Received: 13 July 2021 | Accepted: 18 November 2021

DOI: 10.1111/2041-210X.13780

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

An assessment of minimum sequence copy thresholds for identifying and reducing the prevalence of artefacts in dietary metabarcoding data

Lorna E. Drake[1] | Jordan P. Cuff[1,2] | Rebecca E. Young[1] | Angela Marchbank[1] | Elizabeth A. Chadwick[1] | William O. C. Symondson[1]

[1]School of Biosciences, Cardiff University, Cardiff, UK

[2]Rothamsted Insect Survey, Rothamsted Research, West Common, Harpenden, UK

Correspondence
Lorna E. Drake
Email: drake.lorna@gmail.com

Funding information
Biotechnology and Biological Sciences Research Council, Grant/Award Number: BB/M009122/1; Knowledge Economy Skills Scholarships (KESS 2); Welsh Government's European Social Fund (ESF)

Handling Editor: Andrew Mahon

Abstract

1. Metabarcoding provides a powerful tool for investigating biodiversity and trophic interactions, but the high sensitivity of this methodology makes it vulnerable to errors, resulting in artefacts in the final data. Metabarcoding studies thus often utilise minimum sequence copy thresholds (MSCTs) to remove artefacts that remain in datasets; however, there is no consensus on best practice for the use of MSCTs.
2. To mitigate erroneous reporting of results and inconsistencies, this study discusses and provides guidance for best-practice filtering of metabarcoding data for the ascertainment of conservative and accurate data. Several of the most commonly used MSCTs were applied to example datasets of Eurasian otter *Lutra lutra* and cereal crop spider (Araneae: Linyphiidae and Lycosidae) diets.
3. Changes in both the method and threshold value considerably affected the resultant data. Of the MSCTs tested, it was concluded that the optimal method for the examples given combined a sample-based threshold with removal of maximum taxon contamination, providing stringent filtering of artefacts while retaining target data.
4. Choice of threshold value differed between datasets due to variation in artefact abundance and sequencing depth, thus studies should employ controls (mock communities, negative controls with no DNA and unused MID tag combinations) to select threshold values appropriate for each individual study.

KEYWORDS
contamination, diet, false positives, high-throughput sequencing, trophic interactions

1 | INTRODUCTION

Metabarcoding provides a powerful tool for ecological studies of biodiversity and trophic interactions (Deiner et al., 2017; Taberlet et al., 2018). By combining high-throughput sequencing (HTS) with DNA barcoding, large volumes of high-resolution data can be generated from many samples simultaneously (Taberlet et al., 2018). As an accurate means of detecting and identifying not just common species, but also cryptic and rare species, metabarcoding has in many cases superseded traditional methods such as morphological analysis

of prey remains in gut contents and faeces, and direct observation (Bowser et al., 2013; Elbrecht et al., 2017; Roslin & Majaneva, 2016). The high sensitivity of metabarcoding does, however, render it vulnerable to error (Alberdi et al., 2018; Jusino et al., 2019), with differences in the treatment of samples producing distinct data, and thus conclusions, from the same samples (Alberdi et al., 2018, 2019). Better guidelines on best practice for data processing are thus required for metabarcoding studies as they become increasingly commonplace.

False positives, or 'artefacts', are detections of taxa in samples within which that taxon's DNA was not likely to be present at the point of collection (Darling et al., 2021). These can be introduced at any stage of the metabarcoding process, from sample collection through to bioinformatic analysis (Alberdi et al., 2019; Jusino et al., 2019). These can occur through contamination from environmental or laboratory sources (Czurda et al., 2016; Leonard et al., 2007; Siddall et al., 2009), tag-jumping and sample misassignment (transfer of sample-specific tags between samples; Schnell et al., 2015) or PCR and sequencing errors (chimeras or misidentified sequences; Bjørnsgaard Aas et al., 2017; Shin et al., 2014). Artefacts may also be produced through errors in reference databases (such as GenBank and BOLD; Valentini et al., 2009), resulting in sequences being assigned to the wrong taxon (Keskin et al., 2016; Rulik et al., 2017; Taberlet et al., 2018). Many of these artefacts can be limited through careful study design (e.g. pre- and post-PCR workstations; King et al., 2008; Murray et al., 2015) or the use of bioinformatics software to detect and remove erroneous sequences, the latter now possible through various different bioinformatic pipelines (e.g. UNOISE: Edgar, 2016; DADA2: Callahan et al., 2016). However, it is likely that some artefacts will remain regardless of precautionary steps taken (Nakagawa et al., 2018; Weyrich et al., 2019), potentially inflating species richness (Clare et al., 2016; Schnell et al., 2015; Zinger et al., 2019) and distorting data interpretation.

Minimum sequence copy thresholds (MSCTs) are one adaptable method commonly used to reduce the prevalence of artefacts (e.g. Hänfling et al., 2016). The choice of threshold must be carefully considered as it can considerably impact the data; low thresholds will be unsuccessful at removing artefacts, leaving false positives in the resultant data, whereas high thresholds may remove too much data, resulting in false negatives (Hänfling et al., 2016). This is especially true for dietary studies in which DNA of the focal consumer can be present at much higher concentrations than that of the food items (i.e. prey) and is undegraded, often resulting in its greater degree of amplification, depending on the PCR primers used. Other considerations include the amplification of non-target taxa (e.g. fungi, bacteria and symbionts/parasites in studies of carnivorous diet), or disproportionate representation of accidentally or secondarily consumed taxa, particularly problematic in omnivores (Tercel et al., 2021). The use of general primers that amplify the consumer will result in a lower proportion of each sample being assigned to food item DNA, whereas specific primers that avoid amplifying the consumer may reduce the amplification of some food items over others due to primer bias (Piñol et al., 2014). This variation increases the risk of target sequences being excluded if inappropriate filtering thresholds are selected.

Experimental controls are valuable components for empirically assigning MSCTs, as they provide a mechanism for estimating the proportion of artefacts within a dataset (Alberdi et al., 2019; Taberlet et al., 2018). Theoretically, negative controls [e.g. extraction blanks, PCR blanks and unused MID tag (molecular identifier tag) combinations] should contain no DNA, and positive controls (e.g. mock communities) should only contain DNA from selected taxa. This is, however, rarely the case, and these unexpected reads facilitate effective determination of optimal thresholds for data clean-up. Reads in negative controls may represent otherwise undetected contamination present in other samples (predominately identified using extraction and PCR blanks; Alberdi et al., 2019; Czurda et al., 2016; Leonard et al., 2007) or may occur due to tag-jumping or sequence mis-assignment (predominately identified using unused MID tag combinations; Schnell et al., 2015). Such artefacts are impossible to identify with certainty without negative controls since they are mostly assigned to taxa that occur in high read abundances across many samples and are thus indistinguishable from target DNA (Carew et al., 2018; Jensen et al., 2015; Sepulveda et al., 2020). Further artefacts are detected through the presence of positive control taxa in samples and sample taxa in positive controls, likely through tag-jumping, mis-assignment or sample cross-contamination. Unexpected reads in positive controls also allow low abundance artefacts from contaminants and PCR or sequencing errors, which may occur across samples too, to be identified. Control samples thus highlight artefact prevalent throughout unfiltered data, with those identified through negative controls otherwise increasing the frequency of occurrence of taxa, those identified through positive controls inflating sample diversity and both contributing to higher total read counts and, ultimately, false positives.

The application of MSCTs, and the use of controls for assessing thresholds, remains ambiguous and non-standardised, with many studies employing entirely distinct methodologies and thresholds (e.g. Gebremedhin et al., 2016; Guardiola et al., 2016; McInnes, Alderman, Lea, et al., 2017). Here we compared common practices for removing artefacts from metabarcoding data using example datasets of Eurasian otter *Lutra lutra* (Linnaeus, 1758) and cereal crop spider (Araneae: Linyphiidae and Lycosidae) dietary DNA. Samples were processed alongside experimental controls, allowing the practicality of controls for selecting filtering thresholds to be assessed. Through these examples, distinctions in the data outputs when using different techniques are highlighted, providing a basis for standardisation and outlining optimal solutions for the use of MSCTs on metabarcoding datasets. We hypothesised that: (a) data with MSCTs applied would still contain artefacts; (b) the extent of artefact removal would differ depending on the method of MSCT applied, with different MSCTs removing artefacts from different sources (e.g. artefacts in blanks vs. those in mock communities); (c) thresholds will require a fine context-dependent balance between low filtering thresholds which fail to remove many artefacts

and high thresholds which remove too much data, thus each dataset will require a unique threshold to optimally remove artefacts; (d) using multiple MSCTs simultaneously would remove more artefacts than MSCTs applied on their own; and (e) experimental controls would greatly benefit the choice of filtering method and threshold through identification of known target sequences and artefacts.

2 | MATERIALS AND METHODS

To assess existing artefact removal methodologies in use for DNA metabarcoding data, the methods used in 154 studies conducting metabarcoding on eukaryotic DNA for environmental monitoring or dietary analysis were tabulated (Table S1). Given the focus of this study on the clean-up of dietary metabarcoding data, which presents many unique challenges, each method was applied to four different datasets from two dietary studies carried out by the authors of this study: a dietary study of the Eurasian otter *Lutra lutra* (one COI and one 16S dataset) and a dietary study of cereal crop money spiders (two COI datasets).

2.1 | Example dataset 1: British otter diet

Faecal samples were collected during otter post-mortems by the Cardiff University Otter Project. Extracted faecal DNA was amplified using two metabarcoding primer pairs, new to this study, designed to amplify regions of the 16S rRNA and cytochrome *c* oxidase subunit I (COI) genes, each primer having 10 base pair molecular identifier tags (MID tags) to facilitate post-bioinformatic sample identification. Extraction and PCR negative controls, unused MID tag combinations, repeat samples and mock communities were included alongside the focal samples. Mock communities comprised standardised mixtures of DNA of marine species not previously detected in the diet of Eurasian otters (Table S2; Supporting Information 1). The resultant DNA libraries for each marker were sequenced on separate MiSeq V2 chips with 2 × 250 bp paired-end reads. Greater detail regarding sample processing, amplification and sequencing is provided in Supporting Information 2.

2.2 | Example dataset 2: Cereal crop spider diet

Money spiders (*Bathyphantes*, *Erigone*, *Microlinyphia* and *Tenuiphantes*; Araneae: Linyphiidae) and wolf spiders (*Pardosa*; Araneae: Lycosidae) were visually located on transects through barley fields. Gut DNA, extracted from the whole spider abdomen, was amplified using two COI metabarcoding primer pairs. One primer pair was selected for broad amplification of all invertebrates present, including the predator, and the other designed to exclude spider DNA to avoid predator amplification, each primer having 10 base pair MID tags to facilitate post-bioinformatic sample identification. Extraction and PCR negative controls, unused MID tag combinations, repeat samples and mock communities were included alongside the focal samples. Mock communities comprised standardised mixtures of DNA of exotic species not previously recorded in Britain (Table S2; Supporting Information 1). The resultant DNA libraries for each marker were sequenced on a MiSeq V3 chip with 2x300bp paired-end reads. Greater detail regarding sample processing, amplification and sequencing is provided in Supporting Information 3.

2.3 | Sequence analysis

Bioinformatic analyses were carried out using a custom pipeline. Sequences were first checked for truncation of MID tags by determining the proportion of sequence files containing exactly 10 bp before their respective primer. In all cases, the degree of truncation was deemed acceptable (≤10%).

FastP (Chen et al., 2018) was used to check the quality of reads, discard poor quality reads (<Q30, <125 bp long or too many unqualified bases, denoted by 'N') and merge read pairs from MiSeq files (R1 and R2). Merged reads were assigned a sample ID based on the MID tags associated with each primer using the 'trim.seqs' function of Mothur (Schloss et al., 2009); this also removed the MID tag and primer sequences from the reads. Using the files created by Mothur, reads were demultiplexed to obtain one file per sample ID. Read headers were modified for each file to include the sample ID and reads were then concatenated back into one file. Sequences were denoised (removal of PCR and sequencing errors), clustered into amplicon sequence variants (ASVs) and an ASV table was created using the commands 'fastx_uniques', 'unoise3' and 'otutab' in Usearch (v. 11; Edgar, 2016; Edgar, 2020). Taxonomic assignment for each ASV was obtained using the 'blastn' command in BLAST+, using a threshold of 97% similarity and e-value of 0.00001, against a downloaded database of DNA barcoding sequences submitted to online databases (e.g. GenBank; Camacho et al., 2009; National Center for Biotechnology Information, 2008).

Before assigning taxonomic identities to each ASV, BLAST results were filtered using the DPLYR package in R (version 3.6.0) using R Studio (version 1.2.1335; R Core Team, 2019). This was used to retain only accession codes with the top BIT score for each ASV. These data were then processed via MEGAN (version 6.12.3; Huson et al., 2016) to assign taxonomic names to each ASV. As erroneous entries on online databases can prevent species-level assignments, ASVs for which the top BLAST hit (i.e. top BIT score) was not resolved to species level were thus manually checked and assigned the most appropriate taxon. Taxonomic identity for each ASV was added to the ASV table produced by Usearch and reads were aggregated by taxonomic identity for each sample in R using the 'aggregate' function with a sum base function. ASVs were allocated taxonomic identities to overcome issues such as over-splitting of taxonomic groups, and to facilitate ecological interpretation of the data, particularly regarding identification of artefacts (e.g. identifying marine species in non-coastal otters).

2.4 | Minimum sequence copy thresholds

Seven of the most commonly used MSCTs (Table 1) were tested and their efficacy in cleaning all datasets compared. Filtering methods were enacted in excel using IF formulae.

If the read count (i.e. number of reads per sample per taxon) did not pass the designated threshold, then it was converted to zero (rather than subtracting the threshold, thus not altering the remaining read counts). For proportional methods (5–7, Table 1), a variety of thresholds, based around the proportional prevalence of different false positive instances, were tested to explore how choice of threshold can affect data output. The range of thresholds tested were chosen based on artefacts identified in control samples; we started with a low threshold and increased the value until most of the identifiable artefacts were removed. We also explored the effectiveness of using different MSCTs in pairwise combinations; this involved simultaneously applying 'Max Contamination' with each proportional threshold method (5–7), and 'Sample %' with 'Taxon %'.

Basic statistics were calculated to assess the effectiveness of each filtering method; total read count was used to assess the loss of reads across the whole dataset, the presence of singleton reads was used to assess the removal of PCR and sequencing errors, reads in blanks (negative controls and unused MID tags) were used to assess the levels of contamination and tag-jumping, and mock communities were used to assess the presence of false positives within samples. Artefacts could also be identified through taxa unexpectedly occurring in samples, such as taxa from dietary samples in controls, marine taxa associated with otters that did not have access to marine habitats, exotic taxa in British spider samples and mock community taxa in negative controls, unused MID tags or dietary samples.

To visualise the results of each method, tables of reads were converted into heat charts using the GGPLOT2 package (Wickham, 2016) in R. The frequency of occurrence for each taxon across all MID tag combinations was also calculated for each filtering method and used to create heat charts. Relative frequencies were calculated by dividing the frequency of occurrence by the total number of MID tag combinations; these values then underwent non-metric multidimensional scaling (NMDS) to visualise dissimilarity between the taxa present following application of each MSCT. This was conducted using the 'metaMDS' function in the VEGAN package (Oksanen et al., 2013) with two dimensions (stress <0.1) and a Bray–Curtis dissimilarity calculation (Bray & Curtis, 1957). Ellipses were created using the 'ordiellipse' function with the default 'sd' setting (standard deviation).

3 | RESULTS

3.1 | Sequencing output

Sequencing yielded 17.6, 13.7, 11.2 and 11.0 million paired-end reads, for the otter 16S and COI, and spider general and exclusion datasets, respectively, which decreased to 11.7, 7.9, 7.9 and 7.4 million, respectively, following bioinformatic analysis. Comparison of post-bioinformatic clean-up methods produced the same general patterns across the four datasets (otter 16S, otter COI, spider general COI and spider exclusion COI). We therefore used the simplest dataset (otter 16S) to graphically represent artefact removal (Figures 1 and 2; Table 2), with Supporting Information presenting the same data for the other datasets (otter COI, spider general COI and spider exclusion COI; Figures S1–S3; Tables S2–S4), as well as graphs depicting read counts per sample (Figures S4–S7) and the spatial distribution of otter faecal samples with marine taxa presences (Figures S8 and

TABLE 1 Seven post-bioinformatic filtering methods often applied to metabarcoding datasets, selected from those identified in metabarcoding studies (Table S1). The 'method name', herein used to refer to these methods, is given alongside the description (how the methods are executed) and the aim of each

Method name	Method description	Method aim
1. No filter	No OTU/ASV or sample filtering	No clean-up/maximum preservation of data
2. Singleton	Remove any read counts of one	Remove extremely low-frequency artefacts (e.g. sequencing artefacts)
3. <10	Remove any read counts that are less than 10	Remove low-frequency artefacts (e.g. sequencing artefacts, low-lying PCR contamination)
4. Max contamination	Remove any read counts within each OTU/ASV that are lower than the highest read count within a negative/blank control for that OTU/ASV	Remove contamination detected by the negative controls (e.g. extraction/PCR contamination, tag-jumping)
5. Total %	Remove any read counts less than a proportion of the total dataset read count for all reads	Remove low-frequency artefacts (e.g. sequencing artefacts, PCR contamination)
6. Sample %	Remove any read counts within a sample that are less than a proportion of the total sample read count for that sample	Remove sample contamination (e.g. environmental, extraction or PCR contamination)
7. Taxon %	Remove read counts with an abundance less than a proportion of the total OTU/ASV read count for that OTU/ASV	Remove cross-contamination (e.g. cross-contamination, tag-jumping)

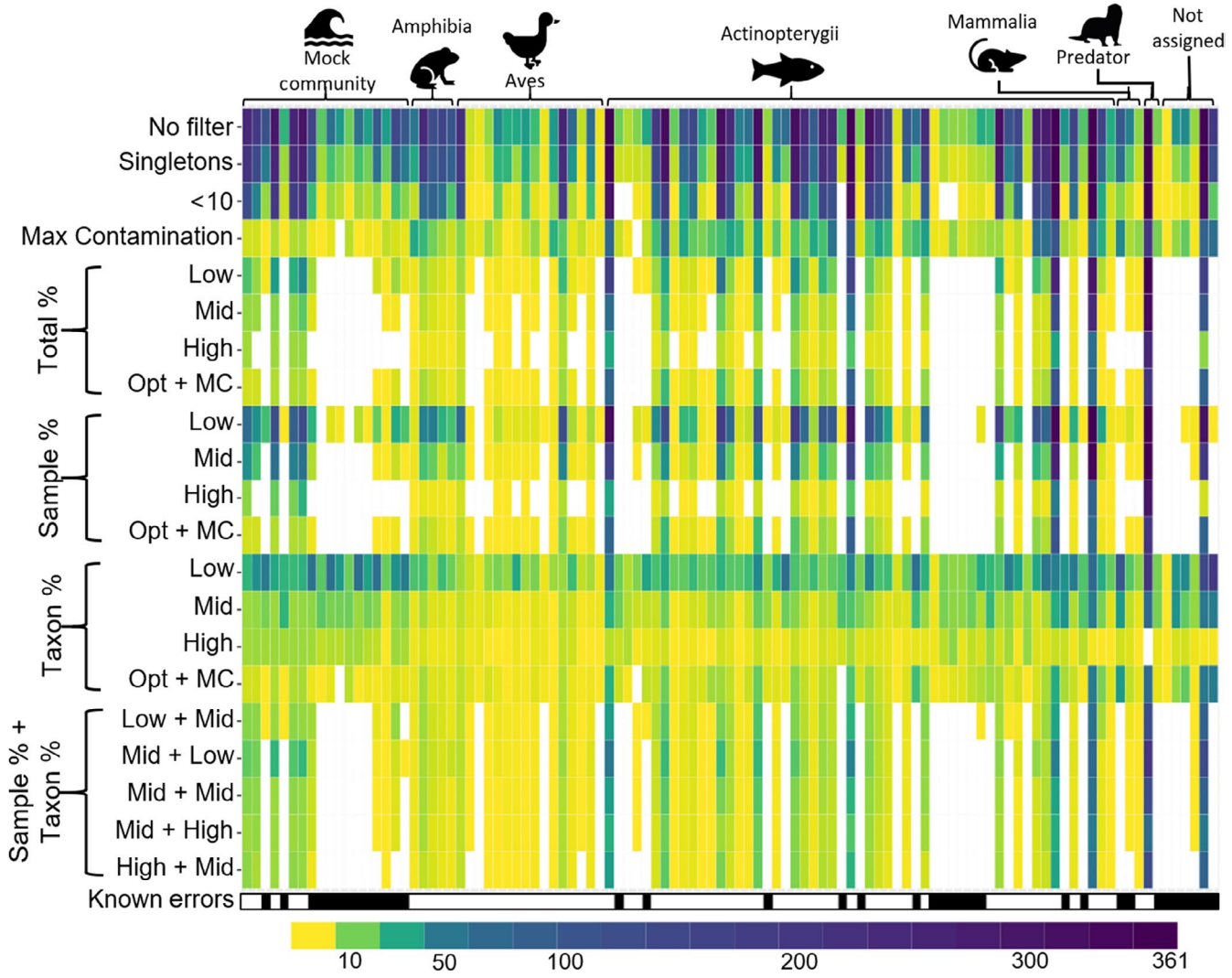


FIGURE 1 Otter diet 16S counts. The number of presences of each taxon is displayed for each method (low count = yellow, high count = purple) along with the number of taxa in each dataset following clean-up. Differences in common taxa, mock communities, predator amplification and erroneous taxa can be observed. 'Low', 'Mid' and 'High' depict the context-dependent range of values utilised for proportional thresholds ('Total %', 'Sample %' and 'Taxon %'), with 'Opt + MC' denoting the threshold deemed 'optimal' combined with the 'Max Contamination' method (for specific values, see Table S3). The same figure is available for three other datasets (otter COI, spider general COI and spider exclusion COI) in Figures S1–S3

S9). The effectiveness of each clean-up method across all datasets is also summarised in Table 3.

3.2 | No filter applied ('No Filter')

The highest read counts and occurrence of artefacts were observed in data with no MSCT applied. False positives in mock communities, reads in blanks, mock community taxa present in blanks and samples, taxa from samples occurring in control samples and, obviously, erroneously present taxa (e.g. marine taxa occurring in faecal samples from otters with no access to marine habitats) all occurred frequently across the datasets (Figure 1; Table 2). Artefacts appeared to be much more prevalent for taxa with high total read counts (e.g. mock community taxa, taxa commonly consumed by the predator and the focal predator itself). Many low abundance reads, including singletons, were also observed in the unfiltered data, possibly representing rare species but likely also sequencing errors.

3.3 | Remove singleton reads ('Singletons')

Removing singleton reads resulted in data very similar to that of unfiltered data in all cases, with only few artefacts removed (Figure 1; Table 2).

3.4 | Remove read counts <10 ('<10')

Removing reads with an abundance less than 10 reduced the occurrence of artefacts in blanks, mock communities and the presence

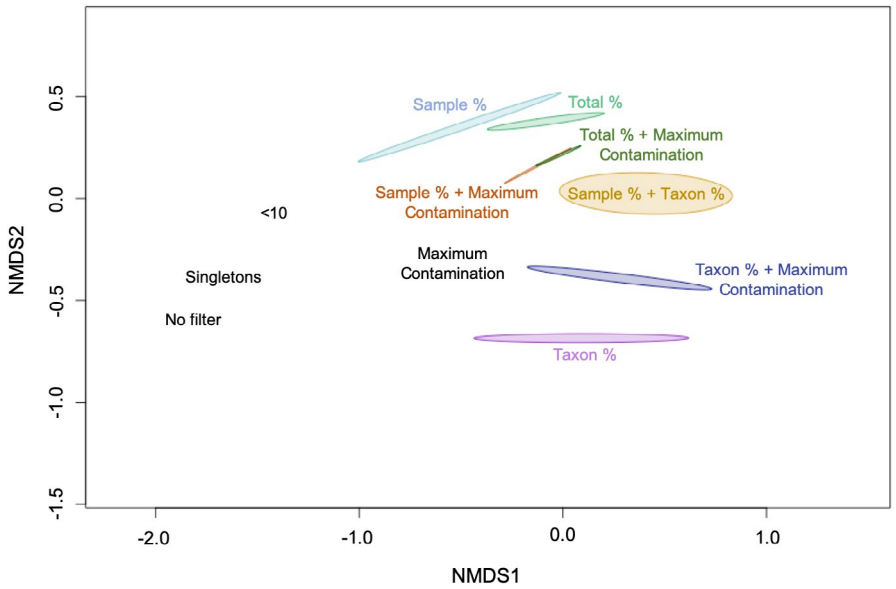


FIGURE 2 Otter 16S non-metric multidimensional scaling of relative frequency of occurrence for each taxon following the application of different minimum sequence copy thresholds, including different methods and thresholds where possible. Ellipse colours denote each method with None, Singletons, <10 and Maximum Contamination not having ellipses given the lack of modifiable threshold. The same figure is available for three other datasets (otter COI, spider general COI and spider exclusion COI) in Figures S7–S9

TABLE 2 Performance of different minimum sequence copy thresholds on otter 16S data. 'Low', 'Mid' and 'High' depict the context-dependent range of values utilised for proportional thresholds ('Total %', 'Sample %' and 'Taxon %'), with 'Opt + MC' denoting the threshold deemed 'optimal' combined with the 'Max Contamination' method (for specific values see Table S3). Expected presences of marine taxa (~) were defined by the number of Eurasian otters *Lutra lutra* displaying reads for each marine taxon that was located along the coast or near an estuary. Similar tables were produced for three other datasets (otter COI, spider general COI and spider exclusion COI) and are presented in Tables S2–S4

	Total		Singletons	Blanks	
Minimum sequence copy threshold	Summed read count	Taxa	Number of presences	Summed read count	Average read count
No filter	11,723,871	105	2,767	117,460	1,864
Singletons	11,721,104	105	0	117,032	1,858
<10	11,705,943	99	0	114,675	1,820
Maximum Contamination	10,938,496	102	63	0	0
Low Total %	11,534,535	71	0	96,869	1,538
Mid Total %	11,349,821	60	0	78,023	1,238
High Total %	10,733,900	46	0	35,916	570
Opt Total % + MC	10,874,148	63	0	0	0
Low Sample %	11,659,268	89	218	116,737	1,853
Mid Sample %	11,478,669	68	0	113,804	1,806
High Sample %	10,631,707	46	0	86,797	1,378
Opt Sample % + MC	10,875,890	65	0	0	0
Low Taxon %	11,031,736	105	742	45,985	730
Mid Taxon %	8,669,244	105	267	30,812	489
High Taxon %	3,660,086	104	25	30,645	486
Opt Taxon % + MC	8,569,029	102	0	0	0
Low Sample % + Mid Taxon %	10,187,214	72	2	30,851	490
Mid Sample % + Low Taxon %	10,959,369	68	0	44,471	706
Mid Sample % + Mid Taxon %	10,177,475	67	0	30,434	483
Mid Sample % + High Taxon %	8,647,191	67	0	29,865	474
High Sample % + Mid Taxon %	10,155,032	60	0	29,886	474

of mock community taxa in other samples. However, artefacts persisted in all controls and samples, producing data very similar to unfiltered data (Figure 1; Table 2).

3.5 | Remove maximum taxon contamination ('Max Contamination')

Removing reads less than or equal to the maximum read count in blanks per taxon removed no reads from some taxa and high values from others (otter 16S: minimum read removal = 0, maximum = 8,757 and average = 394; otter COI: minimum = 0, maximum = 23,413 and average = 117; spider amplification: minimum = 0, maximum = 5,851 and average = 136; and spider exclusion: minimum = 0, maximum = 10,764 and average = 155). Taxa experiencing high levels of read removal were often those with high total read counts. This cleared all reads from blanks (Tables 2 and 3), all mock community taxa from samples and taxa with high read abundances in samples from controls (Figure 1). False positives were still present in mock communities though (Figure 1), as were singleton reads. This method also cleared several erroneously located taxa, such as marine species associated with inland otters, but not all (Figure 1; Table 2).

3.6 | Proportion of total read count ('Total %')

This method removed artefacts present in blanks (Table 2), false positives in mock communities and erroneously located taxa (Figure 1; Table 2). Mock community taxa were cleared from blanks and samples to an extent, but some were still present even at high thresholds (Table 2). Taxa from dietary samples with high read abundances were not filtered efficaciously though, with many occurring in controls even at high thresholds. Thresholds tested across the

Mock communities			Marine taxa presences		
Average false positive read count	Average false positive presences	Presences in samples/blanks	*T. bubalis* (~1–3)	Pleuronectidae (~10–15)	*E. viperia* (~1)
3,121	38	295	166	324	37
3,113	30	259	84	291	28
3,066	19	198	38	194	7
314	4	0	36	14	7
2,498	5	38	11	36	1
2,018	2	11	1	14	1
220	0	2	1	10	0
115	0	0	1	14	1
3,290	10	126	40	172	3
2,113	2	51	6	38	1
0	0	21	1	8	1
96	0	0	3	14	1
376	13	21	36	27	22
163	8	2	19	12	5
99	5	1	1	7	1
96	2	0	19	12	5
15	0	2	12	13	1
140	0	19	4	16	1
124	0	2	4	13	1
124	0	2	4	12	1
0	0	2	2	13	1

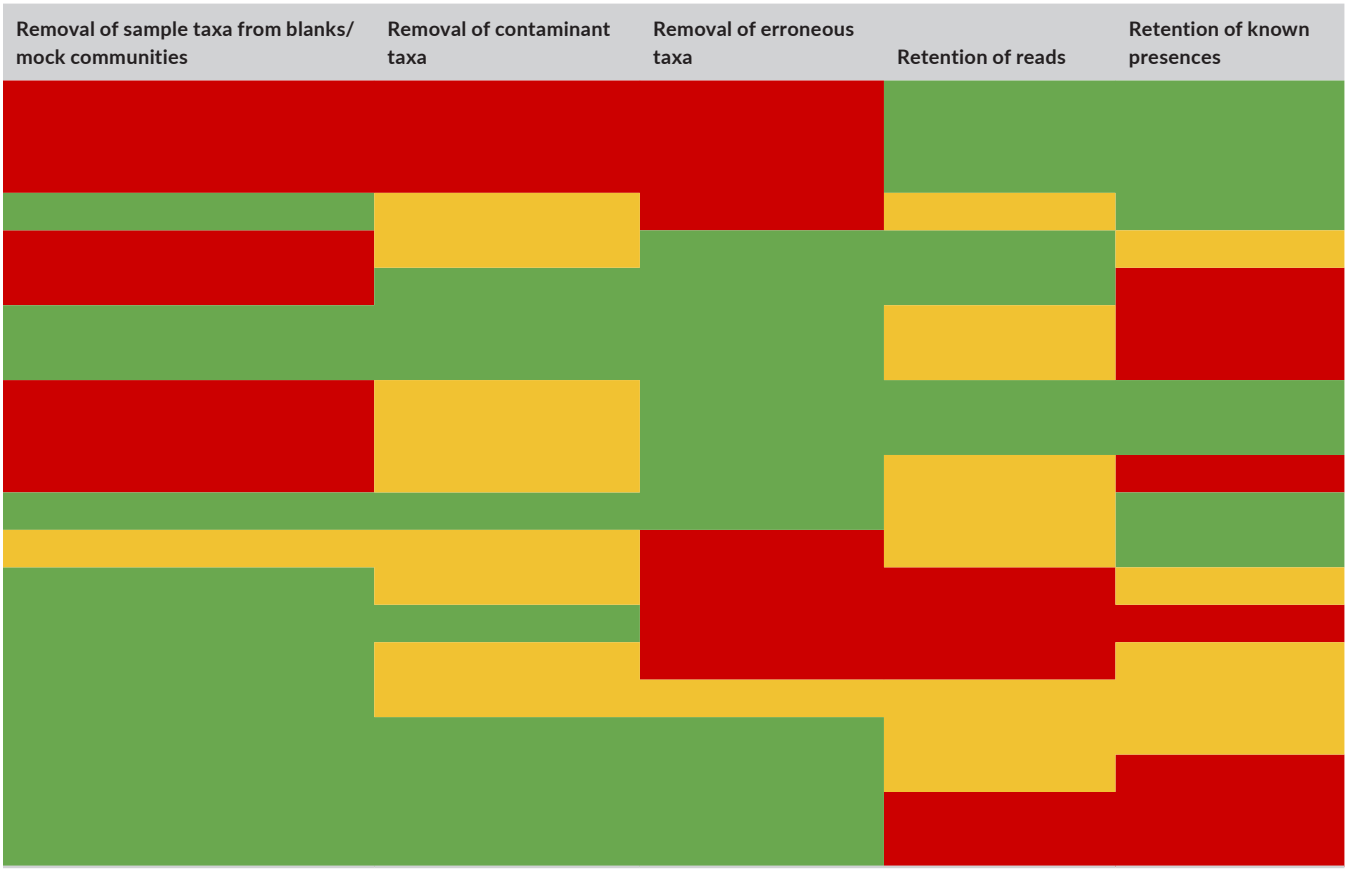
TABLE 3 Success of different filtering methods in achieving the key objectives of post-bioinformatic data clean-up. Green, orange and red denote positive, neutral and negative outcomes, respectively, determined by subjective inspection of the data output. The rating of the outcome is based on the relative ability of each method to remove false positives while preserving perceivably true positives. 'Low', 'Mid' and 'High' depict the value utilised for proportional thresholds ('Total %', 'Sample %' and 'Taxon %'), with 'Opt + MC' denoting the 'optimal' threshold combined with 'Max Contamination' methods (for specific values, see Table S3)

		Removal of singletons	Clearance of blanks	Removal of artefacts in mock communities	Removal of mock community taxa in blanks/ samples
	No filter				
	Singletons				
	<10				
	Max Contam				
Total %	Low				
	Mid				
	High				
	Opt + Max Contam				
Sample %	Low				
	Mid				
	High				
	Opt + Max Contam				
Taxon %	Low				
	Mid				
	High				
	Opt + Max Contam				
Sample % + Taxon %	Low + Mid				
	Mid + Low				
	Mid + Mid				
	Mid + High				
	High + Mid				

datasets ranged between removing reads that contributed to less than 0.0001% and 0.02% of the total read count. The lowest thresholds only filtered out a proportion of the artefacts, while the highest thresholds filtered out all false positives within mock communities and almost all reads in blanks (Figure 1; Table 2); however, the latter also removed target reads, shown by the loss of mock community taxa within mock communities. A lower threshold was therefore necessary to give a balance between false positives and false negatives. The optimal threshold was identified as 0.003%, 0.0008%, 0.0005% and 0.005% for otter 16S, otter COI, spider general amplification and spider exclusion, respectively, removing reads with abundances less than 79, 352, 39 and 236 respectively.

3.7 | Proportion of read count per sample ('Sample %')

This method removed false positives from mock communities (Figure 1) and erroneously located taxa (Table 2). Low abundance taxa (e.g. foreign taxa occurring through sequencing errors) were less prevalent (Figure 1), as were singletons. Taxa with high total read abundances (e.g. mock community taxa and common taxa in dietary samples) and reads present in blanks were only filtered to an extent (Figure 1; Table 2), resulting in artefacts from both being prevalent in filtered data regardless of the threshold utilised. This method removed fewer reads from samples with low total read counts, therefore these samples were more likely to still contain artefacts. Thresholds tested across the datasets included removing reads that contributed less than 0.01% to 8% of a sample's reads. The highest thresholds were required to remove all false positives from mock communities. A much higher threshold was required for some datasets (e.g. otter 16S) when they contained taxa with greater relative read counts. The high thresholds required to clear mock communities of false positives also removed many target reads (highlighted by the loss of mock community taxa), thus lower thresholds effectively balanced false positives and false negatives. The optimal threshold was identified as 1%, 0.3%, 0.38% and 1% for otter 16S, otter COI, spider general amplification and spider exclusion respectively. These thresholds removed reads to a varying degree (otter 16S: minimum read removal for a sample = 0, maximum = 8,757 and average = 394; otter COI: minimum = 0, maximum = 23,413 and average = 117; spider general amplification: minimum = 1, maximum = 240 and



average = 80; and spider exclusion: minimum = 1, maximum = 1,704 and average = 199).

3.8 | Proportion of read count per taxon ('Taxon %')

This method filtered out reads in blanks (Figure 1; Table 2), as well as artefacts from taxa with high read abundances, clearing most of these from the datasets when using sufficient thresholds. A large proportion of reads were removed using this method (Figure 1; Table 2), especially from taxa with high total read counts. Taxa with low read counts had fewer reads removed, resulting in these containing more artefacts, highlighted by the prevalence of singleton reads and taxa identified as PCR or sequencing errors (e.g. foreign taxa; Figure 1). This method proved insufficient at removing false positives from samples, with false positives prevalent in mock communities regardless of the threshold used, and erroneously located taxa were only removed when using a high threshold (Figure 1; Table 2). Thresholds tested included removing reads that contributed to <0.1%–3% of a taxon's reads. With low thresholds applied, many more artefacts were observed in blanks, but a threshold of 3% cleared most of these artefacts from the datasets in most cases. The highest thresholds removed a high proportion of reads, therefore lower thresholds were selected to give a balance between clearing out artefacts and not losing too many reads; this was 0.5%, 0.8%, 0.5% and 1% for otter 16S, otter COI, spider general amplification and spider exclusion respectively. These thresholds removed reads to different extents (otter 16S: minimum read removal for a taxa = 0, maximum = 26,039 and average = 553; otter COI: minimum = 0, maximum = 2,040 and average = 49; spider general amplification: minimum = 0, maximum = 306 and average = 28; and spider exclusion: minimum = 0, maximum = 1,286 and average = 76).

3.9 | Combining methods

Many of the thresholds tested for MSCTs based on read counts ('Total %', 'Sample %' and 'Taxon %') did not clear all artefacts, particularly regarding clearance of blanks. Proportional methods were thus also combined with 'Max Contamination' to overcome this issue. 'Sample %' thresholds were also combined with 'Taxon %' thresholds given

their complementary removal of artefacts. Combining methods removed more artefacts than using just one method. 'Total %' thresholds or 'Sample %' thresholds combined with 'Max Contamination' left very few artefacts in the data. These methods were highly complementary, with proportional thresholds clearing most false positives from mock communities and erroneously located taxa (Figure 1; Table 2), while the contamination threshold cleared reads in blanks and artefacts from taxa with high read counts (e.g. mock community taxa in non-mock community samples and faecal taxa in controls; Figure 1; Table 2). These combinations also cleared singletons and taxa suspected to be PCR or sequencing errors (Figure 1; Table 2). Combining these methods sometimes allowed lower thresholds to be used concurrently for optimal results, but in other cases did not change the thresholds required (otter 16S: optimal sample % = 0.5%, optimal total % = 0.002%; otter COI: optimal sample % = 0.2%, optimal total % = 0.0008%; spider general amplification: optimal sample % = 0.38%, optimal total % = 0.005%; and spider exclusion: optimal sample % = 0.39%, optimal total % = 0.005%).

'Taxon %' thresholds combined with 'Max Contamination' still contained many artefacts; all reads in blanks and singletons were removed, but false positives were still present in mock communities as were erroneously located taxa (although in lower abundances compared to either filter alone; Figure 1; Table 2). This is likely due to the similar action of both filters. Combining 'Taxon %' thresholds with 'Sample %' thresholds removed more artefacts and performed similar to MSCTs combining 'Sample %' thresholds with 'Max Contamination'. Combining these methods cleared the majority of reads from blanks, all singleton reads, artefacts from taxa with high read counts and most false positives in mock communities (Figure 1; Table 2); however, there were still artefacts present in the negative controls and erroneously located taxa were still present (Table 2). Combining these methods also removed many overall reads. The optimal combination of thresholds changed between datasets (otter 16S: sample = 0.5%, taxon = 0.3%; otter COI: sample = 0.2%, taxon = 0.3%; spider general amplification: sample = 0.5%, taxon = 0.3%; and spider exclusion: sample = 0.5%, taxon = 0.3%). Lowering the sample threshold introduced more false positives to the data, while increasing the threshold removed target reads. Lowering the taxon threshold retained more reads in blanks and artefacts from taxa with high total read counts, while increasing the taxon threshold greatly decreased the total read count, resulting in loss of target reads.

3.10 | NMDS analysis

The choice of MSCT method greatly affected the final composition of the data across all four datasets, as shown by NMDS (Figure 2; Figures S8–S10). The application of 'No Filter', 'Singletons' and '<10' MSCTs produced similar outcomes, with the '<10' threshold also appearing to elicit similar effects to MSCTs based on 'Total %' and 'Sample %'. 'Sample %' and 'Total %' thresholds were the most similar and gave results distinct from those of taxon MSCTs ('Taxon %' and 'Maximum Contamination'). By combining taxon MSCTs with either 'Sample %' or 'Total %' thresholds, an intermediate result was obtained. All combinations of taxon filters with 'Sample %' or 'Total %' thresholds performed similar to one another; however, with the otter 16S data those that combined 'Sample %' or 'Total %' with 'Maximum Contamination' were more dissimilar to taxon methods than combinations between 'Sample %' and 'Taxon %'.

4 | DISCUSSION

Here we have illustrated the efficacy of different filtering methods and thresholds for the removal of artefacts from dietary metabarcoding data, allowing us to identify an optimal method for artefact removal; utilising a threshold that removes a proportion of read counts per sample, combined with a threshold that removes reads less than the maximum read count identified in blanks per taxon ('Opt sample % + MC'; Table 3). For optimisation of thresholds, previous studies have disproportionately emphasised the importance of mock communities (e.g. Elbrecht & Leese, 2017; Jusino et al., 2019); however, since the biases affecting true unknown mixtures of DNA are almost impossible to experimentally replicate (Alberdi et al., 2018), data cannot be adequately filtered using only mock communities. By sequencing and analysing mock communities, blanks and samples together, it was possible to fully assess which filters and thresholds were optimal in cleaning metabarcoding data of this nature.

4.1 | Previous studies

Inspection of a number of relevant studies (Table S1) revealed a large proportion did not employ MSCTs and those which did often used entirely distinct methodologies and thresholds, with no optimal method apparent. Studies utilising one threshold across all read counts were commonly used, but often employed largely arbitrary thresholds (e.g. removal of reads with an abundance of <10) that did not consider the variation in artefact prevalence that can occur through differences in sequencing depth (De Barba et al., 2014; Elbrecht & Leese, 2015). While some studies circumvent this issue by using relative thresholds, each of these methods is likely to have removed artefacts to a different extent, introducing inconsistencies between datasets as a consequence. This study shows how using different MSCTs can drastically affect metabarcoding data, and in turn ecological interpretations of such data, therefore highlighting the need for more stringent removal of artefact across metabarcoding studies. Furthermore, the disparity in terminology and methodological descriptions between studies obviates confident inter-study comparison and undermines an overall requirement for scientific transparency. By comparing existing filtering methodologies, this study thus also provides effective descriptions for such methods which can be applied to mitigate this disparity. Importantly, other methods and strategies exist, such as the use of PCR triplicates and exclusion of any taxa which do not occur in at least two samples.

Similarly, Olds et al. (2016) required the detection of each taxon by multiple markers; this decreases the likelihood of laboratory-based contaminants persisting. The oftentimes severe taxonomic biases imposed by different PCR primers may, however, obscure the detection of some taxa with some markers, requiring thorough in silico and in vitro validations of the consistency of bias. This is particularly problematic for dietary studies which often intentionally employ taxonomic biases for the restriction of consumer amplification, as is the case for some of the datasets highlighted in this study. In such cases, the design of a single primer pair with ideal taxonomic biases can be difficult (e.g. Cuff et al., 2021; Lafage et al., 2019), rendering the replication of this infeasible for many studies. The use of different markers for the identification of taxonomically distinct compartments of the diet in such studies (e.g. plants and animals in the diet of omnivores; Tercel et al., 2021) would also confound this approach for dietary analyses, requiring fourfold multiplications of PCR costs and sequencing depths for already multiplexed experimental designs. This approach also insufficiently accounts for sequencing-based false positives such as those introduced between libraries sequenced together (Olds et al., 2016), whereas MSCTs can account for such false positives.

The use of alternative bioinformatic protocols may also have profound effects on the detection and mitigation of contamination, but this study is focused only on those methods which can be employed following typical bioinformatics processes. This study, for example, used an ASV-based approach (i.e. only identical sequences are clustered together), whereas traditional percentage identity-based clustering methods may generate different taxonomic diversities which could impact the perception of contamination (e.g. by obscuring some instances of contamination where these are taxonomically close to sample taxa). Similarly, the quality of reads used could affect the prevalence of often poorer quality sequences; these are typically removed bioinformatically, but some bioinformatic pipelines may better account for this than others. Truncation of reads, for example, could result in incorrect assignment of reads to samples, or poor quality sequences may be mis-assigned to taxa. Importantly, increased prevalence of these poor quality reads is particularly debilitating for arbitrary cut-offs (e.g. removal of singletons or read counts less than ten) which are not adjusted to account for differences in their prevalence between datasets or bioinformatics processes.

Many studies, particularly historically, have not employed the full set of measures that are presented as best practice in this manuscript (e.g. positive and negative controls). While this may preclude evidence-based application of some of the methods presented in this manuscript, it does not obviate filtering altogether. In the absence of negative controls (which is true of many published studies), the use of % taxon thresholds may be a viable means of reducing cross-contamination and tag-jumping. The use of a % sample threshold would arguably still be an effective means for the removal of low-level contaminants (e.g. sequencing errors, environmental contamination). Ideally these thresholds would require optimisation based on the primer pairs used, the consumer studied, the system in which that consumer exists and the generality of that consumer's foraging. Identifiable sources of contamination are critical in setting these thresholds (e.g. ecologically infeasible presences in samples, the presence of known laboratory contaminants), but this might be difficult to confidently determine in many cases. In such cases, a conservative threshold for a % sample and % taxon combined approach would likely provide an effective means for limiting contamination risk from multiple possible sources. We have demonstrated that thresholds around 0.5% are typically quite effective, but that thresholds around 1% will eliminate a very high proportion of contamination, albeit at the cost of false negatives. Without adequate controls, conservative data thresholds would be the only means of safeguarding against false positives, even if this was at the cost of a greater incidence of false negatives.

This study focuses on the use of MSCTs for the removal of contaminants in metabarcoding data, which are currently widely adopted. Other methods for the removal of contaminants are worth noting, but may not be feasible in some contexts. For example, Olds et al. (2016) built error distributions and calculate probabilities of detection to flag non-target taxa and false positives; this poses the additional benefit of somewhat mitigating the removal of true positives. Additional putative sources of contamination were, however, still identified from negative controls which would require alternative means of intervention (as in Evans et al., 2017 using a MSCT). Importantly, methods used for the detection of contaminants from environmental DNA samples (e.g. water, soil, air) can utilise a distinct set of principles given the disparate nature of sampling. For example, the taxa found in environmental samples will mostly be spatially constrained (i.e. present at that site), whereas dietary samples can include taxa spatially removed from the collection site, but only within the remit of that taxon's feasible trophic interactions (although instances such as secondary predation and scavenging, indistinguishable from predation events, will obscure the interpretations of ecologically feasible interactions). Care must thus be taken to ensure that translation of environmental DNA-focused approaches to dietary applications are appropriate and relevant.

4.2 | Identifying artefacts

Despite all appropriate precautionary steps being taken to reduce contamination (e.g. screening negative controls, pre- and post-PCR workstations), and bioinformatic programs used to remove erroneous sequences, artefacts were still observed in the unfiltered data. Such contamination is, however, largely unavoidable when using a method so broad-spectrum and sensitive (Alberdi et al., 2018; Jusino et al., 2019). Artefacts primarily manifested as unexpected reads in control samples, but also as erroneous taxa and mis-assigned reads. Erroneous taxa, usually existing in low read counts in the unfiltered data (De Barba et al., 2014; Ficetola et al., 2015), are, in this case, taxa produced through PCR or sequencing errors that are ecologically highly unlikely to appear in their respective samples (e.g. foreign species), thus rendering them easy to identify and eliminate. Mis-assigned reads were more difficult to identify, primarily

detected through mock community taxa occurring in samples and vice versa; however, some datasets also allow the detection of mis-assignment between samples through the presence of, for example, marine taxa in land-locked sites (Figures S1–S6). In such cases, reads were assumed to be derived from other samples through cross-contamination, tag-jumping or mis-assignment (Alberdi et al., 2019; Schnell et al., 2015). If easily identifiable, this can be fortuitous for threshold determination, but where samples share taxa that could theoretically co-occur, they will remain undetected.

The detection of artefacts is facilitated through the presence of unexpected reads in controls. Such reads in negative controls may occur due to low levels of contamination (e.g. from reagents or samples; Alberdi et al., 2019; Czurda et al., 2016; Leonard et al., 2007) that went undetected during screening of samples and may be present throughout only a few, or potentially all samples. Reads present in blanks may also occur due to tag-jumping or mis-assignment (Schnell et al., 2015), which are primarily identifiable through unused MID tag combinations. These artefacts are hard to detect without blanks because they are frequently assigned to taxa that legitimately occur in high read abundances across many samples (Carew et al., 2018; Jensen et al., 2015), such as mock community taxa and common taxa in samples (e.g. commonly consumed taxa or the consumer itself). Further artefacts were detected through the presence of mock community taxa in samples and common sample taxa in mock communities; these were concluded to be primarily due to tag-jumping or mis-assignment rather than sample cross-contamination because samples and mock community samples were processed separately. Unexpected reads in mock communities also allowed low abundance artefacts from contaminants and PCR or sequencing errors to be identified, which may have occurred across the samples. Control samples showed artefacts were prevalent throughout the unfiltered data, with those identified through blanks increasing the frequency of occurrence of taxa, those identified through mock communities inflating sample diversity and both contributing to higher total read counts and, ultimately, false positives.

The composition of mock communities is of great importance to the process of identifying artefacts. If the mock communities are comprised of species that may feasibly occur in the samples taken from the focal study system, the utility of those controls is reduced. Although the mock communities in this study comprised species considered highly unlikely to appear in the corresponding samples, distinct problems were encountered for all datasets. For the otter dietary analysis, the mock communities contained marine taxa unlikely to have been consumed by otters, yet high read counts were observed in the COI mock communities for brill *Scophthalmus rhombus*, a species known to be consumed by otters and not included in the mock community mixtures. The marine samples from which DNA was extracted were collected as part of a larger marine surveying initiative and, while care was taken by the practitioners responsible for the collection, cross-contamination between species was possible. Since this taxon could legitimately occur in both mock communities and samples, false presences are harder to confirm, but its marine origin meant that in areas lacking access to marine prey by otters, reads could still be identified as artefacts. The mock community mixtures used for the spider dietary analysis included exotic species from Round Island, Mauritius, collected as part of a separate study. These were selected for their absence in Britain and taxonomic relevance to the expected prey species (also small invertebrates). Given the poorly described entomological fauna of Round Island, Mauritius, the identities of a minority of these species were not resolved in the bioinformatics process, resulting in their designation as 'not assigned' and thus their exclusion from the filtering process alongside other unassigned taxa.

4.3 | Performance of MSCTs

Artefacts were removed to varying extents depending on the filtering method and threshold utilised. Basic MSCTs commonly used in the literature, such as removing singletons (e.g. Oliverio et al., 2018) or reads with an abundance less than 10 (e.g. Gebremedhin et al., 2016), removed very few artefacts. This will, however, vary with sequencing depth, with relatively greater depths increasing the likelihood of artefacts having more than 10 occurrences (De Barba et al., 2014; Elbrecht & Leese, 2015). MSCTs removing reads with an abundance below a proportion of the total read count performed better, reducing abundance of all detectable artefacts; however, applying one threshold across all read counts potentially indiscriminately removes target reads with low abundances and retains abundant artefacts. This bias can be overcome by using MSCTs based on sample read counts, as the read count will inevitably vary between samples despite best efforts to facilitate consistent sample read depths (Deagle et al., 2019). Sample MSCTs efficaciously removed artefacts from within samples, with lowered levels of cross-contamination and erroneous taxa, but did not clear artefacts from blanks, nor abundant taxa.

Minimum sequence copy thresholds that removed reads less than the maximum read count present in the blanks for each taxon ('Max Contamination'), and those which removed reads less than a given proportion of the total read count for that taxon ('Taxon %') removed artefacts from blanks and abundant taxa, but not mock communities or erroneous taxa. Of these two methods, removal of maximum taxon contamination was more suitable as it removed all artefacts from negative controls and taxa with high read counts without removing too many reads overall. To achieve the same result using thresholds based on taxon read counts resulted in much greater read losses, increasing the likelihood of false negatives. Proportional taxon thresholds also showed a strong bias towards removing reads from abundant taxa. While helping to remove artefacts produced through tag-jumping, this would potentially produce false negatives if taxa legitimately occurred in many samples. Comparing proportional taxon thresholds to others that cleared out similar amounts of artefacts revealed that proportional taxon thresholds produced the highest loss of reads, thus making this method more likely to lead to false negatives. Removal of maximum taxon contamination is logically superior given that the taxa for which the greatest number

of reads will be removed will be based on those that are verifiably contaminating the blanks. Care must, however, be taken to ensure that the protocols followed to produce the blanks are sufficiently stringent but not unnecessarily conservative (e.g. negative control volumes included being based on the average volume pooled per plate vs. the maximum volume pooled per plate), since this will cause this filtering method to produce many false negatives through overly strict removal of data.

4.4 | Combining MSCTs

Combining different MSCTs improved the performance of all filters, leading to a greater reduction in artefact presence. The weakest combination used proportional taxon thresholds with removal of maximum taxon contamination ('Taxon %' with 'Maximum Contamination'); these analogous methods removed artefacts in similar ways (i.e. removal based on reads present across taxa, rather than across samples), with neither sufficiently mitigating artefacts within samples. Artefacts persisting in blanks, following application of total read count thresholds, were removed by combining this method with removal of maximum taxon contamination; however, this combination may introduce biases by not accounting for read depth variation between samples, thus providing overly conservative filtering to some samples and insufficient filtering to others. Taxon-based thresholds were complementary to sample-based thresholds, with one removing artefacts identified through blanks and abundant taxa and the other removing artefacts within samples, including erroneous taxa. Combining sample-based thresholds with removal of maximum taxon contamination performed better than combinations with proportional taxon thresholds, as a greater proportion of artefacts were removed with a lower total read loss, reducing the likelihood of false negatives. A range of sample % threshold values were originally tested alongside maximum contamination to determine optimal thresholds. As can be observed with % sample thresholds alone, more conservative thresholds removed greater diversity from the resultant data, whereas lower thresholds, even when combined with the maximum contamination threshold, were insufficient at removing known contaminants. Due to its consistently improved performance over other MSCTs across all four metabarcoding datasets, we conclude that combining a sample-based threshold with removal of maximum taxon contamination is the optimal method for stringent filtering of metabarcoding data while retaining target data.

4.5 | Choosing an appropriate threshold

In metabarcoding studies, removal of false positives tends to be prioritised over false negatives due to the assumption that reads prove taxon presence while a lack of reads does not prove absence because false negatives can occur due to experimental biases (e.g. sampling or primer bias; Oehm et al., 2011; Piñol et al., 2015). A trade-off exists whereby the removal of false positives leads to an increase in false negatives (Alberdi et al., 2019; Zepeda-Mendoza et al., 2016), observed here when utilising high thresholds which removed many artefacts but also removed target reads, biasing results to taxa with high read abundance. Ultimately though, not all false positives are identifiable, meaning some artefacts may persist despite appropriate filtering removing all known artefacts. A balance can be achieved by which a high proportion of false positives are removed while retaining only very few false negatives that are easily disregarded (Clare et al., 2016; Hänfling et al., 2016; Zizka et al., 2019), thus better reflecting the true diversity within samples. The threshold at which this balance is achieved varies between studies depending on the sequencing depth and breadth of taxa. Appropriate thresholds should be chosen based on artefact removal from control samples. The aim of the study should, however, also be considered. Studies concerning commonly detected taxa can employ more stringent filters that remove more artefacts at the expense of losing rare taxa that may not be of interest anyway (e.g. studies of major sources of nutrition to a predator). However, studies concerning rare taxa should consider refining their thresholds to optimally remove artefacts while retaining the greatest amount of sequencing data (e.g. surveys of species richness).

In this study, we chose to assess the effectiveness of different thresholds using taxa read counts as well as occurrences (count data converted to presence or absence). Occurrence data are often assumed to be a conservative method of assessing metabarcoding data, as recovery biases (e.g. primer bias, starting amount of DNA) have a lower impact on such data (Deagle et al., 2019). Although occurrence data can inflate the importance of taxa that occur at low read counts (e.g. rare taxa or taxa consumed in small amounts; Deagle et al., 2019), and therefore also artefacts, particularly in studies with small sample sizes, we found it provided a simple and concise method for assessing artefact prevalence. Other methods, such as relative read abundance (RRA), may provide an alternative method for assessing abundance of artefacts and their impact on metabarcoding datasets by considering the proportion of reads each taxon contributes to a sample's total read count (this is analogous to the 'Sample %' MSCT). However, conversion of reads to RRA can produce misleading results due to biases such as differential digestion rates or primer amplifications (Alberdi et al., 2018; Clare, 2014; Elbrecht & Leese, 2015; Elbrecht et al., 2017; Piñol et al., 2014; Taberlet et al., 2012; Thomas et al., 2014), while the loss of read count data can potentially obscure interpretations of overall data loss. For these reasons, we chose not to convert read count data into RRA in this study but instead use raw read counts to assess the use of different MSCTs, thus allowing both artefact abundance and overall loss of reads to be assessed and directly compared. Future developments may make RRA a useful tool for artefact detection and removal though, allowing identification of artefacts that are having a proportionally large impact on metabarcoding data.

5 | CONCLUSIONS

Here we have shown that artefacts persist in metabarcoding data even following stringent laboratory and bioinformatic procedures. Although artefacts often occur in low abundances, they can create a disproportionate representation of biodiversity and produce misleading results, highlighting the need for read count filters. MSCTs removed artefacts to differing extents, but combining sample-based thresholds with removal of maximum taxon contamination provided an optimal outcome. While the optimal method was the same for all four datasets, thresholds applied differed due to variation in sequencing depth and differential taxon amplification. The choice of thresholds must thus depend on the individual study, taking into consideration the sequencing depth, breadth of taxa amplified, artefact abundance and the fundamental question under investigation. Control samples were crucial in assessing filters and selecting appropriate thresholds, providing a means for assessing the removal of artefacts and target reads. We recommend that future metabarcoding studies include mock communities and blanks, and, if possible, use taxa detected within samples that can be used to identify artefacts in the resultant metabarcoding data (e.g. marine taxa in inland samples) to facilitate identification of appropriate thresholds. Given the broad variation in MSCTs applied to metabarcoding studies, inconsistent results between these studies are inevitable. To mitigate erroneous reporting of results and inconsistencies, effective guidance for best-practice filtering of metabarcoding data for the ascertainment of conservative and accurate data should be followed.

ACKNOWLEDGEMENTS

The analysis of dietary data from otter samples was conducted by LED and funded by Knowledge Economy Skills Scholarship (KESS) and the Wildlife Trust of South and West Wales who partnered this project. KESS is a pan-Wales higher level skills initiative led by Bangor University on behalf of the HE sector in Wales and part funded by the Welsh Government's European Social Fund (ESF) convergence program for West Wales and the Valleys. The analysis of dietary data from spider samples was conducted by J.P.C. and funded by the Biotechnology and Biological Sciences Research Council through the South West Biosciences Doctoral Training Partnership (grant BB/M009122/1). Sample collection of otter faeces was conducted by Cardiff University Otter Project employees and placement students during post-mortems of otter carcasses. Samples for primer testing were provided by Manuel Nicolaus from CEFAS (marine species), Shaun Leonard at Wild Trout Trust (freshwater fish), Francois Edward and Monika Juergens at CEH (freshwater invertebrates) and Derek Gow Consultancy (mammals). The authors thank Robert Reader for allowing access to his farm for collection of terrestrial invertebrate samples, and Helen Hipperson and the NBAF team at the University of Sheffield for their initial training in bioinformatic analyses.

CONFLICT OF INTEREST

None declared.

AUTHORS' CONTRIBUTIONS

L.E.D., J.P.C., E.A.C. and W.O.C.S. conceived the ideas and oversaw the project; L.E.D. and J.P.C. generated the data; A.M. carried out the sequencing and advised on hypothetical implications for different data management strategies; L.E.D., J.P.C. and R.E.Y. analysed the data; L.E.D. led writing the manuscript. All authors commented on and contributed to the drafts and approved the final manuscript for publication.

LICENCES AND PERMITS

Samples for otter diet were collected from carcasses as part of the Cardiff University Otter Project National Monitoring Programme and thus did not require licences or permits to carry out the work in this study. Fieldwork carried out to collect spiders did not require licences or permits.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13780.

DATA AVAILABILITY STATEMENT

The data relevant to this publication, as well as the bioinformatics code used to generate the data and R code used to generate the results are publicly available via Dryad Digital Repository https://doi.org/10.5061/dryad.2jm63xsp4 (Drake et al., 2021).

ORCID

Lorna E. Drake https://orcid.org/0000-0003-0860-555X
Jordan P. Cuff https://orcid.org/0000-0002-0198-4940
Elizabeth A. Chadwick https://orcid.org/0000-0002-6662-6343

REFERENCES

Alberdi, A., Aizpurua, O., Bohmann, K., Gopalakrishnan, S., Lynggaard, C., Nielsen, M., & Gilbert, M. T. P. (2019). Promises and pitfalls of using high-throughput sequencing for diet analysis. *Molecular Ecology Resources*, *19*(2), 327–348. https://doi.org/10.1111/1755-0998.12960

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

Bjørnsgaard Aas, A., Davey, M. L., & Kauserud, H. (2017). ITS all right mama: Investigating the formation of chimeric sequences in the ITS2 region by DNA metabarcoding analyses of fungal mock communities of different complexities. *Molecular Ecology Resources*, *17*(4), 730–741. https://doi.org/10.1111/1755-0998.12622

Bowser, A. K., Diamond, A. W., & Addison, J. A. (2013). From puffins to plankton: A DNA-based analysis of a seabird food chain in the Northern Gulf of Maine. *PLoS One*, *8*(12), 1–16. https://doi.org/10.1371/journal.pone.0083152

Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, *27*, 325–349. https://doi.org/10.2307/1942268

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*, 581–583. https://doi.org/10.1038/nmeth.3869

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 1–9. https://doi.org/10.1186/1471-2105-10-421

Carew, M. E., Coleman, R. A., & Hoffmann, A. A. (2018). Can non-destructive DNA extraction of bulk invertebrate samples be used for metabarcoding? *PeerJ*, *2018*(6). https://doi.org/10.7717/peerj.4980

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Clare, E. L. (2014). Molecular detection of trophic interactions: Emerging trends, distinct advantages, significant considerations and conservation applications. *Evolutionary Applications*, *7*(9), 1144–1157.

Clare, E. L., Chain, F. J. J., Littlefair, J. E., & Cristescu, M. E. (2016). The effects of parameter choice on defining molecular operational taxonomic units and resulting ecological analyses of metabarcoding data. *Genome*, *59*(11), 981–990. https://doi.org/10.1139/gen-2015-0184

Cuff, J. P., Drake, L. E., Tercel, M. P., Stockdale, J. E., Orozco-terWengel, P., Bell, J. R., Vaughan, I. P., Müller, C. T., & Symondson, W. O. C. (2021). Money spider dietary choice in pre- and post-harvest cereal crops using metabarcoding. *Ecological Entomology*, *46*, 249–261. https://doi.org/10.1111/een.12957

Czurda, S., Smelik, S., Preuner-Stix, S., Nogueira, F., & Lion, T. (2016). Occurrence of fungal DNA contamination in PCR reagents: Approaches to control and decontamination. *Journal of Clinical Microbiology*, *54*(1), 148–152. https://doi.org/10.1128/JCM.02112-15

Darling, J. A., Jerde, C. L., & Sepulveda, A. J. (2021). What do you mean by false positive? *Environmental DNA*. In press.

De Barba, M., Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E., & Taberlet, P. (2014). DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: Application to omnivorous diet. *Molecular Ecology Resources*, *14*(2), 306–323

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. https://doi.org/10.1111/mec.14734

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E., & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*(21), 5872–5895. https://doi.org/10.1111/mec.14350

Drake, L. E., Cuff, J. P., Young, R. E., Marchbank, A., Chadwick, E. A., & Symondson, W. O. C. (2021). Data from: Post-bioinformatic methods to identify and reduce the prevalence of artefacts in metabarcoding data. *Dryad Digital Repository*, https://doi.org/10.5061/dryad.2jm63xsp4

Edgar, R. (2020). Usearch V11: Generating OTUs and ZOTUs. Retrieved from https://www.drive5.com/usearch/manual/pipe_otus.html

Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv*. https://doi.org/10.1101/003723

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), 1–16. https://doi.org/10.1371/journal.pone.0130324

Elbrecht, V., & Leese, F. (2017). Validation and development of COI metabarcoding primers for freshwater macroinvertebrate bioassessment. *Frontiers in Environmental Science*, *5*(April), 1–11. https://doi.org/10.3389/fenvs.2017.00011/full

Elbrecht, V., Vamos, E. E., Meissner, K., Aroviita, J., & Leese, F. (2017). Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. *Methods in Ecology and Evolution*, *8*(10), 1265–1275. https://doi.org/10.1111/2041-210X.12789

Evans, N. T., Li, Y., Renshaw, M. A., Olds, B. P., Deiner, K., Turner, C. R., Jerde, C. L., Lodge, D. M., Lamberti, G. A., & Pfrender, M. E. (2017). Fish community assessment with eDNA metabarcoding: Effects of sampling design and bioinformatic filtering. *Canadian Journal of Fisheries and Aquatic Sciences*, *74*(9), 1362–1374. https://doi.org/10.1139/cjfas-2016-0306

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

Gebremedhin, B., Flagstad, O., Bekele, A., Chala, D., Bakkestuen, V., Boessenkool, S., Popp, M., Gussarova, G., Schrøder-Nielsen, A., Nemomissa, S., Brochmann, C., Stenseth, N. C., & Epp, L. S. E. (2016). DNA metabarcoding reveals diet overlap between the endangered walia ibex and domestic goats - Implications for conservation. *PLoS One*, *11*(7). https://doi.org/10.1371/journal.pone.0159133

Guardiola, M., Wangensteen, O. S., Taberlet, P., Coissac, E., Uriz, M. J., & Turon, X. (2016). Spatio-temporal monitoring of deep-sea communities using metabarcoding of sediment DNA and RNA. *PeerJ*, *2016*(12), 1–31. https://doi.org/10.7717/peerj.2807

Hänfling, B., Handley, L. L., Read, D. S., Hahn, C., Li, J., Nichols, P., Blackman, R. C., Oliver, A., & Winfield, I. J. (2016). Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology*, *25*(13), 3101–3119. https://doi.org/10.1111/mec.13660

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H. J., & Tappu, R. (2016). MEGAN community edition - Interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Computational Biology*, *12*(6), 1–12. https://doi.org/10.1371/journal.pcbi.1004957

Jensen, R. H., Mollerup, S., Mourier, T., Hansen, T. A., Fridholm, H., Nielsen, L. P., Willerslev, E., Hansen, A. J., & Vinner, L. (2015). Target-dependent enrichment of virions determines the reduction of high-throughput sequencing in virus discovery. *PLoS One*, *10*(4), 1–18. https://doi.org/10.1371/journal.pone.0122636

Jusino, M. A., Banik, M. T., Palmer, J. M., Wray, A. K., Xiao, L., Pelton, E., Barber, J. R., Kawahara, A. Y., Gratton, C., Peery, M. Z., & Lindner, D. L. (2019). An improved method for utilizing high-throughput amplicon sequencing to determine the diets of insectivorous animals. *Molecular Ecology Resources*, *19*(1), 176–190. https://doi.org/10.1111/1755-0998.12951

Keskin, E., Unal, E. M., & Atar, H. H. (2016). Detection of rare and invasive freshwater fish species using eDNA pyrosequencing: Lake Iznik ichthyofauna revised. *Biochemical Systematics and Ecology*, *67*, 29–36. https://doi.org/10.1016/j.bse.2016.05.020

King, R. A., Read, D. S., Traugott, M., & Symondson, W. O. C. (2008). Molecular analysis of predation: A review of best practice for DNA-based approaches. *Molecular Ecology*, *17*(4), 947–963. https://doi.org/10.1111/j.1365-294X.2007.03613.x

Klymus, K. E., Richter, C. A., Thompson, N., & Hinck, J. E. (2017). Metabarcoding of environmental DNA samples to explore the use of uranium mine containment ponds as a water source for wildlife. *Diversity*, *9*(4), 1–18. https://doi.org/10.3390/d9040054

Lafage, D., Elbrecht, V., Cuff, J. P., Steinke, D., Hambäck, P. A., & Erlandsson, A. (2019). A new primer for metabarcoding of

spider gut contents. *Environmental DNA*, *2*(2), 234–243. https://doi.org/10.1002/edn3.62

Leonard, J. A., Shanks, O., Hofreiter, M., Kreuz, E., Hodges, L., Ream, W., Wayne, R. K., & Fleischer, R. C. (2007). Animal DNA in PCR reagents plagues ancient DNA research. *Journal of Archaeological Science*, *34*(9), 1361–1366. https://doi.org/10.1016/j.jas.2006.10.023

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. https://doi.org/10.1111/2041-210X.12677

McInnes, J. C., Alderman, R., Lea, M. A., Raymond, B., Deagle, B. E., Phillips, R. A., Stanworth, A., Thompson, D. R., Catry, P., Weimerskirch, H., Suazo, C. G., Gras, M., & Jarman, S. N. (2017). High occurrence of jellyfish predation by black-browed and Campbell albatross identified by DNA metabarcoding. *Molecular Ecology*, *26*(18), 4831–4845. https://doi.org/10.1111/mec.14245

Murray, D. C., Coghlan, M. L., & Bunce, M. (2015). From benchtop to desktop: Important considerations when designing amplicon sequencing workflows. *PLoS One*, *10*(4), 1–21. https://doi.org/10.1371/journal.pone.0124671

Nakagawa, H., Yamamoto, S., Sato, Y., Sado, T., Minamoto, T., & Miya, M. (2018). Comparing local- and regional-scale estimations of the diversity of stream fish using eDNA metabarcoding and conventional observation methods. *Freshwater Biology*, *63*(6), 569–580. https://doi.org/10.1111/fwb.13094

National Center for Biotechnology Information. (2008). *BLAST® Command line applications user manual*. www.ncbi.nlm.nih.gov/books/NBK279690/

Oehm, J., Juen, A., Nagiller, K., Neuhauser, S., & Traugott, M. (2011). Molecular scatology: How to improve prey DNA detection success in avian faeces? *Molecular Ecology Resources*, *11*(4), 620–628. https://doi.org/10.1111/j.1755-0998.2011.03001.x

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., & Wagner, H. (2013). *vegan: Community ecology package*.

Olds, B. P., Jerde, C. L., Renshaw, M. A., Li, Y., Evans, N. T., Turner, C. T., Deiner, K., Mahon, A. R., Brueseke, M. A., Shirey, P. D., Pfrender, M. E., Lodge, D. M., & Lamberti, G. A. (2016). Estimating species richness using environmental DNA. *Ecology and Evolution*, *6*(12), 4214–4226. https://doi.org/10.1002/ece3.2186

Oliverio, A. M., Gan, H., Wickings, K., & Fierer, N. (2018). A DNA metabarcoding approach to characterize soil arthropod communities. *Soil Biology and Biochemistry*, *125*(March), 37–43. https://doi.org/10.1016/j.soilbio.2018.06.026

Piñol, J., Mir, G., Gomez-Polo, P., & Agustí, N. (2015). Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular Ecology Resources*, *15*(4), 819–830. https://doi.org/10.1111/1755-0998.12355

Piñol, J., San Andrés, V., Clare, E. L., Mir, G., & Symondson, W. O. C. (2014). A pragmatic approach to the analysis of diets of generalist predators: The use of next-generation sequencing with no blocking probes. *Molecular Ecology Resources*, *14*(1), 18–26. https://doi.org/10.1111/1755-0998.12156

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Roslin, T., & Majaneva, S. (2016). The use of DNA barcodes in food web construction-terrestrial and aquatic ecologists unite! *Genome*, *59*(9), 603–628. https://doi.org/10.1139/gen-2015-0229

Rulik, B., Eberle, J., von der Mark, L., Thormann, J., Jung, M., Köhler, F., Apfel, W., Weigel, A., Kopetz, A., Köhler, J., Fritzlar, F., Hartmann, M., Hadulla, K., Schmidt, J., Hörren, T., Krebs, D., Theves, F., Eulitz, U., Skale, A., … Ahrens, D. (2017). Using taxonomic consistency with semi-automated data pre-processing for high quality DNA barcodes. *Methods in Ecology and Evolution*, *8*(12), 1878–1887. https://doi.org/10.1111/2041-210X.12824

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., & Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09

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. https://doi.org/10.1111/1755-0998.12402

Sepulveda, A. J., Hutchins, P. R., Forstchen, M., Mckeefry, M. N., & Swigris, A. M. (2020). The elephant in the lab (and field): Contamination in aquatic environmental DNA studies. *Frontiers in Ecology and Evolution*, *8*, 440. https://doi.org/10.3389/fevo.2020.609973

Shin, S., Lee, T. K., Han, J. M., & Park, J. (2014). Regional effects on chimera formation in 454 pyrosequenced amplicons from a mock community. *Journal of Microbiology*, *52*(7), 566–573. https://doi.org/10.1007/s12275-014-3485-6

Siddall, M. E., Fontanella, F. M., Watson, S. C., Kvist, S., & Erséus, C. (2009). Barcoding bamboozled by bacteria: Convergence to metazoan mitochondrial primer targets by marine microbes. *Systematic Biology*, *58*(4), 445–451. https://doi.org/10.1093/sysbio/syp033

Taberlet, P., Bonin, A., Coissac, E., Zinger, L., & Lucie, Z. (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.

Tercel, M. P. T. G., Symondson, W. O. C., & Cuff, J. P. (2021). The problem of omnivory: A synthesis on omnivory and DNA metabarcoding. *Molecular Ecology*, *30*(10), 2199–2206. https://doi.org/10.1111/mec.15903

Thomas, A. C., Jarman, S. N., Haman, K. H., Trites, A. W., & Deagle, B. E. (2014). Improving accuracy of DNA diet estimates using food tissue control materials and an evaluation of proxies for digestion bias. *Molecular Ecology*, *23*(15), 3706–3718. https://doi.org/10.1111/mec.12523

Valentini, A., Miquel, C., Nawaz, M. A., Bellemain, E., Coissac, E., Pompanon, F., Gielly, L., Cruaud, C., Nascetti, G., Wincker, P., Swenson, J. E., & Taberlet, P. (2009). New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: The trnL approach. *Molecular Ecology Resources*, *9*(1), 51–60. https://doi.org/10.1111/j.1755-0998.2008.02352.x

Weyrich, L. S., Farrer, A. G., Eisenhofer, R., Arriola, L. A., Young, J., Selway, C. A., Handsley-Davis, M., Adler, C. J., Breen, J., & Cooper, A. (2019). Laboratory contamination over time during low-biomass sample analysis. *Molecular Ecology Resources*, *19*(4), 982–996. https://doi.org/10.1111/1755-0998.13011

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.

Zepeda-Mendoza, M. L., Bohmann, K., Carmona Baez, A., & Gilbert, M. T. P. (2016). DAMe: A toolkit for the initial processing of datasets with PCR replicates of double-tagged amplicons for DNA metabarcoding analyses. *BMC Research Notes*, *9*(1), 1–13. https://doi.org/10.1186/s13104-016-2064-9

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. https://doi.org/10.1111/mec.15060

Zizka, V. M. A., Elbrecht, V., Macher, J. N., & Leese, F. (2019). Assessing the influence of sample tagging and library preparation on DNA metabarcoding. *Molecular Ecology Resources*, *19*(4), 893–899. https://doi.org/10.1111/1755-0998.13018

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Drake, L. E., Cuff, J. P., Young, R. E., Marchbank, A., Chadwick, E. A., & Symondson, W. O. C. (2021). An assessment of minimum sequence copy thresholds for identifying and reducing the prevalence of artefacts in dietary metabarcoding data. *Methods in Ecology and Evolution*, 00, 1–17. https://doi.org/10.1111/2041-210X.13780