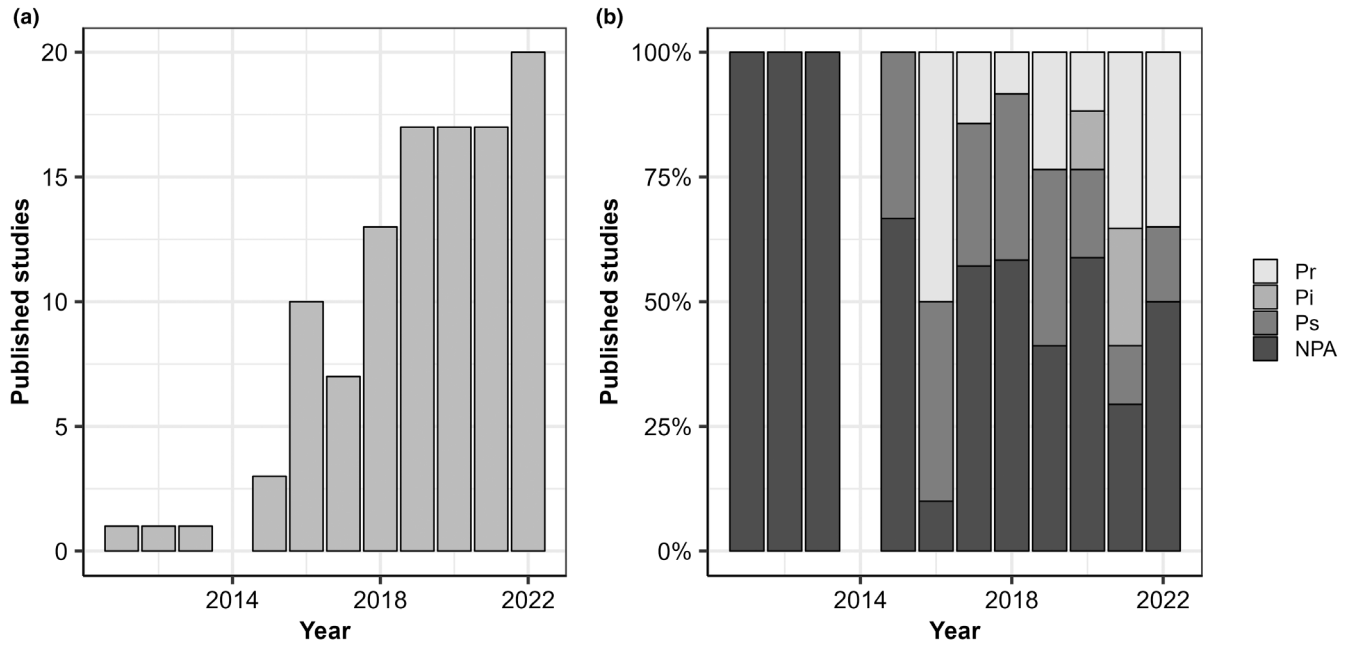

FIGURE 2 Yearly trends in the number of publications reporting a correlation between environmental DNA concentrations and species abundance/biomass (a) and the data availability for published studies over the years (b). Data availability categories are: Not publicly available (NPA), publicly available but the data is summarised (Ps), publicly available but the data is incomplete (Pi) or the complete and unmodified data is publicly available (Pr).

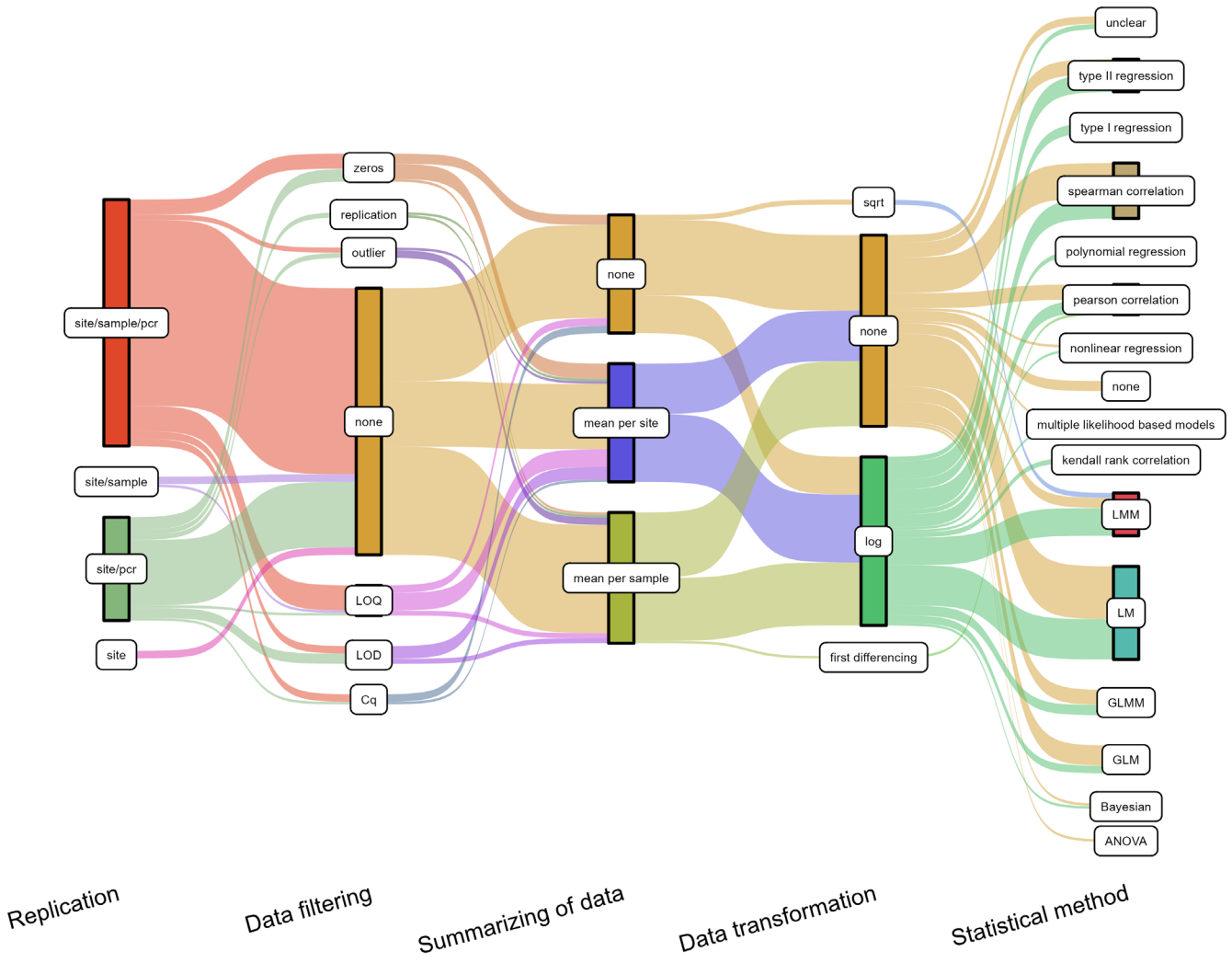

FIGURE 3 Sankey diagram visualising the key analytical choices found in the literature review of species-specific quantitative environmental DNA studies.

applying statistical tests (Figure 3). Most studies do not use a data transformation (ca. 55%) but the second most popular strategy is to log transform eDNA concentrations prior to statistical analyses (ca. 44%) (Figure 3). Finally, multiple statistical tests have been used in the current literature, falling into 15 different categories (Figure 3). Overall, linear and generalised linear models, with or without random effects, appear to be most popular (ca. 53%). Amongst the others, Spearman (ca. 15%) and Pearson (ca. 7%) correlations are frequently used to statistically test correlations (Figure 3).

3 | EVALUATING COMMON ANALYTICAL CHOICES

3.1 | Distributional assumptions

We considered several approaches for modelling the distribution of quantitative eDNA data. These data typically have the following count characteristics: (i) consisting of positive values (e.g. number of copies per L); (ii) exhibiting overdispersion; and (iii) often having an excess of zeros. Accounting for excess zeros in statistical modelling can be achieved through the use of zero-inflated models (Brooks et al., 2017). The added benefit of these models is that they effectively use a qualitative modelling for all zero values which has previously been recommended when eDNA concentrations fall below the LOQ (Klymus et al., 2020). We assume that a Negative Binomial model would be the 'optimal' model as it can account for both overdispersion and, when including a zero-inflation model, the excess of zeros in count data (Blasco-Moreno et al., 2019). Additional distributions frequently used to model quantitative eDNA were compared to this benchmark. For instance, a zero-inflated Log-Normal distribution (i.e. applying a $\log(y+1)$ transformation of the response variable) can accommodate overdispersion and excess zeros but only when overdispersion is limited (O'Hara & Kotze, 2010). Alternatively, a zero-inflated Poisson distribution can account for excess zeros but not overdispersion. Finally, normal distributions have previously been used to model quantitative eDNA data but we did not consider this in the current analyses

given the obvious inability to adequately model data with the earlier described characteristics.

3.2 | Data structure and pre-processing

Appropriately incorporating the hierarchical data structure can be done through: (i) summarising the data by averaging over sample or PCR replicates or (ii) using appropriate hierarchical models. While summarising the data prior to analyses effectively removes the data structure and also reduces the impact of excess zeros, it may reduce statistical power (Jo, 2023). In contrast, mixed-effect models can accommodate a hierarchical data structure and current statistical packages can fit such models using a wide range of distributions while also accounting for zero-inflation (Arnqvist, 2020; Brooks et al., 2017). Finally, setting lower bound threshold values for quantitative results and thus removing all zeros is a common data filtering approach which effectively removes zero-inflation but may remove potentially relevant data.

3.3 | Re-analyses of empirical datasets

Models fitting was done in R version 4.2.3 using the R packages glmmTMB (Brooks et al., 2017), performance (Lüdecke et al., 2021) and model performance was assessed using the packages broom.mixed (Bolker & Robinson, 2019) and emmeans (Lenth, 2024). Final analyses were conducted on a subset of studies where replication was present across all hierarchical levels (i.e. 16 studies reporting 81 individual regressions) and eDNA concentrations were measured as C/L (i.e. 12 out of 16 studies reporting 60 individual regressions). Studies reporting quantitative eDNA measures as ng/L were excluded (5 out of 16 studies reporting 21 individual regressions) as they contain zero-inflated positive decimal values which are notoriously challenging to model correctly. For final analyses, some studies or subgroups were excluded due to the presence of a high number of zero values (Plough et al., 2018) or limited variation in species A/B estimates (Takahashi et al., 2020) (Figures S3–S6). Removal of these data alleviated initial convergence problems when fitting models and resulted in a final dataset of 11 studies reporting 55 individual regressions (i.e. C/L ~ A: 10 studies reporting 45 regressions, C/L ~ B: 3 studies reporting 15 regressions; Figure 1).

Commonly used data summarising (i.e. no summary or taking means per sample or per site) or filtering approaches (i.e. no filtering or removal of zeros) were applied prior to model fitting thus generating six different datasets for each individual regression. We fitted models separately for each individual regression between eDNA copy numbers and species A or B and each of the data pre-processing strategies. Firstly, a Negative Binomial regression model was used to determine optimal model structure (see Figure 1). Copy numbers per litre were modelled using the log transformed A or B estimates as a fixed effect. Random effects (i.e. samples nested within sites when no data pre-processing was used) and/or a zero-inflation model (i.e. considering only A/B as fixed effect or including both fixed and random effects) were subsequently included to account for the hierarchical data structure and zero-inflation (see Supporting Information for full model descriptions). Random effects and zero-inflation models were adjusted according to data pre-processing strategies (e.g. sample effects were removed from the random effects if means were taken per sample and zero-inflation models were excluded if zeros were removed prior to model fitting). Akaike Information Criterion (AIC) values (Akaike, 1974, 1978) were used to select the optimal model structure, which was subsequently used to assess model fit using two commonly used alternative distributions, namely, a Log-Normal (i.e. applying a log[C/L + 1] transformation) and Poisson distribution. Model parameter estimates, associated p-values and AIC values (i.e. corrected in the case of the Log-Normal regression) were extracted. A Wilcoxon signed-rank test for paired samples was used to assess if the Negative Binomial distribution provides a better fit compared to the alternative distributions (i.e. lower AIC values) for all combinations of data pre-processing strategies. Finally, Spearman and Pearson tests were performed to assess the correlations between eDNA concentrations and species A/B for all individual regressions and the different combinations of data pre-processing strategies. Pearson correlations were also performed after log transforming response and predictor variables (i.e. log–log transformation). This transformation was not considered for the Spearman correlations as this is a non-parametric rank-based test and more robust to potential outliers.

For all individual regressions, explanatory variables (A/B), data pre-processing steps (i.e. filtering and summarising) and statistical tests; the ability to detect a significant positive correlation between eDNA concentrations and species A/B at $\alpha = 0.05$ was encoded as binomial data. For this, we only considered the conditional, and not the zero-inflation, part of the models as the latter effectively models eDNA detection probabilities. This approach also facilitated comparisons between different data pre-processing strategies, which may exclude the need for a zero-inflation model. We subsequently used a generalised linear mixed-effect model to assess the impacts of the different choices in the analytical workflow. The probability to detect a significant positive correlation was modelled as a function of the statistical tests and the two main data pre-processing steps used. Individual regressions nested within studies were included as random effects to account for between-study variance. AIC-based forward model selection was used to evaluate whether the inclusion of two-way interactions between fixed effects improved model fit. Pairwise comparisons using estimated marginal means were conducted to assess the significance of between-group effects relative to the "optimal" model, after applying a false discovery rate correction (Benjamini & Hochberg, 1995). The above analyses were rerun using an α-value of 0.01 to determine the significance of regression slope estimates to assess if more conservative thresholds would influence the results. Finally, a similar approach was used to evaluate the impact of the statistical choices on model regression slope estimates as a proxy for effect size using a linear mixed-effect

model and an ordered quantile normalisation transformation of the response variable to ensure normality of the data.

3.3.1 | The neglected impact of distributional assumptions

Comparisons of AIC values revealed that a Negative Binomial distribution best describes raw quantitative eDNA data as AIC values are significantly lower when compared to models assuming a Log-Normal or Poisson distribution ($p < 0.0001$; Figure 4). When plotting the predicted versus observed values for a random subset of 50 regressions, we found that the Negative Binomial distribution provided a good fit to the data (Figure S7). Pairwise comparisons of the probability to detect significant positive correlations and the estimated regression slopes were conducted using the Negative Binomial model without any prior data modifications as a benchmark. Comparisons between different statistical tests based on the raw data revealed an increase in type I errors (i.e. false positives) when using Spearman and Pearson regressions, and more conservative significance estimates only enhanced these patterns (Figure 5a,b). No significant effects on the probability to detect a significant positive correlation and the estimates of regression slopes were observed for the Log-Normal and Poisson models, and the use of more conservative alpha values reduced the occurrence of type I and II (i.e. false negatives) errors (Figure 5).

3.3.2 | Data structure and pre-processing affect model inferences

Comparison of AIC values showed that a Log-Normal distribution fitted the quantitative eDNA data better when mean values are taken per sample or site, while the Negative Binomial model outperformed the Poisson model when applying common data pre-processing strategies ($p < 0.0001$; Figure 4). Considering the pairwise comparisons between mixed-effect model parameters, the general observed pattern was a decreasing ability to detect

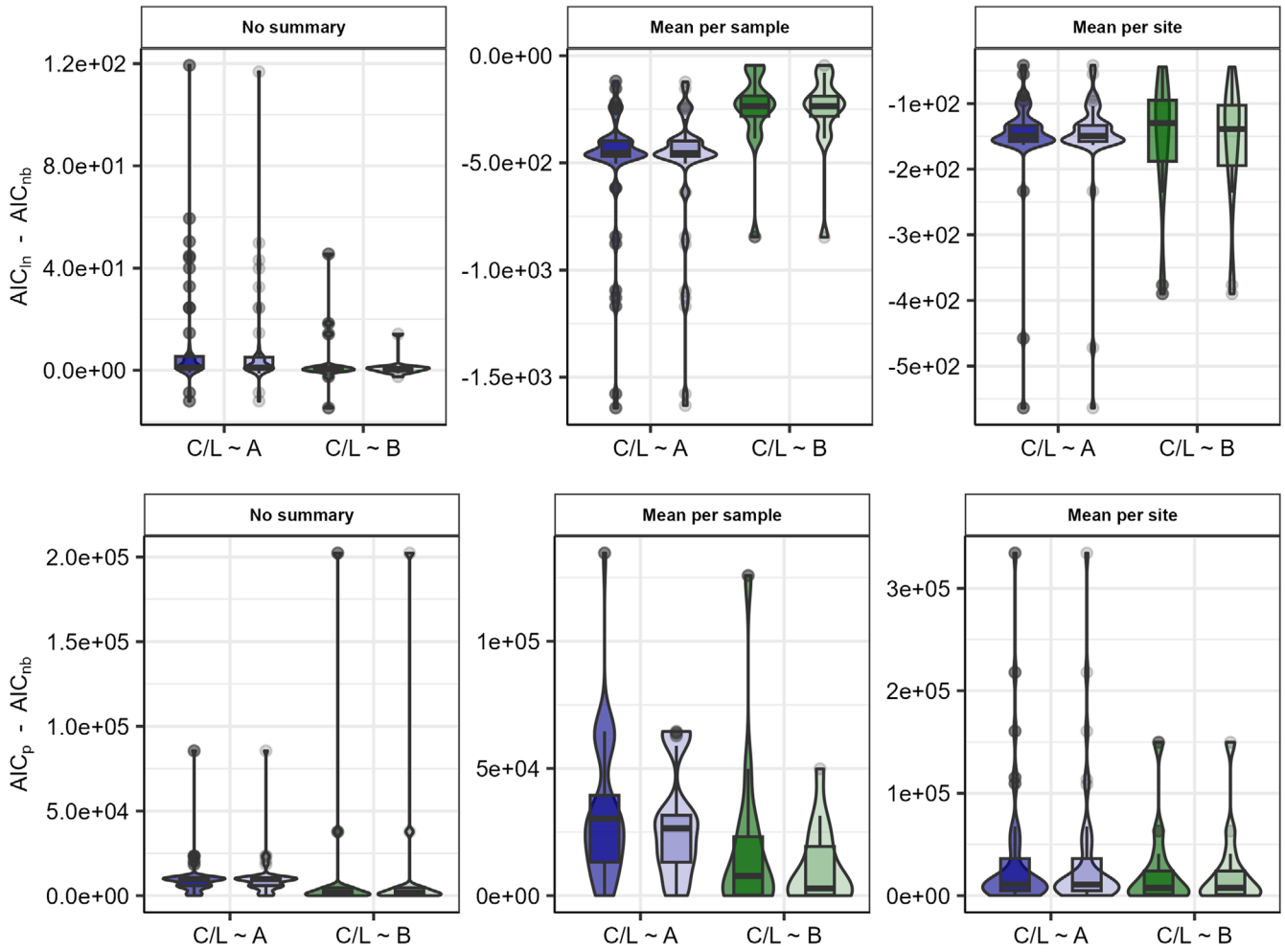

FIGURE 4 Delta Akaike Information Criterion (AIC) values for the alternative distributions used in the mixed-effect models (i.e. the log-normal (AIC_{ln}) for the upper panels and Poisson distribution (AIC_p) for the lower pnels) relative to the negative binomial distribution (AIC_{nb}). Results are given for different regression types (*x*-axis and different colours) and the different possible strategies for summarising (horizontal panels) and filtering (i.e. darker colours: No filtering, lighter colours: Removal of zeros) the data.

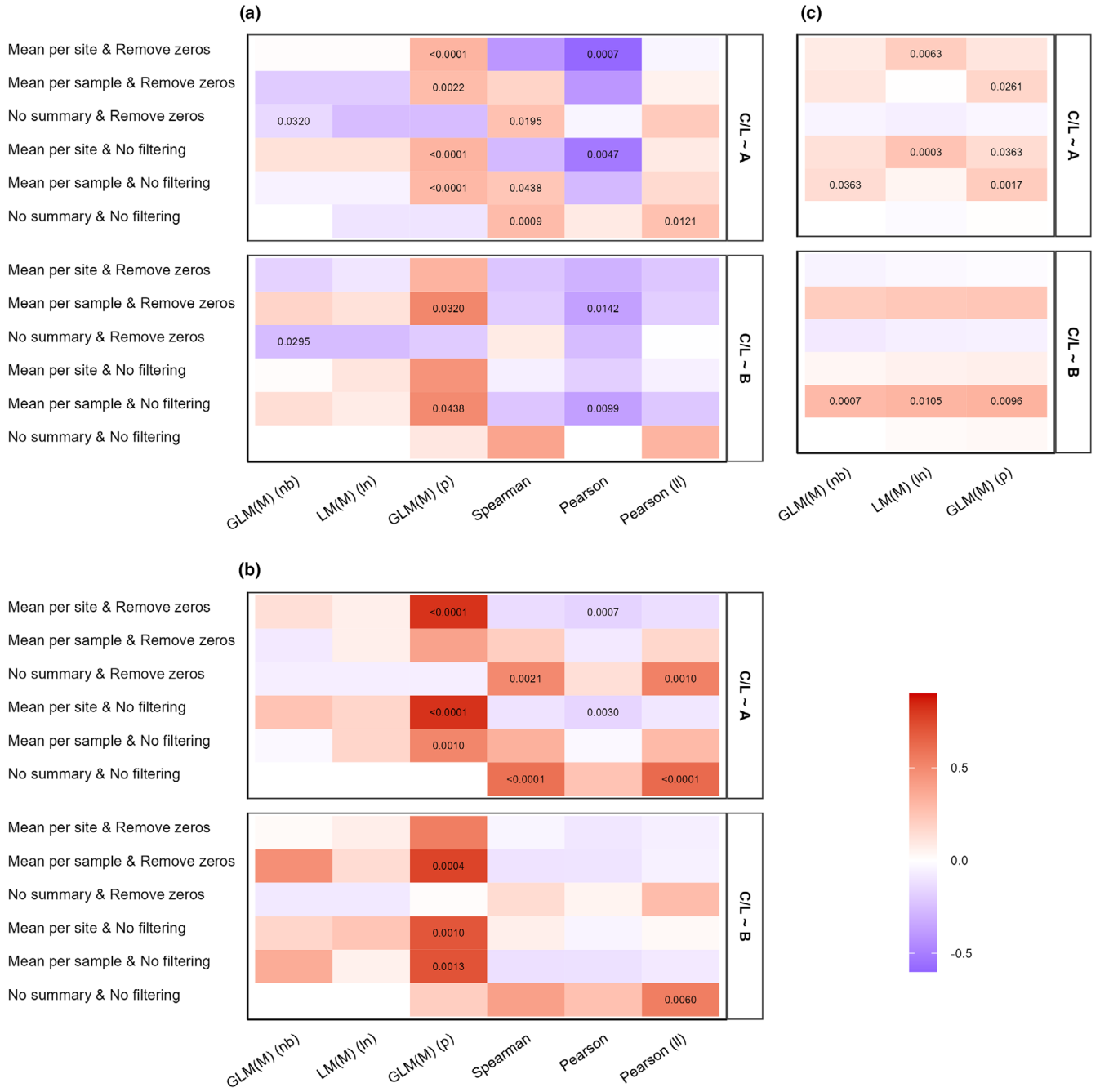

FIGURE 5 Influence of different combinations of statistical choices, relative to "optimal" model using a Negative Binomial distribution and without prior data filtering or summarising, on the probability of detecting a significant positive correlation (a: $\alpha = 0.05$, b: $\alpha = 0.01$) and the estimates of the regression slope ($\alpha = 0.05$) (c). Commonly used statistical test are given on the *x*-axis while the different strategies for summarising and filtering the data prior to statistical analyse are given on the *y*-axis. Colours indicate the relative decrease (blue) or increase (red) in mean estimates relative to the "optimal" model for each regression type (vertical panels). When relative differences where statistically significant ($\alpha = 0.05$), *p*-values are given.

significant positive correlations and lower slope estimates when removing zero values prior to analyses (Figure 5). Taking the means per sample or site resulted in an overestimation of significant positive correlations and regression slope estimates, with these effects being most pronounced when using a Poisson model (Figure 5). Applying pre-processing strategies prior to Spearman and Pearson correlations tended to reduce the impact of type I errors, especially when taking mean values per sample or site (Figure 5). In particular, applying Pearson analyses to the untransformed data and taking the mean values per sample or site may result in an increase in type II errors. While the use of more conservative alpha values to assess the significance of correlations reduces the prevalence of type II errors, it increased type I errors when using a Poisson model distribution or Spearman and Pearson correlations (Figure 5).

4 | GENERAL RECOMMENDATIONS

Since 2011, the number of quantitative species-specific eDNA studies has steadily increased. This growth resulted in a wide variety of data pre-processing approaches and statistical methods that have been applied to assess potential correlations between eDNA concentrations and species A/B. Surprisingly, a formal evaluation of the impact of these different strategies and formulating some general guidelines to handle such data is lacking. A re-analysis of available empirical datasets shows that the choice of pre-processing strategy or statistical test can have profound effects on type I and II error rates and estimated effect sizes (i.e. regression slope). Harmonising analytical workflows is urgently needed to increase the reliability of quantitative eDNA studies and to ensure that sound conclusions can be drawn from future studies, literature reviews and/or meta-analyses (e.g. Carvalho et al., 2022; Rourke et al., 2021). Several key recommendations arise from our study, which can significantly improve the reliability and credibility of quantitative eDNA analyses, and thus, their application in conservation programmes and policy making.

4.1 | Increased efforts are needed to provide clear descriptions of statistical workflows and make raw unmodified datasets publicly available

The first standard step that should be promoted is providing full and detailed descriptions of statistical analyses and making raw data publicly available (Reichman et al., 2011). Unfortunately, many current studies lack appropriate (meta-)data standards, making it a challenging endeavour to understand the available data or reproduce the analysis. Moreover, despite the availability of multiple free data repositories for storing raw research data and the requirement for a data accessibility statement by an increasing number of journals, only a quarter of the studies reviewed here published their raw data. During our search we came across instances where the data was not available through the provided links, the metadata were unavailable or too limited to fully understand all reported measures, the data was available but in a summarised format or only a subset of the data was available. While the unavailability of raw data remains an issue across various fields (Miyakawa, 2020), we urge the eDNA community to make an extra effort to store unmodified data in public repositories and ensure there is well-documented metadata, even if this is not required by the journal.

4.2 | Negative binomial mixed-effect models that account for both the hierarchical data structure and zero-inflation should be preferred to assess relationships between eDNA concentrations and species A/B

The use of mixed-effect models, particularly those with a Negative Binomial distribution, is the most appropriate standard for modelling eDNA copy numbers. Our study confirms the findings by Chambert et al. (2018) that a Negative Binomial distribution better fits quantitative eDNA data. However, alternative modelling distributions do not seem to have a significant impact on model estimates as long as data were not modified and appropriate measures are taken to account for the hierarchical data structure and the impact of zero-inflation. Patterns in the pairwise comparisons are generally consistent between the Negative Binomial and Log-Normal models, thus indicating that Log-Normal models can be a valuable alternative to Negative Binomial models when the latter prove challenging to fit. This aligns with Ives (2015) who found that, when testing for the significance of regressions, a log transformation of count data can be used. Our review indicates that a wide variety of statistical tests have been used to assess correlations between eDNA concentrations and species A/B, highlighting the need for standardisation of statistical tests to facilitate more direct comparisons between studies. Overall, we recommend against the use of Spearman and Pearson correlations due to the increased risk of type I errors and the inability to account for data structure and zero-inflation in the analyses.

4.3 | Pre-processing of quantitative eDNA data should be avoided, especially in combination with sub-optimal statistical tests

Data pre-processing strategies should, as much as possible, be avoided. Our results indicate that common pre-processing strategies can increase type I and type II errors and affect effect size estimates. The impact of these pre-processing strategies is especially profound when combined with sub-optimal statistical analyses (i.e. Poisson models, Spearman and Pearson correlations). Our literature review revealed that most studies applied at least one data pre-processing step. While filtering out values (e.g. based on LOD, LOQ, outliers, etc.) was relatively uncommon, many studies averaged data across sample or site replicates. Summarising data prior to statistical analyses was found to increase type I errors especially when using mixed-effect Poisson models for regression analyses. While taking mean values across PCR and/or sample replicates is common and likely to give a value closer to the 'true' eDNA concentrations, especially for low concentrations (Ellison et al., 2006), this ignores the variability inherent to the data generating processes which can be effectively accounted for through mixed-effect models. Earlier work has also indicated that pooling measurements reduces the likelihood of detecting significant correlations and attributed this to a reduced sample size (Jo, 2023). Due to the limited availability of LOD and LOQ estimates, we did not assess the effects of discarding results below the LOD or LOQ. However, previous research has recommended that eDNA concentration below the LOQ should only be considered qualitatively (Klymus et al., 2020). In this regard, it is noteworthy that the use of zero-inflated mixed-effect models effectively addresses this issue for all zero values. Overall, given the range of pre-processing choices, the use of models that can account for the data characteristics (i.e. counts, excess zeros and hierarchical structure) should be preferred (Arnqvist, 2020; St-Pierre et al., 2018; Stroup, 2015).

AUTHOR CONTRIBUTIONS

Jonas Bylemans and Teun Everts designed the study and performed the literature review. Jonas Bylemans performed the data compilation, analyses, and visualisations with significant input from all co-authors. Jonas Bylemans and Teun Everts led the writing of the manuscript with significant contributions from Rein Brys and Richard P. Duncan. All authors revised and approved the final manuscript for publication.

ACKNOWLEDGEMENTS

Jonas Bylemans was funded by AnaEE France and the Pôle R&D Écla during the time of this research. Teun Everts was funded by Research Foundation Flanders grants (S01822N).

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to declare.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.70064.

DATA AVAILABILITY STATEMENT

All data and R scripts are publicly available through the figshare data repository https://doi.org/10.6084/m9.figshare.28937870.v1 (Bylemans et al., 2025).

ORCID

Jonas Bylemans https://orcid.org/0000-0001-6263-0874
Teun Everts https://orcid.org/0000-0001-7862-4209
Rein Brys https://orcid.org/0000-0002-0688-3268
Richard P. Duncan https://orcid.org/0000-0003-2295-449X

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723. https://doi.org/10.1109/TAC.1974.1100705

Akaike, H. (1978). On the likelihood of a time series model. *Journal of the Royal Statistical Society. Series D*, *27*, 217–235. https://doi.org/10.2307/2988185

Arnqvist, G. (2020). Mixed models offer no freedom from degrees of freedom. *Trends in Ecology & Evolution*, *35*, 329–335. https://doi.org/10.1016/j.tree.2019.12.004

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bista, I., Carvalho, G. R., Walsh, K., Seymour, M., Hajibabaei, M., Lallias, D., Christmas, M., & Creer, S. (2017). Annual time-series analysis of aqueous eDNA reveals ecologically relevant dynamics of lake ecosystem biodiversity. *Nature Communications*, *8*, 14087. https://doi.org/10.1038/ncomms14087

Blasco-Moreno, A., Pérez-Casany, M., Puig, P., Morante, M., & Castells, E. (2019). What does a zero mean? Understanding false, random and structural zeros in ecology. *Methods in Ecology and Evolution*, *10*, 949–959. https://doi.org/10.1111/2041-210X.13185

Bliss, C. I., & Fisher, R. A. (1953). Fitting the negative binomial distribution to biological data. *Biometrics*, *9*, 176–200. https://doi.org/10.2307/3001850

Bolker, B., & Robinson, D. (2019). broom.mixed: Tidying methods for mixed models. R Package Version, 667(0.2), 4.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*, 378. https://doi.org/10.32614/RJ-2017-066

Buxton, A., Matechou, E., Griffin, J., Diana, A., & Griffiths, R. A. (2021). Optimising sampling and analysis protocols in environmental DNA studies. *Scientific Reports*, *11*, 11637. https://doi.org/10.1038/s41598-021-91166-7

Bylemans, J., Everts, T., Brys, R., & Duncan, R. P. (2025). Bylemans_et_al_2025_MEE.zip. Dataset. *figshare*. https://doi.org/10.6084/m9.figshare.28937870.v1

Carvalho, C. S., de Oliveira, M. E., Rodriguez-Castro, K. G., Saranholi, B. H., & Galetti, P. M. (2022). Efficiency of eDNA and iDNA in assessing vertebrate diversity and its abundance. *Molecular Ecology Resources*, *22*, 1262–1273. https://doi.org/10.1111/1755-0998.13543

Chambert, T., Pilliod, D. S., Goldberg, C. S., Doi, H., & Takahara, T. (2018). An analytical framework for estimating species density from environmental DNA. *Ecology and Evolution*, *8*, 3468–3477. https://doi.org/10.1002/ece3.3764

Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W. K., Potter, C., & Bik, H. M. (2016). The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution*, *7*, 1008–1018. https://doi.org/10.1111/2041-210X.12574

Doi, H., Inui, R., Akamatsu, Y., Kanno, K., Yamanaka, H., Takahara, T., & Minamoto, T. (2017). Environmental DNA analysis for estimating the abundance and biomass of stream fish. *Freshwater Biology*, *62*, 30–39. https://doi.org/10.1111/fwb.12846

Doi, H., Uchii, K., Takahara, T., Matsuhashi, S., Yamanaka, H., & Minamoto, T. (2015). Use of droplet digital PCR for estimation of fish abundance and biomass in environmental DNA surveys. *PLoS One*, *10*, e0122763.

Dougherty, M. M., Larson, E. R., Renshaw, M. A., Gantz, C. A., Egan, S. P., Erickson, D. M., & Lodge, D. M. (2016). Environmental DNA (eDNA) detects the invasive rusty crayfish Orconectes rusticus at low abundances. *Journal of Applied Ecology*, *53*, 722–732. https://doi.org/10.1111/1365-2664.12621

Dunn, N., Priestley, V., Herraiz, A., Arnold, R., & Savolainen, V. (2017). Behavior and season affect crayfish detection and density inference using environmental DNA. *Ecology and Evolution*, *7*, 7777–7785. https://doi.org/10.1002/ece3.3316

Eichmiller, J. J., Miller, L. M., & Sorensen, P. W. (2016). Optimizing techniques to capture and extract environmental DNA for detection and quantification of fish. *Molecular Ecology Resources*, *16*, 56–68. https://doi.org/10.1111/1755-0998.12421

Ellison, S. L., English, C. A., Burns, M. J., & Keer, J. T. (2006). Routes to improving the reliability of low level DNA analysis using real-time PCR. *BMC Biotechnology*, *6*, 33. https://doi.org/10.1186/1472-6750-6-33

Erickson, R. A., Rees, C. B., Coulter, A. A., Merkes, C. M., McCalla, S. G., Touzinsky, K. F., Walleser, L., Goforth, R. R., & Amberg, J. J. (2016). Detecting the movement and spawning activity of bigheaded carps with environmental DNA. *Molecular Ecology Resources*, *16*, 957–965. https://doi.org/10.1111/1755-0998.12533

Ficetola, G. F., Miaud, C., Pompanon, F., & Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biology Letters*, *4*, 423–425. https://doi.org/10.1098/rsbl.2008.0118

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 presence / absence from eDNA metabarcoding data. *Molecular Ecology Resources*, *15*, 543–556. https://doi.org/10.1111/1755-0998.12338

Furlan, E. M., Gleeson, D., Hardy, C. M., & Duncan, R. P. (2016). A framework for estimating the sensitivity of eDNA surveys. *Molecular Ecology Resources*, *16*, 641–654. https://doi.org/10.1111/1755-0998.12483

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., Spear, S. F., McKee, A., Oyler-McCance, S. J., Cornman, R. S., Laramie, M. B., Mahon, A. R., Lance, R. F., Pilliod, D. S., Strickler, K. M., Waits, L. P., Fremier, A. K., Takahara, T., Herder, J. E., … Gilbert, M. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution*, *7*(11), 1299–1307. https://doi.org/10.1111/2041-210X.12595

Gould, E., Fraser, H. S., Parker, T. H., Nakagawa, S., Griffith, S. C., Vesk, P. A., Fidler, F., Hamilton, D. G., Abbey-Lee, R. N., Abbott, J. K., Aguirre, L. A., Alcaraz, C., Aloni, I., Altschul, D., Arekar, K., Atkins, J. W., Atkinson, J., Baker, C. M., Barrett, M., … Whelan, S. (2025). Same data, different analysts: Variation in effect sizes due to analytical decisions in ecology and evolutionary biology. *BMC Biology*, *23*(1), 23–35. https://doi.org/10.1186/s12915-024-02101-x

Hakimzadeh, A., Abdala Asbun, A., Albanese, D., Bernard, M., Buchner, D., Callahan, B., Caporaso, J. G., Curd, E., Djemiel, C., Brandström Durling, M., Elbrecht, V., Gold, Z., Gweon, H. S., Hajibabaei, M., Hildebrand, F., Mikryukov, V., Normandeau, E., Özkurt, E., M Palmer, J., … Anslan, S. (2023). A pile of pipelines: An overview of the bioinformatics software for metabarcoding data analyses. *Molecular Ecology Resources*, *24*, e13847. https://doi.org/10.1111/1755-0998.13847

Hänfling, B., Handley, L. L., Read, D. S., Hahn, C., & Li, J. (2016). Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology*, *25*, 3101–3119. https://doi.org/10.1111/mec.13660

Harzing, A. W. (2007). *Publish or Perish*. [Computer software]. https://harzing.com/resources/publish-or-perish

Heilbron, D. C. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*, *36*, 531–547. https://doi.org/10.1002/bimj.4710360505

Hinlo, R., Lintermans, M., Gleeson, D., Broadhurst, B., & Furlan, E. (2018). Performance of eDNA assays to detect and quantify an elusive benthic fish in upland streams. *Biological Invasions*, *20*, 3079–3093. https://doi.org/10.1007/s10530-018-1760-x

Ives, A. R. (2015). For testing the significance of regression coefficients, go ahead and log-transform count data. *Methods in Ecology and Evolution*, *6*, 828–835. https://doi.org/10.1111/2041-210X.12386

Jo, T. S. (2023). Pooling of intra-site measurements inflates variability of the correlation between environmental DNA concentration and organism abundance. *Environmental Monitoring and Assessment*, *195*, 936. https://doi.org/10.1007/s10661-023-11539-5

Klymus, K. E., Merkes, C. M., Allison, M. J., Goldberg, C. S., Helbing, C. C., Hunter, M. E., Jackson, C. A., Lance, R. F., Mangan, A. M., Monroe, E. M., Piaggio, A. J., Stokdyk, J. P., Wilson, C. C., & Richter, C. A. (2020). Reporting the limits of detection and quantification for environmental DNA assays. *Environmental DNA*, *2*, 271–282. https://doi.org/10.1002/edn3.29

Kutti, T., Johnsen, I. A., Skaar, K. S., Ray, J. L., Husa, V., & Dahlgren, T. G. (2020). Quantification of eDNA to map the distribution of cold-water coral reefs. *Frontiers in Marine Science*, *7*, 446. https://doi.org/10.3389/fmars.2020.00446

Lacoursière-Roussel, A., Rosabal, M., & Bernatchez, L. (2016). Estimating fish abundance and biomass from eDNA concentrations: Variability among capture methods and environmental conditions. *Molecular Ecology Resources*, *16*, 1401–1414. https://doi.org/10.1111/1755-0998.12522

Lenth, R. V. (2024). emmeans: Estimated marginal means, aka least-squares means. R package version 1.10.5. https://doi.org/10.32614/CRAN.package.emmeans

Loeza-Quintana, T., Abbott, C. L., Heath, D. D., Bernatchez, L., & Hanner, R. H. (2020). Pathway to increase standards and competency of eDNA surveys (PISCeS)-advancing collaboration and standardization efforts in the field of eDNA. *Environmental DNA*, *2*, 255–260. https://doi.org/10.1002/edn3.112

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, *6*(60), 3139. https://doi.org/10.21105/joss.03139

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*, 2565–2579. https://doi.org/10.1111/1755-0998.13430

Miyakawa, T. (2020). No raw data, no science: Another possible source of the reproducibility crisis. *Molecular Brain*, *13*, 24. https://doi.org/10.1186/s13041-020-0552-2

O'Hara, R. B., & Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, *1*, 118–122. https://doi.org/10.1111/j.2041-210X.2010.00021.x

Pilliod, D. S., Goldberg, C. S., Arkle, R. S., & Waits, L. P. (2013). Estimating occupancy and abundance of stream amphibians using environmental DNA from filtered water samples. *Canadian Journal of Fisheries and Aquatic Sciences*, *10*, 1123–1130. https://doi.org/10.1139/cjfas-2013-0047

Plough, L. V., Ogburn, M. B., Fitzgerald, C. L., Geranio, R., Marafino, G. A., & Richie, K. D. (2018). Environmental DNA analysis of river herring in Chesapeake Bay: A powerful tool for monitoring threatened keystone species. *PLoS One*, *13*, e0205578. https://doi.org/10.1371/journal.pone.0205578

R Development Core Team. (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and apportunities of open data in ecology. *Science*, *331*, 703–705. https://doi.org/10.1126/science.1197962

Rourke, M. L., Fowler, A. M., Hughes, J. M., Broadhurst, M. K., DiBattista, J. D., Fielder, S., Wilkes Walburn, J., & Furlan, E. M. (2021). Environmental DNA (eDNA) as a tool for assessing fish biomass: A review of approaches and future considerations for resource surveys. *Environmental DNA*, *4*, 9–33. https://doi.org/10.1002/edn3.185

Skinner, M., Murdoch, M., Loeza-Quintana, T., Crookes, S., & Hanner, R. (2020). A mesocosm comparison of laboratory-based and on-site eDNA solutions for detection and quantification of striped bass (*Morone saxatilis*) in marine ecosystems. *Environmental DNA*, *2*, 298–308. https://doi.org/10.1002/edn3.61

Spear, M. J., Embke, H. S., Krysan, P. J., & Vander Zanden, M. J. (2021). Application of eDNA as a tool for assessing fish population abundance. *Environmental DNA*, *3*, 83–91. https://doi.org/10.1002/edn3.94

St-Pierre, A. P., Shikon, V., & Schneider, D. C. (2018). Count data in biology—Data transformation or model reformation? *Ecology and Evolution*, *8*, 3077–3085. https://doi.org/10.1002/ece3.3807

Stroup, W. W. (2015). Rethinking the analysis of non-Normal data in plant and soil science. *Agronomy Journal*, *107*, 811–827. https://doi.org/10.2134/agronj2013.0342

Takahara, T., Minamoto, T., Yamanaka, H., Doi, H., & Kawabata, Z. (2012). Estimation of fish biomass using environmental DNA. *PLoS One*, *7*, e35868. https://doi.org/10.1371/journal.pone.0035868

Takahashi, S., Sakata, M. K., Minamoto, T., & Masuda, R. (2020). Comparing the efficiency of open and enclosed filtration systems in environmental DNA quantification for fish and jellyfish. *PLoS One*, *15*, e0231718. https://doi.org/10.1371/journal.pone.0231718

Thalinger, B., Wolf, E., Traugott, M., & Wanzenböck, J. (2019). Monitoring spawning migrations of potamodromous fish species via eDNA. *Scientific Reports*, *9*, 15388. https://doi.org/10.1038/s41598-019-51398-0

Thomsen, P. F., Kielgast, J., Iversen, L. L., Wiuf, C., Rasmussen, M., Gilbert, M. T. P., Orlando, L., & Willerslev, E. (2012). Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, *21*, 2565–2573. https://doi.org/10.1111/j.1365-294X.2011.05418.x

Tsuji, S., Takahara, T., Doi, H., Shibata, N., & Yamanaka, H. (2019). The detection of aquatic macroorganisms using environmental DNA analysis-a review of methods for collection, extraction, and detection. *Environmental DNA*, *1*, 99–108. https://doi.org/10.1002/edn3.21

Warton, D. I., Lyons, M., Stoklosa, J., & Ives, A. R. (2016). Three points to consider when choosing a LM or GLM test for count data. *Methods in Ecology and Evolution*, *7*, 882–890. https://doi.org/10.1111/2041-210X.12552

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*, 1686. https://doi.org/10.21105/joss.01686

SUPPORTING INFORMATION

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

Table S1. Full descriptions of the Negative Binomial models used to model the relationship between environmental DNA (eDNA) concentrations and species abundance (A) or biomass (B).

Figure S1. The percentage of congruence based on the twenty-two independently assessed publications.

Figure S2. Sankey diagrams highlighting the key study systems and target taxa used to study the correlation between environmental DNA (eDNA) concentrations and species abundance/biomass (A) and temporal trends in the methods used for capturing (B) and quantifying (C) eDNA.

Figure S3. Overview of the raw data for all studies (top panel names) and individual regressions (lower panel IDs) reporting correlations analyses between environment DNA copies L^{-1} and species abundance.

Figure S4. Visualisation of the distribution of the environmental DNA concentration measures for all studies (top panel names) and individual regressions (lower panel IDs) reporting correlations analyses between environment DNA copies L^{-1} and species abundance.

Figure S5. Overview of the raw data for all studies (top panel names) and individual regressions (lower panel IDs) reporting correlations analyses between environment DNA copies L^{-1} and species biomass.

Figure S6. Visualisation of the distribution of the environmental DNA concentration measures for all studies (top panel names) and individual regressions (lower panel IDs) reporting correlations analyses between environment DNA copies L^{-1} and species biomass.

Figure S7. Verification of the fit of the Negative Binomial models without any prior treatment of the data by plotting the predicted versus observed values. Plots are created for a random subset of 50 regressions with panel labels indicating the individual regressions and the predictor variable given in between brackets (i.e. species abundance (A) or biomass (B)).

Figure S8. The error rate in function of the *p*-value of the reference model (i.e. model with negative binomial distribution without prior data filtering/summarizing) for regressions between eDNA concentration and species abundance.

Figure S9. The error rate in function of the *p*-value of the reference model (i.e. model with negative binomial distribution without prior data filtering/summarizing) for regressions between eDNA concentration and species biomass.

How to cite this article: Bylemans, J., Everts, T., Brys, R., & Duncan, R. P. (2025). From anarchy to clarity, data pre-processing and statistical choices influence quantitative environmental DNA (eDNA) analyses. *Methods in Ecology and Evolution*, *16*, 1322–1333. https://doi.org/10.1111/2041-210X.70064