NEWS AND VIEWS

REPLY

MOLECULAR ECOLOGY WILEY

Evidence-based strategies to navigate complexity in dietary DNA metabarcoding: A reply

Bethan L. Littleford-Colquhoun[1,2] | Violet I. Sackett[1,2] | Camille V. Tulloss[1,2] | Tyler R. Kartzinel[1,2]

[1]Department of Ecology, Evolution, and Organismal Biology, Brown University, Providence, Rhode Island, USA

[2]Institute at Brown for Environment and Society, Brown University, Providence, Rhode Island, USA

Correspondence
Bethan L. Littleford-Colquhoun and Tyler R. Kartzinel, Department of Ecology, Evolution, and Organismal Biology, Brown University, Providence, RI, USA.
Emails: bethan_littleford-colquhoun@brown.edu; tyler_kartzinel@brown.edu

Handling Editor: Loren Rieseberg

Abstract

It is clearly beneficial to eliminate low-abundance sequences that arise in error during dietary DNA metabarcoding studies, but to purge all low-abundance sequences is to risk eliminating real sequences and complicating ecological analyses. Our prior literature review noted that DNA sequence relative read abundance (RRA) thresholds can help ameliorate false-positive taxon occurrences, but that historical emphasis on this utility has fostered uncertainty about the associated risk of inflating the false-negative rate (Littleford-Colquhoun et al., 2022). To address this, we combined a simulation study and an empirical data set to both illustrate the issue and provide blueprints for simulation studies and sensitivity analyses that can help investigators avoid overcorrecting and thereby bolster confidence in ecological inferences. Awareness of both the costs and the benefits of abundance-filtering is needed because accurately characterizing dietary distributions can be critically important for understanding animal diets, nutrition and trophic networks. Highlighting the need to raise awareness, a critique of our paper emphasized the misleading notion that "false positive interactions between species can present fundamentally incorrect network structures in network ecology, whereas false negatives will provide a correct but incomplete version of the network" (Tercel & Cuff, 2022). Asserting that the reliability of results will be eroded by false positives but resilient to the omission of true positives is risky and runs counter to evidence. Unfortunately, abundance-filtering methods can introduce false negatives at higher rates than they eliminate false positives and thereby undermine the analysis of otherwise reliable sequencing data. Overcorrecting can qualitatively alter and ultimately undermine ecological interpretations.

1 | POINTS OF AGREEMENT

There is broad consensus that careful attention to avoiding contaminants and other spurious sequences that represent "false positives" is critical in DNA metabarcoding studies. Although imperfect, there have been many encouraging improvements to sampling strategies, sequencing technologies and computational methods that can help prevent and mitigate these errors. Readers may refer to numerous prior reviews for more detailed consideration of these developments and specific guidance about how to effectively design and analyse a dietary DNA metabarcoding study (Alberdi et al., 2018; Ando et al., 2018; Carlsen et al., 2012; Cirtwill & Hambäck, 2021; Creer et al., 2016; Deagle et al., 2010, 2019; Mata et al., 2019; McInnes et al., 2017; Zinger et al., 2019).

We found at least three points of agreement in the critique by Tercel & Cuff (2022) that we were disappointed to see framed as points of contention. Our original paper contended: (i) that false-positive sequences are indeed problematic for interpretations of

dietary DNA metabarcoding data; (ii) that practitioners should indeed strive to eliminate false positives in order to ensure valid interpretations; and (iii) that there may indeed be some research utility in employing RRA thresholds (for simplicity, "RRA thresholds" ≈ "minimum copy number thresholds" [MCNTs]). Point iii is particularly important to establish because Tercel and Cuff misquoted our paper, claiming that we discouraged the use of RRA thresholds to exclude low-abundance sequences and suggesting that we would rather rely on junky data that have been left "unchecked" due to "the abandonment" of RRA thresholds "in all studies" (Tercel & Cuff, 2022). But this is not so. We discussed at length how RRA thresholds may be useful in "at least" two common scenarios (Littleford-Colquhoun et al., 2022): (i) to exclude putatively false-positive sequences that occur at low relative abundance (perhaps as a bioinformatic convenience); and (ii) to focus analyses on relatively abundant taxa (perhaps because these taxa are of particular interest). These two scenarios are not mutually exclusive, and in practice they are not always differentiated, but we identified a specific need to focus on evidence concerning the usefulness of RRA thresholds in excluding false positives (a desired effect) vs. the risk of introducing false negatives (an insidious side effect).

2 | THE EVIDENCE CONCERNING MINIMUM READ THRESHOLDS

It is important to establish that there is evidence for both pros and cons when using RRA thresholds to mitigate the impacts of false-positive sequences in DNA metabarcoding studies:

1. The need to filter spurious DNA sequences that arise as contaminants and tag-jumps may be inevitable, but the extent of these sequences can vary based on the number of samples and study system (Dickie et al., 2018; Taberlet et al., 2018; Zinger et al., 2019), the library preparation method (Carøe & Bohmann, 2020; Clarke et al., 2014; Schnell et al., 2015) and the sequencing platform (Schirmer et al., 2015). Spurious sequences that become rampant can undermine the development of any appropriate bioinformatic-filtering strategy, but computational methods can otherwise help mitigate occurrences (Cirtwill & Hambäck, 2021).
2. Using RRA thresholds to exclude false-positive sequences works on the assumption that spurious sequences appear at lower relative abundances than a chosen threshold. It is widely acknowledged that it can be difficult to determine an appropriate threshold or verify its efficacy in practice (Ando et al., 2020; Deagle et al., 2019; Kelly et al., 2019).
3. Spurious sequences can occur at high relative abundances, and thus cannot always be eliminated effectively using this method (Alberdi et al., 2018; Ando et al., 2018, 2020).
4. Some low-abundance sequences may be real, and they may be important components of animal diets and trophic networks; there

is thus significant interest in striving to identify them (Littleford-Colquhoun et al., 2022; Pringle & Hutchinson, 2020).

5. All polymerase chain reaction (PCR)-based methods are prone to amplification bias, error and omission, but converting sequence counts to presence/absence data does not correct these issues: in many cases, converting sequence counts to presence/absence data can amplify the impacts of these errors (Deagle et al., 2019).
6. Using RRA thresholds to exclude sequences based exclusively on their relative abundance will modify dietary diversity distributions and thus the method has potential to distort ecological signals in the data (Littleford-Colquhoun et al., 2022).

Based on the evidence, we contended that the rarity of a sequence alone is insufficient to conclude that it is an error or otherwise unimportant trophic link (Littleford-Colquhoun et al., 2022). In other words, if some low-abundance sequences are true, then not all low-abundance sequences are false. This logical contention is not at odds with the evidence that many low-abundance sequences are indeed errors, but rather acknowledges that positively identifying a sequence as an error may require additional lines of evidence that can be obtained, for example, from algorithms designed to detect chimeras, PCR errors or tag jumps (Ando et al., 2020).

3 | ACCOUNTING FOR ERRORS ARISING FROM FALSE-POSITIVES AND FALSE-NEGATIVES

The central assertion by Tercel and Cuff is that the credibility of DNA metabarcoding studies will be eroded by including false positives but robust to the introduction of false negatives. Presumably, therefore, the benefits of abundance-filtering strategies that are commonly used should tend to outweigh the costs incurred by introducing false negatives when applied to representative data sets. Here, we present a simulation study that accounts for both false positives and false negatives in such a scenario. The results reveal why the assertions by Tercel and Cuff may often be wrong and help provide better intuition about the costs and benefits of abundance-filtering methods.

We will begin with a synopsis of the computer simulations presented in our original study (Littleford-Colquhoun et al., 2022). In those simulations, we produced in silico DNA metabarcoding data to enable comparisons of the effects that abundance-filtering would have on the dietary profiles of a hypothetical generalist and specialist whose true diets differed only in the probability with which they selected available foods (Figure 1a,b). Here, we randomly replaced some of the true sequence reads from those simulated diet profiles with a shared set of contaminants that represented 10 equally abundant taxa (i.e., false positives; Figure 1a,b). We compared scenarios where this contamination occurred at a relatively low level (10 contaminants each at 0.1% relative abundance; each representing 25 out of 25,000 reads) vs. a relatively high level (10 contaminants each at 1% relative abundance; each representing 250 out of 25,000
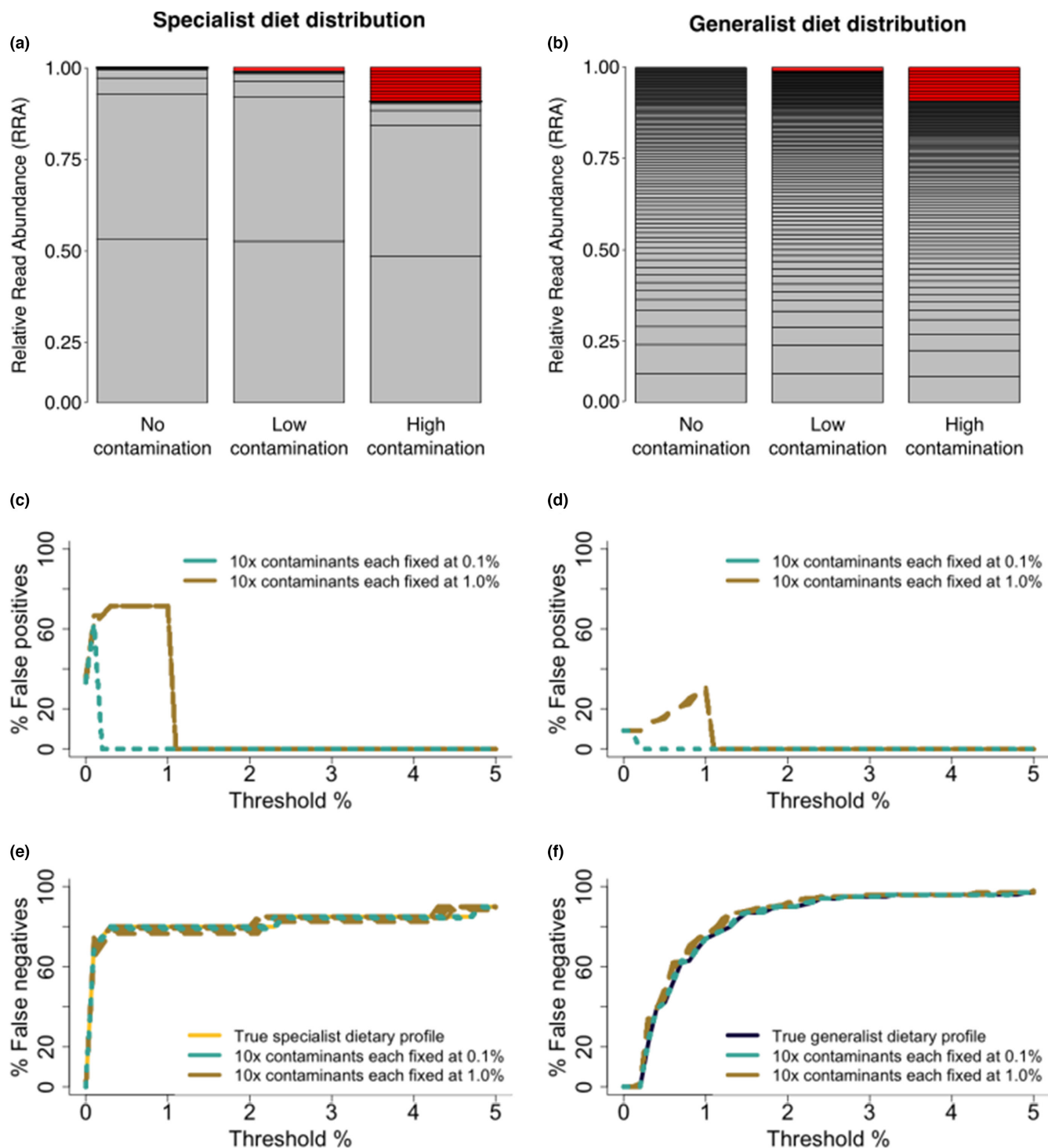
FIGURE 1 Simulated DNA metabarcoding data revealed how abundance-filtering methods eliminate both true and false sequence reads. Analyses began with simulated dietary DNA metabarcoding data for both a (a) specialist and (b) generalist consumer. A subset of true sequence reads was replaced either by relatively low levels of contamination (10 contaminants at 0.1% relative abundance each) or by relatively high levels of contamination (10 contaminants at 1% relative abundance each). The grey segments in the stacked barplots represent true-positive taxa and red segments represent contaminating taxa based on one of the 99 random iterations of the contamination procedure. In both data sets (c, d), the percentage false positives was exaggerated by using low-to-moderate percentage RRA thresholds. With high levels of contamination (brown lines), false positives comprised up to (c) 71% of taxa within the specialist's dietary profile and (d) 29% of taxa within a generalist's dietary profile. Further, high rates of false negatives were produced by low-to-moderate RRA thresholds (0.1%–1%), which led to losses of (e) 69%–79% of true taxa for the specialist and (f) 0%–76% of true taxa for the generalist. Because we conducted 99 random iterations of the low- and high-level contamination procedures, each plot contains a set of 99 nearly overlapping teal and brown lines (c–f)

reads). We repeated this random contamination procedure 99 times, supplanting a different set of true sequences in each iteration. If we attempted to ameliorate the impacts of these contaminants on our dietary data by removing low-abundance sequences using 0%–5% RRA thresholds, we would intuitively eliminate the contaminants when we select thresholds that exceeded their abundance (i.e., 0.1% or 1%; Figure 1c,d). Perhaps counterintuitively, however, we would find that "mild" RRA thresholds increased the percentage false-positive taxa in our results—even though we would not have added

any new false-positive sequences—because we would have eliminated more true sequences than false ones (Figure 1c–f). This result would have been especially pronounced for the specialist because of its highly skewed dietary diversity distribution.

Having shown that the percentage of false positives and percentage of false negatives are both sensitive to differences in the underlying (true) dietary distributions, we will now consider how their different sensitivities can influence measures of dietary overlap in a trophic network. Tercel and Cuff asserted that when spurious
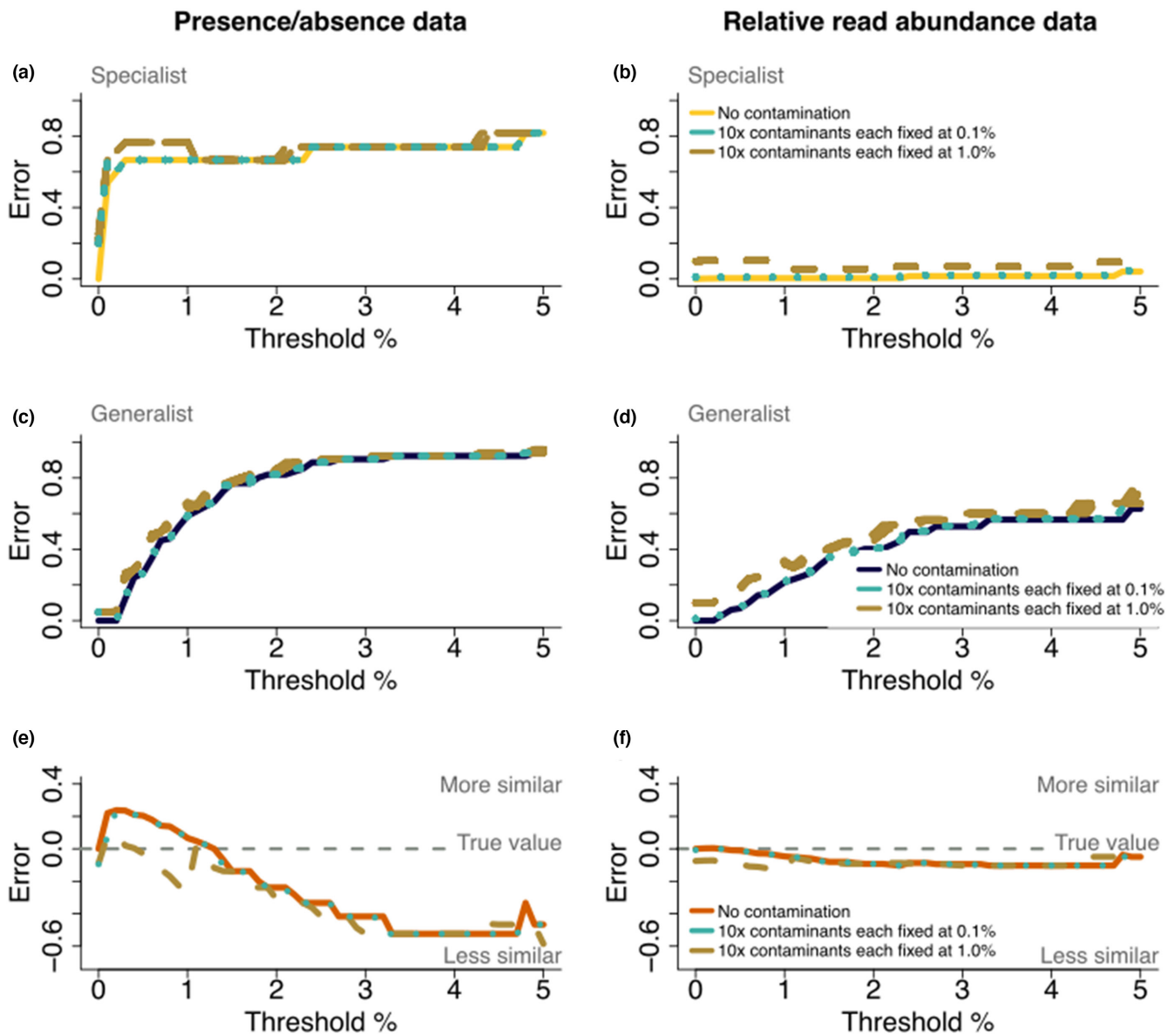
FIGURE 2 When compared to the true simulated dietary distributions, abundance-filtering can introduce greater levels of error than it eliminates. For simulated (a, b) specialists and (c, d) generalists, abundance-filtering using RRA thresholds between 0% and 5% introduced error regardless of the simulated level of contamination. We calculated error as Bray–Curtis dissimilarity between each true dietary profile and the dietary profile that resulted after contamination and/or abundance-filtering. We found (e, f) abundance-filtering had the potential to either increase or decrease the inferred level of similarity between the specialist and generalist compared to the true value. The true level of dietary overlap is indicated by the zero line, such that positive and negative values indicate scenarios where the inferred diets were more or less similar than expected, respectively. Greater levels of error were evident in analyses of presence/absence data (left) compared to RRA data (right) for the specialist diet profile (top), the generalist diet profile (middle), and the overlap between the specialist and generalist (bottom)

sequences are shared by a generalist and specialist, the inferred level of dietary overlap would increase to misleading levels but that this could be corrected by abundance-filtering. Tercel and Cuff did not base this assertion on evidence, but rather drew hypothetical ordinations and speculated (see their Figure 2). Contrary to their conjecture, we found that shared contaminants reaching levels of up to 10% cumulative RRA had almost imperceptible impacts on the resulting measures of dietary composition and overlap (Figure 2a–d). By contrast, the overall percentage of error in our inferred dietary compositions increased drastically after abundance-filtering—especially for the specialist and especially when converting the sequence reads to presence/absence data (Figure 2a,b). Abundance-filtering predominantly (but not exclusively) increased the inferred levels of dietary overlap (Figure 2e,f), again contradicting the assertions made by Tercel and Cuff. Especially when using presence/absence data, abundance-filtering had the potential to alter ecological interpretations: whereas the true Bray–Curtis dissimilarity between the diets of our simulated specialist and generalist was 0.67, abundance-filtering produced results that ranged from 0.08 to 0.91 and thus spanned nearly the entire range of potential values from 0 to 1 (Figure 2e).

4 | NAVIGATING COMPLEXITY

With the goal of producing accurate understanding in ecology, evidence-based approaches can help us navigate the complexity of dietary DNA data. In addition to methodological considerations in vitro, evidence-based approaches in silico may include the use of simulation studies and sensitivity analyses such as the ones that we presented here and in Littleford-Colquhoun et al. (2022). Clearly, there are both pros and cons to using RRA thresholds: they may be useful for focusing analyses on relatively abundant taxa that are of particular interest and/or that ameliorate concerns about the impacts of low-abundance errors, but they may also reshape dietary diversity distributions in ways that can be consequential for analyses that are sensitive to these distributions (e.g., Hill numbers, dissimilarity metrics). The evidence shows that reducing or eliminating our reliance on RRA thresholds may be possible, and investigators may gain confidence in doing so by focusing on robust experimental strategies (e.g., using libraries that enable detection of tag jumps; sequencing PCR replicates, extraction blanks and controls). There will also be benefit in continuing to investigate how abundance-filtering strategies may impact a wider variety of research scenarios, sources of error and bioinformatic pipelines than we have presented here. Important parameters to vary in future studies could include: (i) contamination levels (e.g., the number of contaminants, the levels of contamination, the constancy of contaminants across samples); (ii) experimental designs (e.g., sample size, read depth, amplification biases and any variation in these measures); or (iii) ecological scenarios (e.g., generalists and specialists with various degrees of true dietary overlap). There are undoubtedly other combinations of parameters to consider and controlled experiments to be completed.

Fortunately, however, we do not need to tailor our simulation studies to every empirical scenario in order to test critical assumptions, foster intuition and support evidence-based research outcomes.

AUTHOR CONTRIBUTIONS

B.L.L.C. and T.R.K. conceived and designed the simulations and analyses. B.L.L.C and T.R.K wrote the manuscript with contributions from all authors.

ORCID

Bethan L. Littleford-Colquhoun https://orcid.org/0000-0002-2594-0061
Violet I. Sackett https://orcid.org/0000-0002-2216-0764
Tyler R. Kartzinel https://orcid.org/0000-0002-8488-0580

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.

Ando, H., Fujii, C., Kawanabe, M., Ao, Y., Inoue, T., & Takenaka, A. (2018). Evaluation of plant contamination in metabarcoding diet analysis of a herbivore. *Scientific Reports, 8*(1), 1–10.

Ando, H., Mukai, H., Komura, T., Dewi, T., Ando, M., & Isagi, Y. (2020). Methodological trends and perspectives of animal dietary studies by noninvasive fecal DNA metabarcoding. *Environmental DNA, 2*(4), 391–406.

Carlsen, T., Aas, A. B., Lindner, D., Vrålstad, T., Schumacher, T., & Kauserud, H. (2012). Don't make a mista(g)ke: Is tag switching an overlooked source of error in amplicon pyrosequencing studies? *Fungal Ecology, 5*(6), 747–749.

Carøe, C., & Bohmann, K. (2020). Tagsteady: A metabarcoding library preparation protocol to avoid false assignment of sequences to samples. *Molecular Ecology Resources, 20*(6), 1620–1631.

Cirtwill, A. R., & Hambäck, P. (2021). Building food networks from molecular data: Bayesian or fixed-number thresholds for including links. *Basic and Applied Ecology, 50*, 67–76.

Clarke, L. J., Soubrier, J., Weyrich, L. S., & Cooper, A. (2014). Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. *Molecular Ecology Resources, 14*(6), 1160–1170.

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*(9), 1008–1018.

Deagle, B. E., Chiaradia, A., McInnes, J., & Jarman, S. N. (2010). Pyrosequencing faecal DNA to determine diet of little penguins: Is what goes in what comes out? *Conservation Genetics, 11*(5), 2039–2048.

Deagle, B. E., Thomas, A. C., McInnes, J. C., Clarke, L. J., Vesterinen, E. J., Clare, E. L., Kartzinel, T. R., & Eveson, J. P. (2019). Counting with DNA in metabarcoding studies: How should we convert sequence reads to dietary data? *Molecular Ecology, 28*(2), 391–406.

Dickie, I. A., Boyer, S., Buckley, H. L., Duncan, R. P., Gardner, P. P., Hogg, I. D., Holdaway, R. J., Lear, G., Makiola, A., Morales, S. E., Powell, J. R., & Weaver, L. (2018). Towards robust and repeatable sampling methods in eDNA-based studies. *Molecular Ecology Resources, 18*(5), 940–952.

Kelly, R. P., Shelton, A. O., & Gallego, R. (2019). Understanding PCR processes to draw meaningful conclusions from environmental DNA studies. *Scientific Reports*, *9*(1), 1–14.

Littleford-Colquhoun, B. L., Freeman, P. T., Sackett, V. I., Tulloss, C. V., McGarvey, L. M., Geremia, C., & Kartzinel, T. R. (2022). The precautionary principle and dietary DNA metabarcoding: Commonly used abundance thresholds change ecological interpretation. *Molecular Ecology*, *31*(6), 1615–1626.

Mata, V. A., Rebelo, H., Amorim, F., McCracken, G. F., Jarman, S., & Beja, P. (2019). How much is enough? Effects of technical and biological replication on metabarcoding dietary analysis. *Molecular Ecology*, *28*(2), 165–175.

McInnes, J. C., Alderman, R., Deagle, B. E., Lea, M. A., Raymond, B., & Jarman, S. N. (2017). Optimised scat collection protocols for dietary DNA metabarcoding in vertebrates. *Methods in Ecology and Evolution*, *8*(2), 192–202.

Pringle, R. M., & Hutchinson, M. C. (2020). Resolving food-web structure. *Annual Review of Ecology, Evolution, and Systematics*, *51*, 55–80.

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., & Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, *43*(6), e37.

Schnell, I. B., Bohmann, K., & Gilbert, M. T. P. (2015). Tag jumps illuminated–reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*, *15*(6), 1289–1303.

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). *Environmental DNA: For biodiversity research and monitoring*. Oxford University Press.

Tercel, M. P., & Cuff, J. P. (2022). The complex epistemological challenge of data curation in dietary metabarcoding: Comment on "the precautionary principle and dietary DNA metabarcoding: Commonly used abundance thresholds change ecological interpretation" by Littleford-Colquhoun et al.(2022). *Molecular Ecology*, 1–7.

Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., Chariton, A. A., Creer, S., Coissac, E., Deagle, B. E., De Barba, M., Dickie, I. A., Dumbrell, A. J., Ficetola, G. F., Fierer, N., Fumagalli, L., Gilbert, M. T. P., Jarman, S., Jumpponen, A., … Taberlet, P. (2019). DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. *Molecular Ecology*, *28*(8), 1857–1862.