

## RESEARCH ARTICLE

# Biodiversity Soup II: A bulk-sample metabarcoding pipeline emphasizing error reduction

Chunyan Yang<sup>1</sup> | Kristine Bohmann<sup>2</sup> | Xiaoyang Wang<sup>1</sup> | Wang Cai<sup>1</sup>  |  
Nathan Wales<sup>2,3</sup> | Zhaoli Ding<sup>4</sup> | Shyam Gopalakrishnan<sup>2</sup> | Douglas W. Yu<sup>1,5,6</sup> 

<sup>1</sup>State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China; <sup>2</sup>Section for Evolutionary Genomics, Globe Institute, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark; <sup>3</sup>Department of Archaeology, University of York, York, UK; <sup>4</sup>Biodiversity Genomics Center, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China; <sup>5</sup>School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, UK and <sup>6</sup>Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China

## Correspondence

Douglas W. Yu  
Email: douglas.yu@uea.ac.uk

## Funding information

Chinese Academy of Sciences, Grant/Award Number: XDA20050202; National Natural Science Foundation of China, Grant/Award Number: 41661144002, 31670536, 31400470 and 31500305; Key Research Program of Frontier Sciences, CAS, Grant/Award Number: QYZDY-SSW-SMC024; Bureau of International Cooperation, Grant/Award Number: GJHZ1754; Ministry of Science and Technology of China, Grant/Award Number: 2012FY110800; State Key Laboratory of Genetic Resources and Evolution, Grant/Award Number: GREKF18-04; University of East Anglia; University of Chinese Academy of Sciences; Leverhulme Trust Research Fellowship; Danish Council for Independent Research, Grant/Award Number: DFF-5051-00140

Handling Editor: Nigel Yoccoz

## Abstract

1. Despite widespread recognition of its great promise to aid decision-making in environmental management, the applied use of metabarcoding requires improvements to reduce the multiple errors that arise during PCR amplification, sequencing and library generation. We present a co-designed wet-lab and bioinformatic workflow for metabarcoding bulk samples that removes both false-positive (tag jumps, chimeras, erroneous sequences) and false-negative ('dropout') errors. However, we find that it is not possible to recover relative-abundance information from amplicon data, due to persistent species-specific biases.
2. To present and validate our workflow, we created eight mock arthropod soups, all containing the same 248 arthropod morphospecies but differing in absolute and relative DNA concentrations, and we ran them under five different PCR conditions. Our pipeline includes qPCR-optimized PCR annealing temperature and cycle number, twin-tagging, multiple independent PCR replicates per sample, and negative and positive controls. In the bioinformatic portion, we introduce *Begum*, which is a new version of *DAME* (Zepeda-Mendoza et al., 2016. *BMC Res. Notes* 9:255) that ignores heterogeneity spacers, allows primer mismatches when demultiplexing samples and is more efficient. Like *DAME*, *Begum* removes tag-jumped reads and removes sequence errors by keeping only sequences that appear in more than one PCR above a minimum copy number per PCR. The filtering thresholds are user-configurable.
3. We report that OTU dropout frequency and taxonomic amplification bias are both reduced by using a PCR annealing temperature and cycle number on the low ends of the ranges currently used for the Leray-FolDegenRev primers. We also report that tag jumps and erroneous sequences can be nearly eliminated with *Begum* filtering, at the cost of only a small rise in dropouts. We replicate published findings that uneven size distribution of input biomasses leads to greater dropout

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society

frequency and that OTU size is a poor predictor of species input biomass. Finally, we find no evidence for 'tag-biased' PCR amplification.

4. To aid learning, reproducibility, and the design and testing of alternative metabarcoding pipelines, we provide our Illumina and input-species sequence datasets, scripts, a spreadsheet for designing primer tags and a tutorial.

#### KEYWORDS

bulk-sample DNA metabarcoding, environmental DNA, environmental impact assessment, false negatives, false positives, Illumina high-throughput sequencing, tag bias

## 1 | INTRODUCTION

DNA metabarcoding enables rapid and cost-effective identification of taxa within biological samples, combining amplicon sequencing with DNA taxonomy to identify multiple taxa in bulk samples of whole organisms and in environmental samples such as water, soil and faeces (Deiner et al., 2017; Taberlet, Coissac, Hajibabaei, et al., 2012; Taberlet, Coissac, Pompanon, et al., 2012). Following initial proof-of-concept studies (Fonseca et al., 2010; Hajibabaei et al., 2011; Ji et al., 2013; Thomsen et al., 2012; Yoccoz, 2012; Yu et al., 2012) has come a flood of basic and applied research and even new journals and commercial service providers (Alberdi et al., 2018; Callahan et al., 2016; Murray et al., 2015; Zepeda-Mendoza et al., 2016; Zizka et al., 2019). Two recent and magnificent surveys are Taberlet et al. (2018) and Piper et al. (2019). The big advantage of metabarcoding as a biodiversity survey method is that with appropriate controls and filtering, metabarcoding can estimate species compositions and richnesses from samples in which taxa are not well-characterized a priori or reference databases are incomplete or lacking. However, this is also a disadvantage because we must first spend effort to design reliable and efficient metabarcoding pipelines.

Practitioners are thus confronted by multiple protocols that have been proposed to avoid and mitigate the many sources of error that can arise in metabarcoding (Table 1). These errors can result in false negatives (failures to detect target taxa that are in the sample, 'drop-outs'), false positives (false detections of taxa), poor quantification of biomasses and/or incorrect assignment of taxonomies, which also results in paired false negatives and positives. As a result, despite recognition of its high promise for environmental management (Abrams et al., 2019; Bush et al., 2019; Cordier et al., 2020; Hering et al., 2018; Ji et al., 2013; Piper et al., 2019), the applied use of metabarcoding is still getting started. A comprehensive understanding of costs, the factors that govern the efficiency of target taxon recovery, the degree to which quantitative information can be extracted and the efficacy of methods to minimize error is needed to optimize metabarcoding pipelines (Axtner et al., 2019; Hering et al., 2018; Piper et al., 2019).

Here we consider one of the two main sample types used in metabarcoding: bulk-sample DNA (the other type being environmental DNA, Bohmann et al., 2014). Bulk-sample metabarcoding, such as mass-collected invertebrates, is being studied as a way to generate multi-taxon indicators of environmental quality

(Hering et al., 2018; Lanzén et al., 2016), to track ecological restoration (Barsoum et al., 2019; Cole et al., 2016; Fernandes et al., 2018; Wang et al., 2019), to detect pest species (Piper et al., 2019) and to understand the drivers of species-diversity gradients (Zhang et al., 2016).

We present a co-designed wet-lab and bioinformatic pipeline that uses qPCR-optimized PCR conditions, three independent PCR replicates per sample, twin-tagging, and negative and positive controls to: (a) remove sequence-to-sample misassignment due to tag-jumping, (b) reduce dropout frequency and taxonomic bias in amplification and (c) reduce false-positive frequency.

As part of the pipeline, we introduce a new version of the *DAME* software package (Zepeda-Mendoza et al., 2016), renamed *Begum* (Hindi for 'lady'), to demultiplex samples, remove tag-jumped sequences and filter out erroneous sequences (Alberdi et al., 2018). Regarding the latter, the *DAME/Begum* logic is that true sequences are more likely to appear in multiple, independent PCR replicates and in multiple copies than are erroneous sequences (indels, substitutions, chimeras). Thus, erroneous sequences can be filtered out by keeping only sequences that appear in more than one (or a low number of) PCR replicate(s) at above some minimum copy number per PCR, albeit at a cost of also filtering out some true sequences. *Begum* improves on *DAME* by ignoring heterogeneity spacers in the amplicon, allowing primer mismatches during demultiplexing, and by being more efficient. We note that this logic is less applicable to species represented by trace DNA, such as in water samples, where low concentrations of DNA template are more likely to result in a species truly appearing in only one PCR (Harper et al., 2018; Piaggio et al., 2014).

To test our pipeline, we created eight 'mock' arthropod soups, each consisting of the DNA of the same 248 arthropod taxa mixed together in the laboratory and differing in absolute and relative DNA concentrations, ran them under five different PCR conditions and used *Begum* to filter out erroneous sequences (Figure 1). We then quantified the efficiency of species recovery from bulk arthropod samples, as measured by four metrics:

1. the frequency of false-negative OTUs ('dropouts', i.e. unrecovered input species),
2. the frequency of false-positive OTUs (sequences not from the input species),
3. the recovery of species relative-abundance information (i.e. does OTU size [number of reads] predict input genomic DNA amount per species?) and

**TABLE 1** Four classes of metabarcoding errors and their causes. Not included are software bugs, general laboratory and field errors like mislabelling, sampling biases or inadequate sequencing depth

Main errors	Possible causes	References
False positives (OTU sequences in the final dataset that are not from target taxa)	Sample contamination in the field or laboratory	Champlot et al. (2010) and De Barba et al. (2014)
	PCR errors (substitutions, indels, chimeric sequences)	Deagle et al. (2018)
	Sequencing errors	Eren et al. (2013)
	Incorrect assignment of sequences to samples ('tag jumping')	Esling et al. (2015) and Schnell et al. (2015)
	Intraspecific variability across the marker leading to multiple OTUs from the same species	Virgilio et al. (2010) and Bohmann et al. (2018)
	Incorrect classification of an OTU as a prey item when it was in fact consumed by another prey species in the same gut	Hardy et al. (2017)
	Numts (nuclear copies of mitochondrial genes)	Bensasson et al. (2001)
False negatives ('Dropouts', failure to detect target taxa that are in the sample)	Fragmented DNA leading to failure to PCR amplify	Ziesemer et al. (2015)
	Primer bias (interspecific variability across the marker)	Clarke et al. (2014), Piñol et al. (2015) and Alberdi et al. (2018)
	PCR inhibition	Murray et al. (2015)
	PCR stochasticity	Piñol et al. (2015)
	PCR runaway (loss of diversity caused by some sequences outcompeting others during PCR)	Polz and Cavanaugh (1998)
	Predator and collector DNA dominating the PCR product and causing target taxa (e.g. diet items) to fail to amplify	Deagle et al. (2009) and Shehzad et al. (2012)
	Too many PCR cycles and/or too high annealing temperature, leading to the loss of sequences with low starting DNA	Piñol et al. (2015)
Poor quantification of target species abundances or biomasses	PCR stochasticity	Deagle et al. (2014)
	Primer bias	Piñol et al. (2019)
	Polymerase bias	Nichols et al. (2018)
	PCR inhibition	Murray et al. (2015)
	Too many cycles in the metabarcoding PCR	
Taxonomic assignment errors (a class of error that can result in false positives or negatives, depending on its nature)	Intraspecific variability across the marker leading to multiple OTUs with different taxonomic assignments	Clarke et al. (2014)
	Incomplete reference databases	

4. taxonomic bias (are some taxa more or less likely to be recovered?).

Highest efficiency is achieved by recovering *all* and *only* the input species, in their original frequencies. We show that with *Begum* filtering, metabarcoding efficiency is highest with a PCR cycle number and annealing temperature at the low ends of the ranges currently used in metabarcoding studies, that *Begum* filtering nearly eliminates false-positive OTUs, at the cost of only a small absolute rise in false-negative frequency, that greater species evenness and higher concentrations reduce dropouts (replicating Elbrecht et al., 2017) and that OTU sizes are not reliable estimators of species relative abundances. We also find no evidence for 'tag bias', which is the hypothesis that the sample-identifying nucleotide sequences attached to PCR primers might promote annealing to some template-DNA sequences over others, exacerbating taxonomic bias in PCR (e.g. Berry et al., 2011; O'Donnell et al., 2016). All these results have important implications for using metabarcoding as a biomonitoring tool.

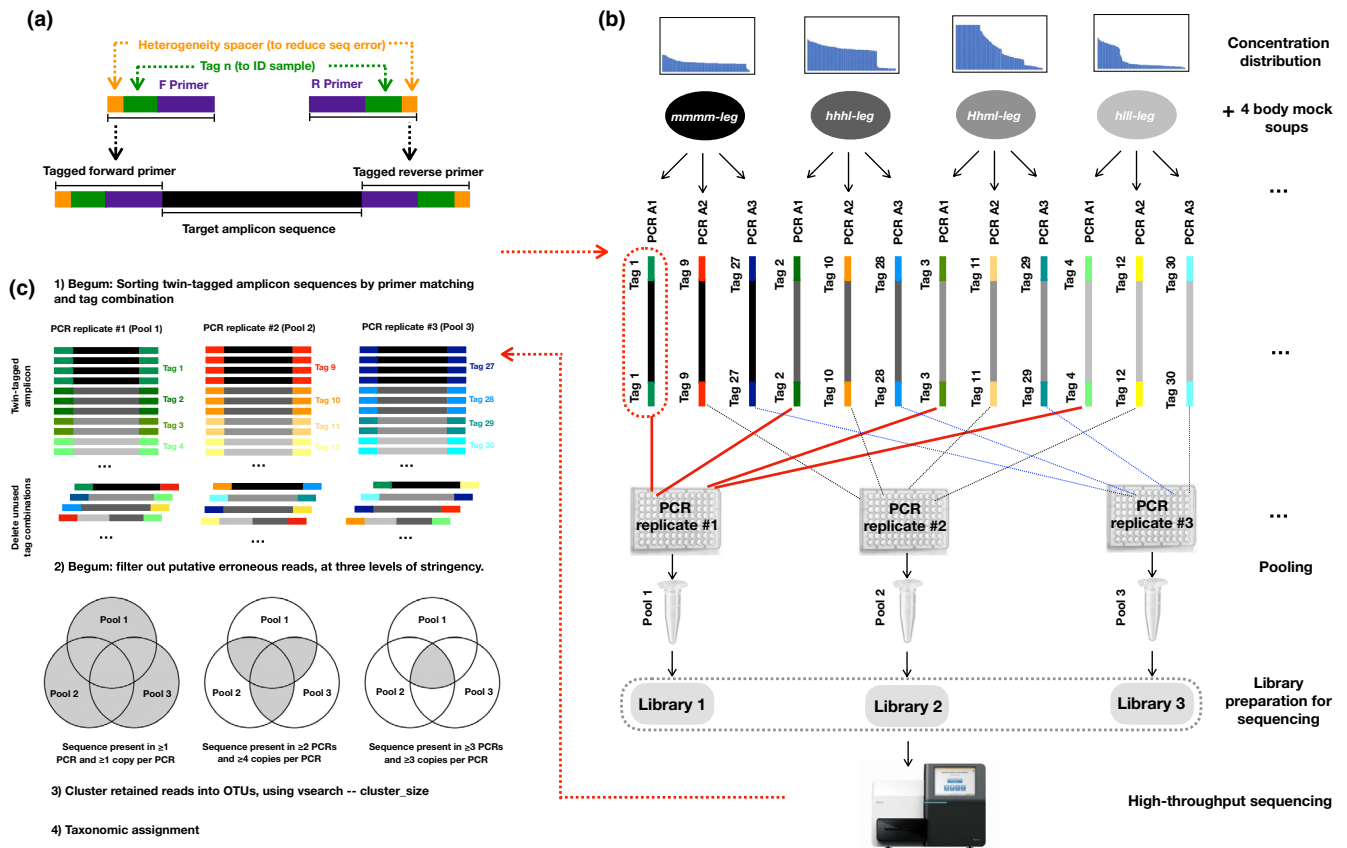
2 | MATERIALS AND METHODS

In S06\_Extended Methods, we present an unabridged version of this Methods section.

2.1 | Mock soup preparation

2.1.1 | Input species

We used Malaise traps to collect arthropods in Gaoligong Mountain, Yunnan province, China. From these, we selected 282 individuals that represented different morphospecies, and from each individual, we separately extracted DNA from the leg and the body. After clustering, we ended up with two hundred and forty-eight 97%-similarity DNA barcodes, which we used as the 'input species' for the mock soups (S07\_MTBFAS.fasta).



**FIGURE 1** Schematic of study. (a) Twin-tagged primers with heterogeneity spacers (above) and final amplicon structure (below). (b) Each mock soup (e.g. *Hhml-leg*) was PCR amplified three times (1, 2, 3) under a given PCR condition (A–H). Each of the three PCRs per soup used a different twin tag, following the *Begum* strategy. There were eight mock soups (*Hhml/hhhl/hlll/mmmm* X *body/leg*), where *H*, *h*, *m* and *l* indicate different DNA concentrations (details in Figure 2). PCR replicates 1 from each of the eight mock soups were pooled into the first amplicon pool (solid red lines), PCR replicates 2 were pooled into the second amplicon pool (black dashes) and PCR replicates 3 were pooled into the third amplicon pool (blue dashes). The entire set-up in B was repeated eight times for the eight PCR experiments (A–H), which thus generated ( $3 \times 8 =$ ) 24 sequencing libraries. (c) Key steps of the *Begum* bioinformatic pipeline. For clarity, primers and heterogeneity spacers not shown. The complete PCR set-up schematic, including positive and negative controls, is in Figure S9

### 2.1.2 | COI and genomic DNA quantification

To create the eight mock soups with different concentration evennesses of the 248 input species, we quantified DNA concentrations of their legs and bodies, using qPCR and a reference standard curve on the QuantStudio 12K Flex Real-Time PCR System (Life Technologies, Singapore) with Leray-FolDegenRev primers (Leray et al., 2013; Yu et al., 2012). We then diluted each species to their target DNA concentrations (Tables 2; Table S3). After dilution, we also measured each species' genomic-DNA concentrations, to test whether species OTU size can predict species genomic DNA masses, which is a proxy measure for animal biomass.

### 2.1.3 | Creation of mock soups

We used 1.0  $\mu$ l aliquots of the appropriately diluted leg and body DNA extracts of the 248 input species to create eight mock soups, achieving different profiles of COI-marker-concentration

evenness: *Hhml*, *hhhl*, *hlll* and *mmmm*, where *H*, *h*, *m* and *l* represent four different concentration levels (Figure 1, Table 2). For instance, in the *Hhml* soups, approximately one fourth of the input species was added at each concentration level (*H*, *h*, *m*, *l*), whereas in the *hlll* soup, three quarters of the species were diluted to the low concentration level before being added. These soups thus represent eight bulk samples with different absolute DNA concentrations (*leg* vs. *body*) and species evennesses (*Hhml*, *Hhml*, *hhhl*, *hlll*).

### 2.1.4 | Primer tag design

For DNA metabarcoding, we also used the Leray-FolDegenRev primer set, which has been shown to result in a high recovery rate of arthropods from mixed DNA soups (Alberdi et al., 2018; Leray et al., 2013), and we used OligoTag (Coissac, 2012) (Table S10) to design 100 unique tags of seven nucleotides in length in which no nucleotide is repeated more than twice, all tag pairs differ by at

**TABLE 2** The eight mock soups, each containing the same 248 arthropod OTUs but differing in absolute (Body/Leg) and relative (*Hhml*, *hhhl*, *hlll* and *mmmm*) DNA concentrations. Numbers in the table are the numbers of OTUs in each concentration category (*H*, *h*, *m*, *l*). Thus, the *Hhml\_body* soup contains 50 species with a DNA concentration between 50 and 200 ng/μl, each added as an aliquot of 1 μl, and so on. The evenness of DNA concentrations in each mock soup is summarized by the Shannon index. Higher values indicate a more even distribution. A few species provided only a low level of DNA concentration but were included in the *mmmm* soup as such

DNA extraction from arthropod body part	DNA concentration evenness	Number of OTUs in each concentration category				Total number of OTUs	Shannon index
		High ( <i>H</i> )	high ( <i>h</i> )	medium ( <i>m</i> )	low ( <i>l</i> )		
		50–200 ng/μl	10–48 ng/μl	1–8 ng/μl	0.001–0.1 ng/μl		
Body	<i>Hhml</i>	50	75	62	61	248	4.56
	<i>hhhl</i>	0	187	0	61	248	5.17
	<i>hlll</i>	0	61	0	187	248	4.08
	<i>mmmm</i>	0	0	247	1	248	5.39
	DNA concentration evenness	High ( <i>H</i> )	high ( <i>h</i> )	medium ( <i>m</i> )	low ( <i>l</i> )	Total number of OTUs	Shannon index
		5–60 ng/μl	0.1–3.0 ng/μl	0.009–0.09 ng/μl	0.0001–0.008 ng/μl		
		5–60 ng/μl	0.1–3.0 ng/μl	0.009–0.09 ng/μl	0.0001–0.008 ng/μl		
Legs	<i>Hhml</i>	69	63	63	53	248	4.21
	<i>hhhl</i>	0	195	0	53	248	5.04
	<i>hlll</i>	0	71	0	177	248	4.13
	<i>mmmm</i>	0	0	238	10	248	5.32

least three nucleotides, no more than three G and C nucleotides are present and none ends in either G or TT (to avoid homopolymers of GGG or TTT when concatenated to the Leray-FolDegenRev primers). We added one or two 'heterogeneity spacer' nucleotides to the 5' end of the forward and reverse primers (De Barba et al., 2014; Fadrosch et al., 2014), which cause sets of amplicons to be sequenced out of phase on the Illumina plate, reducing basecalling errors. The total amplicon length including spacers, tags, primers and markers was expected to be ~382 bp. The primer sequences are listed in Table S10.

## 2.1.5 | PCR optimization

We ran test PCRs using the Leray-FolDegenRev primers with an annealing temperature ( $T_a$ ) gradient of 40–64°C. Based on gel-band strengths, we chose an 'optimal'  $T_a$  of 45.5°C (clear and unique band on an electrophoresis gel) and a 'high'  $T_a$  value of 51.5°C (faint band) to compare their effects on species recovery.

We followed Murray et al. (2015) (see also Bohmann et al., 2018) and first ran the eight mock soups through qPCR to establish the correct dilution per soup so as to minimize PCR inhibition, to assess extraction-negative controls and to estimate the minimum cycle number needed to amplify the target fragment across samples. Based on the qPCR amplifications, we diluted six of the eight soups by five, 10- or 50-fold to minimize inhibition (S06\_Extended Methods), and we observed that the end of the exponential phase for all eight soups was achieved at or near 25 cycles, which we define here as the 'optimal' cycle number. To test the effect of PCR cycle number on species recovery, we also tested a 'low' cycle number

of 21 (i.e. stopping amplification during the exponential phase) and a 'high' cycle number of 30 (i.e. amplifying into the plateau phase).

## 2.1.6 | PCR amplifications of mock soups

We metabarcoded the mock soups under five different PCR conditions:

- A, B. Optimal  $T_a$  (45.5°C) and optimal PCR cycle number (25). A and B are technical replicates.
- C, D. High  $T_a$  (51.5°C) and optimal PCR cycle number (25). C and D are technical replicates.
- E. Optimal  $T_a$  (45.5°C) and low PCR cycle number (21).
- F. Optimal  $T_a$  (45.5°C) and high PCR cycle number (30).
- G, H. Touchdown PCR (Leray & Knowlton, 2015). Sixteen initial cycles: denaturation for 10 s at 95°C, annealing for 30 s at 62°C (–1°C per cycle) and extension for 60 s at 72°C, followed by 20 cycles at an annealing temperature of 46°C. G and H are technical replicates.

Following the *Begum* strategy, for each of the PCR conditions, each mock soup was PCR amplified three times, each time with a different-tag sequence on a different plate (Figure 1). The same-tag sequence was attached to the forward and reverse primers of a given PCR, which we call 'twin-tagging' (e.g. F1-R1, F2-R2, ...), to allow detection and removal of tag-jumped sequences, which produce non-twinning tags (e.g. F1-R2, F2-R3, ...). This lets us remove tag-jumped sequences, which assign species to the wrong samples (Schnell et al., 2015). In each PCR plate, we also included one positive control (with four insect

species), three extraction-negative controls and a row of PCR negative controls. PCR and tag setups are in Table S9.

### 2.1.7 | Illumina high-throughput sequencing

Sequencing libraries were created with the NEXTflex Rapid DNA-Seq Kit for Illumina (Bioo Scientific Corp.), following manufacturer instructions. In total, we generated 24 sequencing libraries (= 8 PCR conditions (A–H) × 3 PCR replicates/condition) (Figure 1), of which 18 were sequenced in one run of Illumina's V3 300 PE kit on a MiSeq at the Southwest Biodiversity Institute, Regional Instrument Center in Kunming. The six libraries from PCR conditions G and H were sequenced on a different run with the same kit type.

### 2.1.8 | Data processing

We removed adapter sequences, trimmed low-quality nucleotides and merged read-pairs with default parameters in *fastp* 0.20.1 (Chen et al., 2018). To allow fair comparison across PCR conditions, we subsampled 350,000 reads from each of the 24 libraries to achieve the same depth.

*Begum* is available at <https://github.com/shyams/Begum> (accessed 13 November 2020). First, we used *Begum's sort.py* (-pm 2 -tm 1) to demultiplex sequences by primers and tags, add the sample information to header lines and strip the spacer, tag and primer sequences. *Sort.py* reports the number of sequences that have novel tag combinations, representing tag-jumping events (mean 3.87%). We then used *Begum's filter.py* to remove sequences <300 bp and to filter out false-positive (erroneous) sequences (PCR and sequencing errors, chimeras, low-level contamination). We filtered at 12 levels of stringency:  $\geq 1$ –3 PCRs ×  $\geq 1$ –4 copies per PCR. For instance,  $\geq 1$  PCR and  $\geq 1$  copy represents no filtering, as this allows even single sequences that appear in only one PCR (i.e. 0\_0\_1, 0\_1\_0 or 1\_0\_0), and  $\geq 2$  PCRs and  $\geq 4$  copies represents moderately stringent filtering, as it allows only sequences that appear in at least two PCRs with at least four copies each (e.g. 32\_4\_0 but not 32\_2\_0).

We used *vsearch* 2.15.0 (Rognes et al., 2016) to remove de novo chimeras (--uchime\_denovo) and to produce a fasta file of representative sequences for 97% similarity Operational Taxonomic Units (OTUs; --cluster\_size) and a sample × OTU table (--otutabout). We assigned high-level taxonomies to the OTUs using *vsearch* (--syntax) on the MIDORI COI database (Leray et al., 2018) and only retained the OTUs assigned to Arthropoda with probability  $\geq 0.80$ . In R 4.0.0 (R Core Team, 2018), we set all cells in the OTU tables that contained only one read to 0 and removed the control samples.

### 2.1.9 | Metabarcoding efficiency

#### *False-negative and false-positive frequencies*

For each of the eight mock soups (Table 2), eight PCRs (A–H), and 12 *Begum* filtering stringencies (Tables 3; Table S05), we used

*vsearch* (--usearch\_global) to match the OTUs against the 248 input species and the four positive-control species (S07\_MTBFA5.fasta), and we removed any OTUs in the mock soups that matched a positive-control species. False negatives (dropouts) are defined as any of the 248 input species that failed to be matched by one or more OTUs at  $\geq 97\%$  similarity, and false positives are defined as OTUs that matched no input species at  $\geq 97\%$  similarity. For clarity, we only display results from the *mmm\_body* soups; results from all soups can be accessed in the DataDryad archive (Yu et al., 2021).

#### *Input DNA concentration and evenness and PCR conditions*

We used non-metric multidimensional scaling (NMDS) (metaMDS (distance="jaccard", binary=FALSE)) in {vegan} 2.5–6 (Oksanen et al., 2017) to visualize differences in OTU composition across the eight mock soups per PCR condition (Figure 1, Table 2). We evaluated the effects of species evenness on species recovery by using a linear mixed-effects model to regress the number of recovered input species on each mock soup's Shannon diversity (Table 2), lme4::lmer(OTUs~Evenness+(1|PCR) (Bates et al., 2015). Finally, we evaluated the information content of OTU size (number of reads) by linearly regressing input genomic-DNA concentration on OTU size.

#### *Taxonomic bias*

To visualize the effects of PCR conditions on taxonomic amplification bias, we used {metacoder} 0.3.4 (Foster et al., 2017) to pairwise compare the compositions of the *mmm\_body* soup under different PCR conditions.

#### *Tag-bias test*

We took advantage of the paired technical replicates in PCRs A&B, C&D and G&H (Table 3) to test for tag bias. For instance, we used the same eight tags in PCRs A1/B1, A2/B2 and A3/B3, and these three pairs should therefore return very similar communities. In contrast, the 12 non-matching pairs (e.g. A1/B2, A2/B1, A3/B1) used different tags and, if there is tag bias, should return differing communities. For each set of PCR replicates (A&B, C&D, G&H), we generated NMDS ordinations and used vegan::protest to calculate the mean Procrustes correlation coefficients for the same-tag ( $n = 3$ ) and different-tag pairs ( $n = 12$ ).

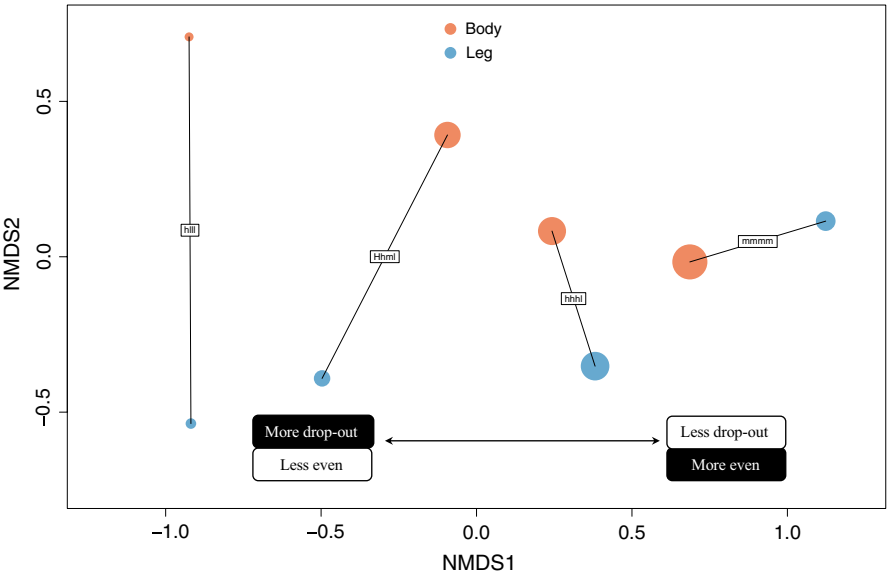
## 3 | RESULTS

The 18 libraries containing PCR sets A–F yielded 7,139,290 total paired-end reads, mean 396,627, and the six libraries of PCR sets G&H yielded 6,356,655 paired-end reads, mean 1,059,442. Each sample (e.g. *Hhml\_body* in PCR\_A) was sequenced in three libraries (Figure 1; Figure S5) and thus was represented by a mean of 132,209 reads (= 396,627 mean reads per library × 3 PCRs/9 samples per library, since each library contains eight mock soups + one positive control) in PCR sets A–F and a mean of 353,147 reads in PCR sets G and H.



**TABLE 3** Species-recovery success by three *Begum* filtering stringency levels and five PCR conditions, using the *mmmm\_body* soup. Recovered species are OTUs that match one of the 248 reference species at  $\geq 97\%$  similarity. False negatives (dropouts) are defined as reference species that fail to be matched by any OTU at  $\geq 97\%$  similarity. False-positive sequences are defined as OTUs that fail to match any reference species at  $\geq 97\%$  similarity. *Begum* filtering strongly reduces false-positive frequencies (dark- to light-orange cells) at the cost of a small rise in dropout frequency (light- to dark-blue cells), especially for optimal PCR conditions (PCRs A, B, E). With non-optimal PCR conditions (PCRs C, D, F, G, H), the trade-off is stronger; filtering to reduce false positives strongly increases dropouts (the blue cells are darker on the right-hand side of the table). See Section 3.1 for more details. Table S5 shows the same information for all 12 *Begum* stringency levels

Begum filtering parameters	Optimum T <sub>a</sub> + optimum cycle number (45.5°C, 25)		Optimum T <sub>a</sub> + <b>low cycle</b> <b>number</b> (45.5°C, 21)	High T <sub>a</sub> + optimum cycle number (51.5°C, 25)		Optimum T <sub>a</sub> + <b>high cycle</b> <b>number</b> (45.5°C, 30)	Touchdown PCR (62-46°C, 16+20 cycles)	
	A	B	E	C	D	F	G	H
Present in ≥1 PCR replicate with ≥1 copies per PCR (i.e. <i>no filtering</i> )	A	B	E	C	D	F	G	H
Recovered species: OTUs matched to Refs (≥97% similarity)	241	243	243	240	239	241	236	235
False-negative sequences (dropouts)	7	5	5	8	9	7	12	13
% False negatives (dropouts)	3%	2%	2%	3%	4%	3%	5%	5%
False-positive sequences	165	161	181	186	132	179	99	124
% False positives	67%	65%	73%	75%	53%	72%	40%	50%
Present in ≥2 PCR replicates with ≥4 copies per PCR	A	B	E	C	D	F	G	H
Recovered species: OTUs matched to Refs (≥97% similarity)	234	229	232	217	204	203	161	171
False-negative sequences (dropouts)	14	19	16	31	44	45	87	77
% False negatives (dropouts)	6%	8%	7%	13%	18%	18%	35%	31%
False-positive sequences	5	5	7	3	2	3	3	4
% False positives	2%	2%	3%	1%	1%	1%	1%	2%
Present in ≥3 PCR replicates with ≥3 copies per PCR	A	B	E	C	D	F	G	H
Recovered species: OTUs matched to Refs (≥97% similarity)	231	228	235	198	192	183	126	136
False-negative sequences (dropouts)	17	20	13	50	56	65	122	112
% False negatives (dropouts)	7%	8%	5%	20%	23%	26%	49%	45%
False-positive sequences	4	4	6	2	2	3	2	1
% False positives	2%	2%	2%	1%	1%	1%	1%	0%



**FIGURE 2** Non-metric multidimensional scaling (NMDS) ordination of eight mock soups, which differ in absolute (Body/Leg) and relative (*Hhml*, *hhhl*, *hlll* and *mmmm*) DNA concentrations of the input species (Table 2). Shown here is the output from the PCR A condition: optimum annealing temperature  $T_a$  (45.5°C) and cycle number (25), at *Begum* filtering stringency  $\geq 2$  PCRs,  $\geq 4$  copies/PCR (Table 3). Point size is scaled to the number of recovered OTUs. Species recovery is lower in samples with more uneven species frequencies (e.g. *hlll*) and, to a lesser extent, lower absolute DNA input (leg)

### 3.1 | Effects of PCR condition and *Begum* filtering

Optimal and near-optimal PCR conditions (PCRs A, B, E) achieved lower false-negative (dropout) frequencies than did non-optimal PCRs (high  $T_a$ , high cycle number or Touchdown) (PCRs C, D, F, G, H) (Table 3; Table S5).

With no *Begum* filtering ( $\geq 1$  PCR &  $\geq 1$  copy), false-positive OTUs were abundant, approaching the number of true OTUs (101–187 false-positive OTUs vs. 248 true OTUs) (Table 3; Table S5). Applying *Begum* filtering at different stringency levels reduced the number of false-positive sequences by 3–90 times. The cost of filtering was a greater loss of true OTUs but only by a small absolute amount in the optimal PCRs (A, B, E), rising from a dropout frequency of  $\sim 2\%$  in the nonfiltered

case to ~4%–6% under all but the two most stringent filtering levels, where dropout frequencies were 5%–11% ( $\geq 3$  PCR &  $\geq 3$  or 4 copies/PCR). In contrast, in the non-optimal PCRs (C, D, F, G, H), *Begum* filtering caused dropout frequencies to rise to much higher levels (5%–55%). In short, it is possible to combine wet-lab and bioinformatic protocols to reduce both false-positive and false-negative errors.

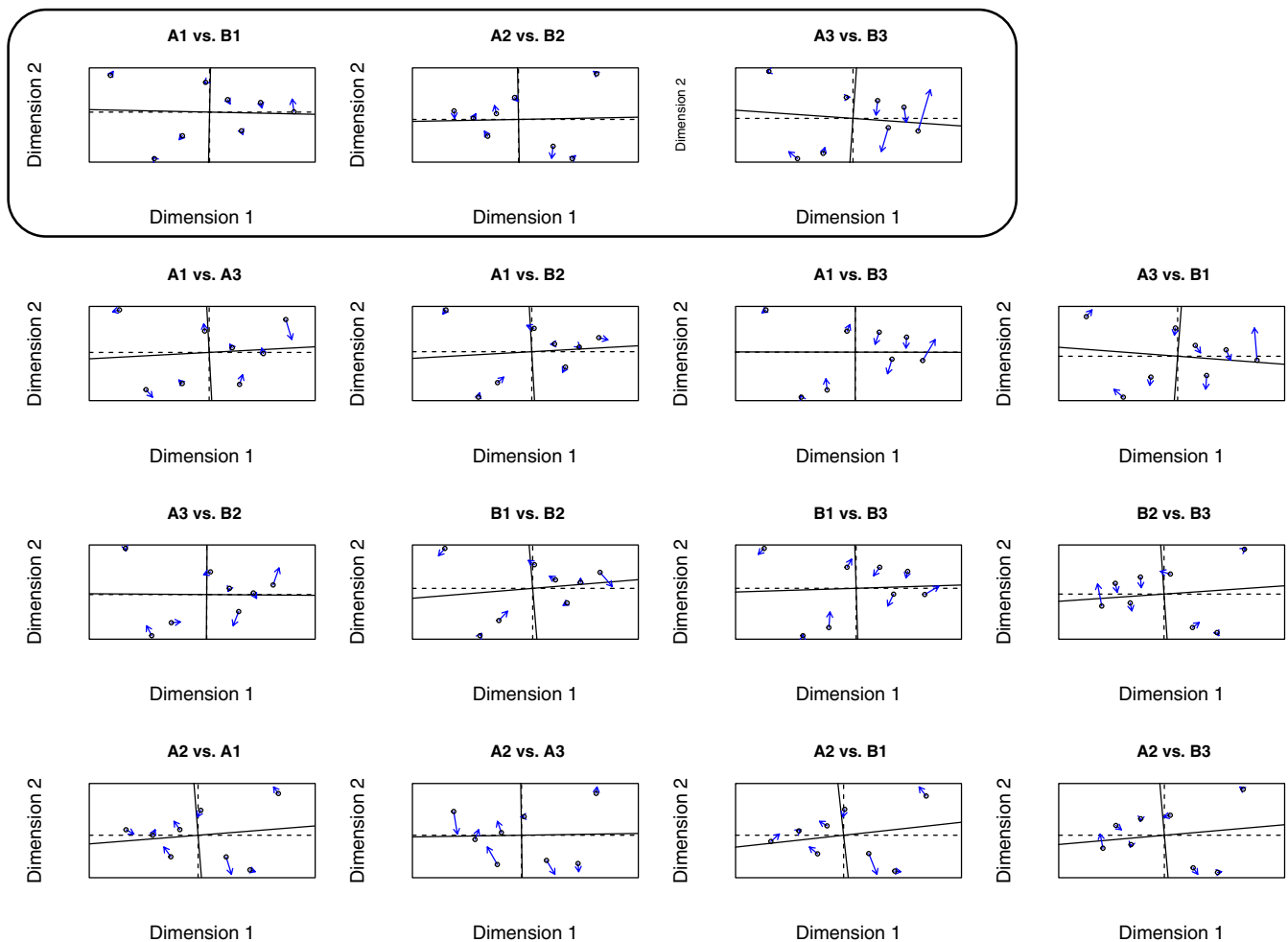
### 3.2 | Effects of input DNA absolute and relative concentrations on OTU recovery

Altering the relative (*Hhml*, *hhhl*, *hlll* and *mmmm*) and absolute (body, leg) input DNA concentrations created quantitative compositional differences in the OTU tables, as shown by NMDS ordination (Figure 2). Soup *hlll*, with the most uneven distribution of input DNA concentrations (Table 2), recovered the fewest OTUs (Figure 2). The same effect was seen by regressing the number of recovered OTUs on species evenness (Figure S1).

As expected, OTU size does a poor job of recovering information on input DNA amount per species (Figure S2). Although there are positive relationships between OTU size and DNA concentrations, the slope of the relationship differs depending on species relative abundances (*Hhml* vs. *hhhl* vs. *hlll*) and source tissues (leg vs. body), which reflects the action of multiple species-specific biases along the metabarcoding pipeline (McLaren et al., 2019). This interaction effect precludes the fitting of a robust model that relates OTU size to DNA concentration, since species-frequency and source-tissue information cannot be known a priori.

### 3.3 | Taxonomic amplification bias

Optimal PCR conditions (PCRs A, B, E) produce larger OTUs than do non-optimal PCR conditions (PCRs C, D, F, G, H), especially for Hymenoptera, Araneae and Hemiptera (Figure 4).



**FIGURE 3** Test for tag bias in the mock soups amplified at optimum annealing temperature  $T_a$  (45.5°C) and optimum cycle number (25) (PCRs A and B). All pairwise Procrustes correlations of PCRs A and B. The top row (box) displays the three same-tag pairwise correlations. The other rows display the 12 different-tag pairwise correlations. If there is tag bias during PCR, the top row should show a greater degree of similarity. However, mean correlations are not significantly different between same-tag and different-tag ordinations (Mean of same-tag correlations:  $0.99 \pm 0.007$  SD,  $n = 3$ . Mean of different-tag correlations:  $0.98 \pm 0.009$  SD,  $n = 12$ .  $p = 0.046$ ,  $df = 3.9$ , Welch's  $t$ -test). In Supplementary Information, we show the results for the high  $T_a$  (PCRs C & D) and Touchdown treatments (PCRs G & H)



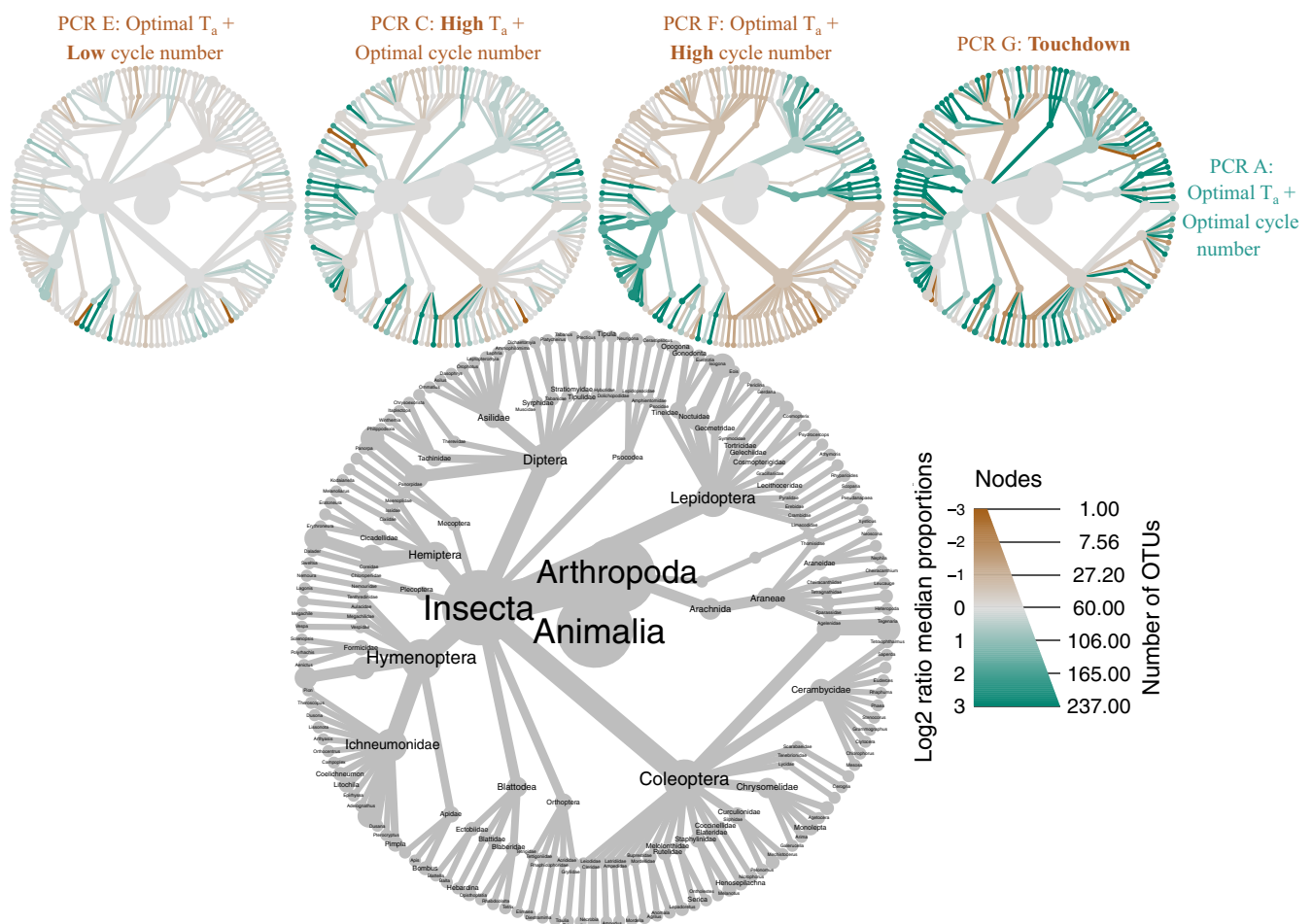
These are the taxa that are at higher risk of failing to be detected by the Leray-FolDegenRev primers under suboptimal PCR conditions.

### 3.4 | Tag-bias test

We found no evidence for tag bias in PCR amplification. For instance, under optimal PCR conditions (A & B), pairs using the same tags (A1/B1, A2/B2, A3/B3) and pairs using different tags (e.g. A1/B2, A2/B1, A3/B2, ...) both generated almost identical NMDS ordinations (Figure 3). Under non-optimal PCRs, we still found no evidence for tag bias, even though at higher annealing temperatures, some tag sequences might be more likely to aid primer annealing (Figures S3 and S4). Note that we did not correct the *p*-values for three tests, underlining the lack of evidence for tag bias.

## 4 | DISCUSSION

In this study, we tested our pipeline with eight mock soups that differed in their absolute and relative DNA concentrations of 248 arthropod taxa (Table 2, Figure 2). We metabarcoded the soups under five different PCR conditions that varied annealing temperatures ( $T_a$ ) and PCR cycles (Table 3), and we used *Begum* to filter the OTUs under different stringencies (Figure 1, Table 3). We define high efficiency in metabarcoding as recovering most of a sample's compositional and quantitative information, which in turn means that both false-negative and false-positive frequencies are low, that OTU sizes predict species relative abundances and that any dropouts are spread evenly over the taxonomic range of the target taxon (here, Arthropoda). This pipeline can of course be applied to other taxa, with appropriate adjustments to primer design, length limits, taxonomic reference database and controls.



**FIGURE 4** Taxonomic amplification bias of non-optimal PCR conditions. Pairwise-comparison heat trees of PCRs E, C, F and G versus the optimal PCR A (Table 3). Green branches indicate that PCR A (right side) produced relatively larger OTUs in those taxa. Brown branches indicate that PCR A produced smaller OTUs. Grey branches indicate similar OTU sizes. There are, on balance, more dark-green branches than dark-brown branches in the three heat trees that compare PCRs C, F and G (suboptimal) with PCR A (optimal), and the green branches are concentrated in the Araneae, Hymenoptera and Lepidoptera, suggesting that these are the taxa at higher risk of failing to be detected by Leray-FolDegenRev primers under suboptimal PCR conditions. Shown here are the *mmmmbody* soups, at *Begum* filtering stringency  $\geq 2$  PCRs,  $\geq 4$  copies per PCR

Our results show that metabarcoding efficiency can be made high for the recovery of species presence-absence, but efficiency is low for the recovery of quantitative information. Efficiency increases when the annealing temperature and PCR cycle number are at the low ends of ranges currently reported in the literature for this primer pair (Table 3, Figure 4). We recovered Elbrecht et al.'s (2017) finding that efficiency is higher when species evenness is higher (Figure 2; Figure S1), and we found that OTU sizes are a poor predictor of input genomic DNA, which confirms the conventional wisdom that OTU size is a poor predictor of species relative abundances (Figure S2; McLaren et al., 2019). Finally, we found no evidence for tag bias during PCR (Figure 3; Figures S3 and S4).

#### 4.1 | Co-designed wet-lab and bioinformatic methods to remove errors

The *Begum* workflow co-designs the wet-lab and bioinformatic components (Figure 1) (Zepeda-Mendoza et al., 2016) to minimize multiple sources of error (Table 1). Apart from the use of qPCR to optimize PCR conditions, the wet-lab and bioinformatic components are designed to work together. Twin-tagging allows removal of tag jumps, which result in sample misassignments. Multiple, independent PCRs per sample allow removal of false-positive sequences caused by PCR and sequencing error and by low-level contamination, at the cost of only a small absolute rise in false-negative error (Tables 3; Table S5). qPCR optimization reduces false negatives caused by PCR runaway, PCR inhibition and annealing failure (Tables 3; Table S5; Figure 4). Moderate dilution appears to be a better solution for inhibition than is increasing cycle number, since the latter increases dropouts (Tables 3; Table S5). qPCR also allows extraction blanks to be screened for contamination. Size sorting (Elbrecht et al., 2017) should reduce false negatives caused by PCR runaway, and the lower recovery of input species in the leg-only soups (Figure 2) argues that large insects should be represented by their heads, not their legs, for DNA extraction.

##### 4.1.1 | Begum filtering and complex positive controls

Increasing the stringency of *Begum* filtering reduces false-positive sequences at the cost of increasing false negatives (dropouts), although fortunately, this trade-off is weakened under optimal PCR conditions (Tables 3; Table S5). The choice of a filtering stringency level for a given study should be informed by complex positive-control samples and should take into account the study's aims. If the aim is to detect a particular taxon, like an invasive pest, it is better to set stringency low to minimize dropout, whereas if the aim is to generate data for an occupancy model, it is better to set stringency high to minimize false positives. Positive controls should be made of diverse taxa not from the study area (Creedy et al., 2019) and span a range of concentrations. Alternatively, a suite of synthetic DNA sequences with appropriate primer-binding regions could be used.

In metabarcoding pipelines, it is common to apply heuristic filters to remove false-positive sequences. For instance, small OTUs are commonly removed (<http://evomics.org/wp-content/uploads/2016/01/phyloseq-Lab-01-Answers.html>, accessed 11 November 2020). We did not do this because we wanted to isolate the effect of *Begum* filtering (and in fact we found that doing so slightly reduced species recovery). We did set to zero all cells in our OTU tables that contained only one read, and the only effect was to greatly reduce the number of false-positive sequences in the case when *Begum* filtering was not applied. Once any level of *Begum* filtering had been applied, those 1-read cells also disappeared (D. Yu, data not shown). Another common correction is to use the R package `LULU` (Frøslev et al., 2017) to combine 'parent' and 'child' OTUs that had failed to cluster. In this study, we could not do this because all input species had been included in all eight soups, which means that OTU co-occurrence could not be used to identify parent-child pairings.

##### 4.1.2 | Future work

*Begum* uses occurrence in multiple, independent PCRs to identify and remove erroneous sequences. This contrasts with solutions such as DADA2 (Callahan et al., 2016) and UNOISE2 (Edgar, 2016) that use only sequence-quality data to remove erroneous sequences. Unique molecular identifiers (UMIs) are also a promising method for the removal of erroneous sequences (Fields et al., 2019). It should be possible to combine some of these methods in the future.

A second area of research is to improve the recovery of quantitative information. Spike-ins and UMIs can be part of the solution (Deagle et al., 2018; Hoshino & Inagaki, 2017; Ji et al., 2020; Smets et al., 2016; Tkacz et al., 2018), but they can only correct for sample-to-sample stochasticity ('row noise') and differences in total DNA mass across samples. Such corrections allow the tracking of *within-species change across samples*, which means tracking how each individual species changes in abundance along a time series or environmental gradient. However, spike-ins and UMIs cannot be used to estimate *species relative abundances within a sample*, because spike-ins do not remove species biases in DNA extraction and primer-binding efficiencies. Thus, we caution against the uncritical use of metabarcoding to identify major and minor diet components (e.g. Deagle et al., 2019). Fortunately, methods for estimating species relative abundances are being developed (Lang et al., 2019; Peel et al., 2019; Williamson et al., 2019).

#### ACKNOWLEDGEMENTS

The authors thank Mr. Zongxu Li in South China Barcoding Center for help with arthropod selection and morphological identification. C.Y., D.W.Y., X.W. and W.C. were supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA20050202), the National Natural Science Foundation of China (41661144002, 31670536, 31400470, 31500305), the Key Research Program of Frontier Sciences, CAS

(QYZDY-SSW-SMC024), the Bureau of International Cooperation (GJHZ1754), the Ministry of Science and Technology of China (2012FY110800), the State Key Laboratory of Genetic Resources and Evolution (GREKF18-04) at the Kunming Institute of Zoology, the University of East Anglia and the University of Chinese Academy of Sciences. D.W.Y. was supported by a Leverhulme Trust Research Fellowship. K.B. was supported by the Danish Council for Independent Research (DFF-5051-00140).

## CONFLICT OF INTEREST

D.W.Y. is a co-founder of NatureMetrics ([www.naturemetrics.co.uk](http://www.naturemetrics.co.uk)), which provides commercial metabarcoding services.

## AUTHORS' CONTRIBUTIONS

D.W.Y. and C.Y. designed the project; C.Y. and K.B. designed the laboratory protocol; C.Y. and W.C. conducted the laboratory work; Z.D. performed the library building and Miseq sequencing; N.W. prepared the primer and tag design Excel spreadsheet; S.G. wrote *Begum*; X.W. wrote additional programs; D.W.Y. and C.Y. wrote the bioinformatic pipeline and performed data analysis; D.W.Y. wrote the first draft of the paper, and C.Y. and K.B. contributed revisions.

## PEER REVIEW

The peer review history for this article is available at <https://publons.com/publon/10.1111/2041-210X.13602>.

## DATA AVAILABILITY STATEMENT

We have archived a tutorial with a reduced sequence dataset and simplified scripts (PCR B only, 253 MB), and we have archived all sequence data, reference files, folder structure, output files and scripts (9.75 GB) on DataDryad (Yu et al., 2021a). To run the scripts from the beginning, remove the output files as instructed in the README file. The scripts are also published on Zenodo (Yu et al., 2021b).

## ORCID

Wang Cai  <https://orcid.org/0000-0002-8132-5821>

Douglas W. Yu  <https://orcid.org/0000-0001-8551-5609>

## REFERENCES

- Abrams, J. F., Horig, L. A., Brozovic, R., Axtner, J., Crampton-Platt, A., Mohamed, A., Wong, S. T., Sollmann, R., Yu, D. W., & Wilting, A. (2019). Shifting up a gear with iDNA: From mammal detection events to standardised surveys. *Journal of Applied Ecology*, 56(7), 1637–1648. <https://doi.org/10.1111/1365-2664.13411>
- 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>
- Axtner, J., Crampton-Platt, A., Horig, L. A., Mohamed, A., Xu, C. C. Y., Yu, D. W., & Wilting, A. (2019). An efficient and robust laboratory workflow and tetrapod database for larger scale environmental DNA studies. *Gigascience*, 8(4), 1–17. <https://doi.org/10.1093/gigascience/giz029>
- Barsoum, N., Bruce, C., Forster, J., Ji, Y. Q., & Yu, D. W. (2019). The devil is in the detail: Metabarcoding of arthropods provides a sensitive measure of biodiversity response to forest stand composition compared with surrogate measures of biodiversity. *Ecological Indicators*, 101, 313–323. <https://doi.org/10.1016/j.ecolind.2019.01.023>
- Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Bensasson, D., Zhang, D. X., Hartl, D. L., & Hewitt, G. M. (2001). Mitochondrial pseudogenes: Evolution's misplaced witnesses. *Trends in Ecology & Evolution*, 16(6), 314–321. [https://doi.org/10.1016/S0169-5347\(01\)02151-6](https://doi.org/10.1016/S0169-5347(01)02151-6)
- Berry, D., Ben Mahfoudh, K., Wagner, M., & Loy, A. (2011). Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Applied and Environmental Microbiology*, 77(21), 7846–7849. <https://doi.org/10.1128/Aem.05220-11>
- Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., Yu, D. W., & de Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Tree*, 29, 358–367.
- Bohmann, K., Gopalakrishnan, S., Nielsen, M., Nielsen, L. D. B., Jones, G., Streicker, D. G., & Gilbert, M. T. P. (2018). Using DNA metabarcoding for simultaneous inference of common vampire bat diet and population structure. *Molecular Ecology Resources*, 18(5), 1050–1063. <https://doi.org/10.1111/1755-0998.12891>
- Bush, A., Zaccacaeus, C., Wendy, M., Teresita, M. P., Royce, S., Erik, E., Nellie, G., Mehrdad, H., Mélanie, R., Mélanie, R., & Donald, J. B. (2019). Studying ecosystems with DNA metabarcoding: Lessons from aquatic biomonitoring. *BioRxiv*, March (578591). <https://doi.org/10.1101/578591>
- 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(7), 581–583. <https://doi.org/10.1038/Nmeth.3869>
- Champlot, S., Berthelot, C., Pruvost, M., Bennett, E. A., Grange, T., & Geigl, E. M. (2010). An efficient multistrategy DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. *PLoS ONE*, 5(9), e13042. <https://doi.org/10.1371/journal.pone.0013042>
- Chen, S. F., Zhou, Y. Q., Chen, Y. R., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17), 884–890. <https://doi.org/10.1093/bioinformatics/bty560>
- 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. <https://doi.org/10.1111/1755-0998.12265>
- Coissac, E. (2012). *OligoTag: A program for designing sets of tags for next-generation sequencing of multiplexed samples*. Humana Press.
- Cole, R. J., Holl, K. D., Zahawi, R. A., Wickey, P., & Townsend, A. R. (2016). Leaf litter arthropod responses to tropical forest restoration. *Ecology and Evolution*, 6(15), 5158–5168. <https://doi.org/10.1002/ece3.2220>
- Cordier, T., Alonso-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., Chariton, A., Creer, S., Frühe, L., Keck, F., Keeley, N., Laroche, O., Leese, F., Pochon, X., Stoeck, T., Pawlowski, J., & Lanzén, A. (2020). Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Molecular Ecology*, 1–22. <https://doi.org/10.1111/mec.15472>
- Creedy, T. J., Ng, W. S., & Vogler, A. P. (2019). Toward accurate species-level metabarcoding of arthropod communities from the tropical forest canopy. *Ecology and Evolution*, 9(6), 3105–3116. <https://doi.org/10.1002/ece3.4839>
- 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. <https://doi.org/10.1111/1755-0998.12188>
- Deagle, B. E., Clarke, L. J., Kitchener, J. A., Polanowski, A. M., & Davidson, A. T. (2018). Genetic monitoring of open ocean biodiversity: An evaluation of DNA metabarcoding for processing continuous plankton

- recorder samples. *Molecular Ecology Resources*, 18(3), 391–406. <https://doi.org/10.1111/1755-0998.12740>
- Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., & Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: Not a perfect match. *Biology Letters*, 10, 20140562. <https://doi.org/10.1098/rsbl.2014.0562>
- Deagle, B. E., Kirkwood, R., & Jarman, S. N. (2009). Analysis of Australian fur seal diet by pyrosequencing prey DNA in faeces. *Molecular Ecology*, 18(9), 2022–2038. <https://doi.org/10.1111/j.1365-294X.2009.04158.x>
- Deagle, B. E., Thomas, A. C., McInnes, J. C., Clarke, L. J., Vesterinen, E. J., Clare, E. L., 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., Machler, E., Seymour, M., Lacoursiere-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>
- Edgar, R. C. (2016). UNOISE2: Improved error-correction for Illumina 16S and ITS amplicon sequencing. *BioRxiv*, <https://doi.org/10.1101/081257>
- Elbrecht, V., Peinert, B., & Leese, F. (2017). Sorting things out: Assessing effects of unequal specimen biomass on DNA metabarcoding. *Ecology and Evolution*, 7(17), 6918–6926. <https://doi.org/10.1002/ece3.3192>
- Eren, A. M., Vineis, J. H., Morrison, H. G., & Sogin, M. L. (2013). A filtering method to generate high quality short reads using illumina paired-end technology. *PLoS ONE*, 8(6), e66643. <https://doi.org/10.1371/journal.pone.0066643>
- Esling, P., Lejzerowicz, F., & Pawlowski, J. (2015). Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Research*, 43(5), 2513–2524. <https://doi.org/10.1093/nar/gkv107>
- Fadrosh, D. W., Ma, B., Gajer, P., Sengamalai, N., Ott, S., Brotman, R. M., & Ravel, J. (2014). An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome*, 2, 6. <https://doi.org/10.1186/2049-2618-2-6>
- Fernandes, K., van der Heyde, M., Bunce, M., Dixon, K., Harris, R. J., Wardell-Johnson, G., & Nevill, P. G. (2018). DNA metabarcoding – A new approach to fauna monitoring in mine site restoration. *Restoration Ecology*, 26(6), 1098–1107. <https://doi.org/10.1111/rec.12868>
- Fields, B., Moeskjær, S., Friman, V.-P., Andersen, S. U., & Young, J. P. W. (2019). MAUI-seq: Multiplexed, high-throughput amplicon diversity profiling using unique molecular identifiers. *BioRxiv*, <https://doi.org/10.1101/538587>
- Fonseca, V. G., Carvalho, G. R., Sung, W., Johnson, H. F., Power, D. M., Neill, S. P., Packer, M., Blaxter, M. L., Lamshead, P. J. D., Thomas, W. K., & Creer, S. (2010). Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature Communications*, 1, 98. <https://doi.org/10.1038/ncomms1095>
- Foster, Z. S. L., Sharpton, T. J., & Grunwald, N. J. (2017). Metacoder: An R package for visualization and manipulation of community taxonomic diversity data. *PLoS Computational Biology*, 13(2), e1005404. <https://doi.org/10.1371/journal.pcbi.1005404>
- Frøslev, T. G., Kjølner, R., Bruun, H. H., Ejrnaes, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A. J. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications*, 8(1), 1188. <https://doi.org/10.1038/s41467-017-01312-x>
- Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. A. C., & Baird, D. J. (2011). Environmental barcoding: A next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE*, 6(4), e17497. <https://doi.org/10.1371/journal.pone.0017497>
- Hardy, N., Berry, T., Kelaher, B. P., Goldsworthy, S. D., Bunce, M., Coleman, M. A., Gillanders, M. A., Connell, S. D., Blewitt, M., & Figueira, W. (2017). Assessing the trophic ecology of top predators across a recolonisation frontier using DNA metabarcoding of diets. *Marine Ecology Progress Series*, 573, 237–254. <https://doi.org/10.3354/meps12165>
- Harper, L. R., Lawson Handley, L., Hahn, C., Boonham, N., Rees, H. C., Gough, K. C., Lewis, E., Adams, I. P., Brotherton, P., Phillips, S., & Hanfling, B. (2018). Needle in a haystack? A comparison of eDNA metabarcoding and targeted qPCR for detection of the great crested newt (*Triturus cristatus*). *Ecology and Evolution*, 8(12), 6330–6341. <https://doi.org/10.1002/ece3.4013>
- Hering, D., Borja, A., Jones, J. I., Pont, D., Boets, P., Bouchez, A., Bruce, K., Drakare, S., Hänfling, B., Kahlert, M., Leese, F., Meissner, K., Mergen, P., Reyjol, Y., Segurado, P., Vogler, A., & Kelly, M. (2018). Implementation options for DNA-based identification into ecological status assessment under the European Water Framework Directive. *Water Research*, 138, 192–205. <https://doi.org/10.1016/j.watres.2018.03.003>
- Hoshino, T., & Inagaki, F. (2017). Application of stochastic labeling with random-sequence barcodes for simultaneous quantification and sequencing of environmental 16S rRNA genes. *PLoS ONE*, 12(1), e0169431. <https://doi.org/10.1371/journal.pone.0169431>
- Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., Kitching, R., Dolman, P. M., Woodcock, P., Edwards, F. A., Larsen, T. H., Hsu, W. W., Benedict, S., Hamer, K. C., Wilcove, D. S., Bruce, C., Wang, X., Levi, T., Lott, M., ... Yur, D. W. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, 16(10), 1245–1257. <https://doi.org/10.1111/ele.12162>
- Ji, Y., Huotari, T., Roslin, T., Schmidt, N. M., Wang, J., Yu, D. W., & Ovaskainen, O. (2020). SPIKEPIPE: A metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes. *Molecular Ecology Resources*, 20(1), 256–267. <https://doi.org/10.1111/1755-0998.13057>
- Lang, D., Tang, M., Hu, J., & Zhou, X. (2019). Genome-skimming provides accurate quantification for pollen mixtures. *Molecular Ecology Resources*, 19(6), 1433–1446. <https://doi.org/10.1111/1755-0998.13061>
- Lanzén, A., Lekang, K., Jonassen, I., Thompson, E. M., & Troedsson, C. (2016). High-throughput metabarcoding of eukaryotic diversity for environmental monitoring of offshore oil-drilling activities. *Molecular Ecology*, 25(17), 4392–4406. <https://doi.org/10.1111/mec.13761>
- Leray, M., Ho, S. L., Lin, I. J., & Machida, R. J. (2018). MIDORI server: A web-server for taxonomic assignment of unknown metazoan mitochondrial-encoded sequences using a curated database. *Bioinformatics*, 34(21), 3753–3754. <https://doi.org/10.1093/bioinformatics/bty454>
- Leray, M., & Knowlton, N. (2015). DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 112(7), 2076–2081. <https://doi.org/10.1073/pnas.1424971112>
- Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., Boehm, J. T., & Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, 10, 34. <https://doi.org/10.1186/1742-9994-10-34>
- McLaren, M. R., Willis, A. D., & Callahan, B. J. (2019). Consistent and correctable bias in metagenomic sequencing experiments. *Elife*, 8, e46923. <https://doi.org/10.7554/eLife.46923>
- Murray, D. C., Coghlan, M. L., & Bunce, M. (2015). From benchtop to desktop: Important considerations when designing amplicon sequencing workflows. *PLoS ONE*, 10(4), e0124671. <https://doi.org/10.1371/journal.pone.0124671>
- Nichols, R. V., Vollmers, C., Newsom, L. A., Wang, Y., Heintzman, P. D., Leighton, M., Green, R. E., & Shapiro, B. (2018). Minimizing polymerase biases in metabarcoding. *Molecular Ecology Resources*, 18(5), 927–939. <https://doi.org/10.1111/1755-0998.12895>



- O'Donnell, J. L., Kelly, R. P., Lowell, N. C., & Port, J. A. (2016). Indexed PCR primers induce template-specific bias in large-scale DNA sequencing studies. *PLoS ONE*, 11(3), e0148698. <https://doi.org/10.1371/journal.pone.0148698>
- Oksanen, J., Blanchet, F. G., Friendly, F., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, H., Szöcs, E., & Wagner, H. (2017). *vegan: Community ecology package. Ordination methods, diversity analysis and other functions for community and vegetation ecologists*. Version 2.4-5. Retrieved from <https://CRAN.R-project.org/package=vegan>
- Peel, N., Dicks, L. V., Clark, M. D., Heavens, D., Percival-Alwyn, L., Cooper, C., Davies, R. G., Leggett, R. M., & Yu, D. W. (2019). Semi-quantitative characterisation of mixed pollen samples using MinION sequencing and Reverse Metagenomics (RevMet). *Methods in Ecology and Evolution*, 10(10), 1690–1701. <https://doi.org/10.1111/2041-210X.13265>
- Piaggio, A. J., Engeman, R. M., Hopken, M. W., Humphrey, J. S., Keacher, K. L., Bruce, W. E., & Avery, M. L. (2014). Detecting an elusive invasive species: A diagnostic PCR to detect Burmese python in Florida waters and an assessment of persistence of environmental DNA. *Molecular Ecology Resources*, 14(2), 374–380. <https://doi.org/10.1111/1755-0998.12180>
- Piñol, J., Mir, G., Gomez Polo, P., & Agustí, N. (2015). Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular Ecology Resources*, 15, 819–830.
- Piñol, J., Senar, M. A., & Symondson, W. O. C. (2019). The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. *Molecular Ecology*, 28(2), 407–419. <https://doi.org/10.1111/mec.14776>
- Piper, A. M., Batovska, J., Cogan, N. O. I., Weiss, J., Cunningham, J. P., Rodoni, B. C., & Blacket, M. J. (2019). Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. *Gigascience*, 8(8), 1–22. <https://doi.org/10.1093/gigascience/giz092>
- Polz, M. F., & Cavanaugh, C. M. (1998). Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology*, 64(10), 3724–3730. <https://doi.org/10.1128/AEM.64.10.3724-3730.1998>
- R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahe, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, 4, e2584. <https://doi.org/10.7717/peerj.2584>
- Schnell, I. B., Bohmann, K., & Gilbert, M. T. (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>
- Shehzad, W., Riaz, T., Nawaz, M. A., Miquel, C., Poillot, C., Shah, S. A., Pompanon, F., Coissac, E., & Taberlet, P. (2012). Carnivore diet analysis based on next-generation sequencing: application to the leopard cat (*Prionailurus bengalensis*) in Pakistan. *Molecular Ecology*, 21(8), 1951–1965. <https://doi.org/10.1111/j.1365-294X.2011.05424.x>
- Smets, W., Leff, J. W., Bradford, M. A., McCulley, R. L., Lebeer, S., & Fierer, N. (2016). A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing. *Soil Biology and Biochemistry*, 96, 145–151. <https://doi.org/10.1016/j.soilbio.2016.02.003>
- Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). *Environmental DNA: For biodiversity research and monitoring* (Vol. 1). Oxford University Press.
- Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L. H. (2012). Environmental DNA. *Molecular Ecology*, 21(8), 1789–1793. <https://doi.org/10.1111/j.1365-294X.2012.05542.x>
- Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, 21(8), 2045–2050. <https://doi.org/10.1111/j.1365-294X.2012.05470.x>
- Thomsen, P. F., Kielgast, J., Iversen, L. L., Wiuf, C., Rasmussen, M., Gilbert, M. T. P., Orlando, L., & Willerslev, E. (2012). Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, 21(11), 2565–2573. <https://doi.org/10.1111/j.1365-294X.2011.05418.x>
- Tkacz, A., Hortal, M., & Poole, P. S. (2018). Absolute quantitation of microbiota abundance in environmental samples. *Microbiome*, 6(1), 110. <https://doi.org/10.1186/s40168-018-0491-7>
- Virgilio, M., Backeljau, T., Nevado, B., & De Meyer, M. (2010). Comparative performances of DNA barcoding across insect orders. *BMC Bioinformatics*, 11(1), 206. <https://doi.org/10.1186/1471-2105-11-206>
- Wang, X., Hua, F., Wang, L., Wilcove, D. S., Yu, D. W., & Burridge, C. (2019). The biodiversity benefit of native forests and mixed-species plantations over monoculture plantations. *Diversity and Distributions*, 25(11), 1721–1735. <https://doi.org/10.1111/ddi.12972>
- Williamson, B. D., Hughes, J. P., & Willis, A. D. (2019). A multi-view model for relative and absolute microbial abundances. *BioRxiv*, <https://doi.org/10.1101/761486>
- Yoccoz, N. G. (2012). The future of environmental DNA in ecology. *Molecular Ecology*, 21(8), 2031–2038. <https://doi.org/10.1111/j.1365-294X.2012.05505.x>
- Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., & Ding, Z. (2012). Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, 3(4), 613–623. <https://doi.org/10.1111/j.2041-210X.2012.00198.x>
- Yu, D. W., Yang, C. Y., Bohmann, K., Wang, X. Y., Wang, C., Wales, N., Ding, Z. L., & Gopalakrishnan, S. (2021a). Biodiversity soup II. *Dryad*, <https://doi.org/10.5061/dryad.ncjxksrc>
- Yu, D. W., Yang, C. Y., Bohmann, K., Wang, X. Y., Wang, C., Wales, N., Ding, Z. L., & Gopalakrishnan, S. (2021b). Biodiversity soup II Zenodo, scripts. *Zenodo*, <https://zenodo.org/badge/latestdoi/250890697>
- Zepeda-Mendoza, M. L., Bohmann, K., Carmona Baez, A., & Gilbert, M. T. (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, 255. <https://doi.org/10.1186/s13104-016-2064-9>
- Zhang, K., Lin, S., Ji, Y., Yang, C., Wang, X., Yang, C., Wang, H., Jiang, H., Harrison, R. D., & Yu, D. W. (2016). Plant diversity accurately predicts insect diversity in two tropical landscapes. *Molecular Ecology*, 25(17), 4407–4419. <https://doi.org/10.1111/mec.13770>
- Ziesemer, K. A., Mann, A. E., Sankaranarayanan, K., Schroeder, H., Ozga, A. T., Brandt, B. W., Zaura, E., Waters-Rist, A., Hoogland, M., Salazar-Garcia, D. C., Aldenderfer, M., Speller, C., Hendy, J., Weston, D. A., MacDonald, S. J., Thomas, G. H., Collins, M. J., Lewis, C. M., Hofman, C., & Warinner, C. (2015). Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. *Scientific Reports*, 5, 16498. <https://doi.org/10.1038/srep16498>
- 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 online in the Supporting Information section.

**How to cite this article:** Yang C, Bohmann K, Wang X, et al. Biodiversity Soup II: A bulk-sample metabarcoding pipeline emphasizing error reduction. *Methods Ecol Evol*. 2021;12: 1252–1264. <https://doi.org/10.1111/2041-210X.13602>