Methods in Ecology and Evolution

MS WANG CAI (Orcid ID : 0000-0002-8132-5821)

DR DOUGLAS W. YU (Orcid ID : 0000-0001-8551-5609)

Article type : Research Article

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

Chunyan Yang[1], Kristine Bohmann[2], Xiaoyang Wang[1], Wang Cai[1], Nathan Wales[2,3], Zhaoli Ding[4], Shyam Gopalakrishnan[2], and Douglas W. Yu[1,5,6]

[1] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China

[2] Section for Evolutionary Genomics, Globe Institute, Faculty of Health and Medical Sciences, University of Copenhagen, 1353 Copenhagen, Denmark

[3] Department of Archaeology, University of York, York, YO10 5DD, United Kingdom

[4] Biodiversity Genomics Center, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China

[5] School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk NR4 7TJ, UK

[6] Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming Yunnan, 650223 China

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 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

生物多样性汤（第二版）：更低错误率的高通量条形码流程

摘要：

1、高通量条形码在环境保护和管理相关决策研究中的适用性已获得广泛的共识，但是该方法若要在更广的领域被应用，还需要进一步减少 PCR 扩增、文库构建以及测序所带来的错误和偏好。本研究设计了一种针对混合生物样本的高通量条形码流程，这一流程改进了实验设计和生物信息学分析，能减少结果中的假阳性（如标签错配、嵌合体、错误序列）和假阴性。研究结果还显示，由于不同物种的扩增效率存在差异，基于扩增子数据的高通量条形码并不能获取准确的物种定量信息。

2、本研究人为构建了 8 个生物多样性汤——节肢动物的混合样本，每一个汤中都含有同样的 248 个节肢动物的形态种，它们的 DNA 被以不同的浓度比例混入这 8 个汤中。然后这 8 个汤在 5 种不同的 PCR 条件下被扩增。本研究的高通量条形码流程采取了双胞胎标记法来双向标记引物，每个样本都进行多次独立的 PCR 扩增，每一批次的 PCR 扩增都包含阳性与阴性对照，并且我们使用 qPCR 来优化高通量条形码 PCR 最终使用的退火温度和循环数。在生物信息学分析部分，我们采用的是 *Begum*（Zepeda Mendoza 等发表的 *DAMe* 的改进版，2016. *BMC Res. Notes* 9:255），它在样本拆分时可以忽略修饰标签的几个碱基，并允许引物序列的错配，且提高了运算速率。*Begum* 能去除由于标签跳动所产生的假阳性序列，以及通过设置多个 PCR 重复和序列的重复出现次数来去除 PCR 和测序等产生的错误序列。

3、本研究结果显示在使用 Leray-Fol-Degen-Rev 的引物对扩增时，采取较低的退火温度和循环数能有效提高 OTU 得率，和减少扩增偏好性。本研究还发现 Begum 能过滤掉几乎所有的标签跳动和错误序列。我们再次证实了样本中物种生物量的不均衡会导致结果中更多的物种信息丢失，而且 OTU 的大小并不能直接用于指示物种的生物量。最后，研究结果表明在优化的 PCR 条件下，并未产生明显的标签偏好现象（即由连接引物的标签不同而引发的 PCR 扩增偏好）。

4、为了让更多人能学习、验证、并设计适合自己的高通量条形码流程，我们提供了本研究用于构建生物多样性汤所有物种的序列、Illumina 测序数据、完整的分析命令脚本，以及用于引物标签设计的表格和指南。

关键词：混合生物样本的高通量条形码；环境 DNA；环境影响评价；假阴性；假阳性；Illumina 高通量测序；标签偏好

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 feces (Taberlet *et al.* 2012a; Taberlet *et al.* 2012b; Deiner *et al.* 2017). Following initial proof-of-concept studies (Fonseca *et al.* 2010; Hajibabaei *et al.* 2011; Thomsen *et al.* 2012; Yoccoz 2012; Yu *et al.* 2012; Ji *et al.* 2013) has come a flood of basic and applied research and even new journals and commercial service providers (Murray, Coghlan & Bunce 2015; Callahan *et al.* 2016; Zepeda-Mendoza *et al.* 2016; Alberdi *et al.* 2018; 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, '*dropouts*'), 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 (Ji *et al.* 2013; Hering *et al.* 2018; Abrams *et al.* 2019; Bush *et al.* 2019; Piper *et al.* 2019; Cordier *et al.* 2020; Cordier 2020), 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 (Hering *et al.* 2018; Axtner *et al.* 2019; 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 (Lanzén *et al.* 2016; Hering *et al.* 2018), to track ecological restoration (Cole *et al.* 2016; Fernandes *et al.* 2018; Barsoum *et al.* 2019; 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: (i) remove sequence-to-sample misassignment due to tag-jumping, (ii) reduce dropout frequency and taxonomic bias in amplification, and (iii) 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 demutiplex 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 (Piaggio *et al.* 2014; Harper *et al.* 2018).

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 lab and differing in absolute and relative DNA concentrations, ran them under five different PCR conditions, and used *Begum* to filter out erroneous sequences (Fig.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

(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, Peinert & Leese 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.

Methods

In S06_Extended Methods, we present an unabridged version of this Methods section.

Mock soup preparation

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 248 97%-similarity DNA barcodes, which we used as the 'input species' for the mock soups (S07_MTBFAS.fasta).

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 (Yu *et al.* 2012; Leray *et al.* 2013). We then diluted each species to their target DNA concentrations (Tables 2,

S03). 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.

Creation of mock-soups. – We used 1.0 µl 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 (Fig. 1, Table 2). For instance, in the *Hhml* soups, approximately one-fourth of the input species were 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*, *hhhl*).

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 (Leray *et al.* 2013; Alberdi *et al.* 2018), and we used OligoTag (Coissac 2012) (Table S10) to design 100 unique tags of 7 nucleotides in length in which no nucleotide is repeated more than twice, all tag pairs differ by at least 3 nucleotides, no more than 3 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; Fadrosh *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.

PCR optimization

We ran test PCRs using the Leray-FolDegenRev primers with an annealing temperature (T_a) gradient of 40 to 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, Coghlan and Bunce (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 minimise 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 6 of the 8 soups by 5, 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).

PCR amplifications of mock soups

We metabarcoded the mock soups under 5 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). 16 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 (Fig. 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-twinned tags (e.g. F1-R2, F2-R3,…). This lets us remove tag-jumped sequences, which assigning species to the wrong samples (Schnell, Bohmann & Gilbert 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 S09.

Ilumina high-throughput sequencing

Sequencing libraries were created with the NEXTflex Rapid DNA-Seq Kit for Illumina (Bioo Scientific Corp., Austin, USA), following manufacturer instructions. In total, we generated 24 sequencing libraries (= 8 PCR conditions (A-H) × 3 PCR replicates/condition) (Fig. 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 6 libraries from PCR conditions G and H were sequenced on a different run with the same kit type.

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 comparion 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/shyamsg/Begum (accessed 13 Nov 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 twelve levels of stringency: ≥1-3 PCRs × ≥1-4 copies per PCR. For instance, ≥1 PCR and ≥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 ≥2 PCRs and ≥4 copies represents moderately stringent filtering, as it allows only sequences that appear in at least 2 PCRs with at least 4 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* (`--sintax`) on the MIDORI COI database (Leray *et al.* 2018) and only retained the OTUs assigned to Arthropoda with probability ≥ 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.

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, S05), we used *vsearch* (`--usearch_global`) to match the OTUs against the 248 input species and the four positive-control species (S07_MTBFAS.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 ≥ 97% similarity, and false positives are defined as OTUs that matched no input species at ≥97% similarity. For clarity, we only display results from the *mmmm_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 visualise differences in OTU composition across the eight mock-soups per PCR condition (Fig. 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, Sharpton & Grunwald 2017) to pairwise-compare the compositions of the *mmmm_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).

Results

The 18 libraries containing PCR sets A-F yielded 7,139,290 total paired-end reads, mean 396,627, and the 6 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 (Figs. 1, S5) and thus was represented by a mean of 132,209 reads (=396,627 mean reads per library X 3 PCRs / 9 samples per library, since each library contains 8 mock soups + 1 positive control.) in PCR sets A-F and a mean of 353,147 reads in PCR sets G and H.

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, S05).

With no *Begum* filtering (≥1 PCR & ≥1 copy), false-positive OTUs were abundant, approaching the number of true OTUs (101-187 false-positive OTUs versus 248 true OTUs) (Table 3, S05). Applying *Begum* filtering at different stringency levels reduced the number of false-positive sequences by 3 to 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 ~2% in the nonfiltered case to ~4-6% under all but the two most stringent filtering levels, where dropout frequencies were 5-11% (≥3 PCRs & ≥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.

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 (Fig. 2). Soup *hlll*, with the most uneven distribution of input DNA concentrations (Table 2), recovered the fewest OTUs (Fig. 2). The same effect was seen by regressing the number of recovered OTUs on species evenness (Fig. S01).

As expected, OTU size does a poor job of recovering information on input DNA amount per species (Fig. S02). 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, Willis & Callahan 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*.

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 (Fig. 4). These are the taxa that are at higher risk of failing to be detected by the Leray-FolDegenRev primers under sub-optimal PCR conditions.

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 (Fig. 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 (Fig. S03, S04). Note that we did not correct the *p*-values for three tests, underlining the lack of evidence for tag bias.

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, Fig. 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 (Fig. 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.

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, Fig. 4). We recovered Elbrecht et al.'s (2017) finding that efficiency is higher when species evenness is higher (Fig. 2, S01), 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 (Fig. S02) (McLaren, Willis & Callahan 2019). Finally, we found no evidence for tag bias during PCR (Figs. 3, S03, S04).

Co-designed wet-lab and bioinformatic methods to remove errors

The *Begum* workflow co-designs the wet-lab and bioinformatic components (Fig. 1) (Zepeda-Mendoza *et al.* 2016) to minimise multiple sources of error (Table 1). Aside from the use of qPCR to optimise 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, S05). qPCR optimization reduces false negatives caused by PCR runaway, PCR inhibition, and annealing failure (Tables 3, S05; Fig. 4). Moderate dilution appears to be a better solution for inhibition than is increasing cycle number, since the latter increases dropouts (Tables 3, S05). qPCR also allows extraction blanks to be screened for contamination. Size sorting (Elbrecht, Peinert & Leese 2017) should reduce false negatives caused by PCR runaway, and the lower recovery of input species in the leg-only soups (Fig. 2) argues that large insects should be represented by their heads, not their legs, for DNA extraction.

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. S05). 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 minimise dropout, whereas if the aim is to generate data for an occupancy model, it is better to set stringency high to minimise false positives. Positive controls should be made of

diverse taxa not from the study area (Creedy, Ng & Vogler 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 Nov 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.

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 (Smets 2016; Hoshino & Inagaki 2017; Deagle *et al.* 2018; Tkacz, Hortala & Poole 2018; Ji *et al.* 2020), 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, Hughes & Willis 2019).

Acknowledgements

We thank Mr. Zongxu Li in South China Barcoding Center for help with arthropod selection and morphological identification. C.Y.Y., D.W.Y., W.X.Y., 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 Insitute 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).

Data Availability

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).

Author contributions

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

Conflict of interest declaration

D.Y. is a co-founder of NatureMetrics (www.naturemetrics.co.uk), which provides commercial metabarcoding services.

References

Abrams, J.F., Horig, L.A., Brozovic, R., Axtner, J., Crampton-Platt, A., Mohamed, A., . . . Wilting, A. (2019) Shifting up a gear with iDNA: From mammal detection events to standardised surveys. *Journal of Applied Ecology,* **56** (7)**,** 1637-1648. doi: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. doi: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. doi: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. doi: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. doi:doi:10.18637/jss.v067.i01.

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. doi:10.1128/Aem.05220-11.

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. doi:10.1111/1755-0998.12891.

Bush, A., Zaccaeus, C., Wendy, M., Teresita, M.P., Royce, S., Erik, E., . . . Donald, J.B. (2019) Studying ecosystems with DNA metabarcoding: lessons from aquatic biomonitoring. *BioRxiv,* **March** (578591). doi:doi: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. doi:10.1038/Nmeth.3869.

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. doi:10.1093/bioinformatics/bty560.

Coissac, E. (2012) *OligoTag: A Program for Designing Sets of Tags for Next-Generation Sequencing of Multiplexed Samples*. Humana Press, Totowa, NJ.

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. doi:10.1002/ece3.2220.

Cordier, T., Alonso-Saez, L., Apotheloz-Perret-Gentil, L., Aylagas, E., Bohan, D.A., Bouchez, A., . . . Lanzen, A. (2020) Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Molecular Ecology,* **00** 1-22. doi:10.1111/mec.15472.

Cordier, T.L.A.S.L.A.-P.-G.E.A.D.A.B.A.B.A.C.S.C.L.F. (2020) Ecosystems monitoring powered by environmental genomics: a review of current strategies with an implementation roadmap. *Molecular Ecology*. doi:doi: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. doi: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. doi: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. doi:10.1111/1755-0998.12740.

Deagle, B.E., Thomas, A.C., McInnes, J.C., Clarke, L.J., Vesterinen, E.J., Clare, E.L., . . . 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. doi:10.1111/mec.14734.

Deiner, K., Bik, H.M., Machler, E., Seymour, M., Lacoursiere-Roussel, A., Altermatt, F., . . . Bernatchez, L. (2017) Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology,* **26** (21)**,** 5872-5895. doi:10.1111/mec.14350.

Edgar, R.C. (2016) UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *BioRxiv*. doi: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. doi:10.1002/ece3.3192.

Fadrosh, D.W., Ma, B., Gajer, P., Sengamalay, 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. doi: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. doi: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*. doi:10.1101/538587.

Fonseca, V.G., Carvalho, G.R., Sung, W., Johnson, H.F., Power, D.M., Neill, S.P., . . . Creer, S. (2010) Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature Communications,* **1** 98. doi:10.1038/ncomms1095.

513 Foster, Z.S.L., Sharpton, T.J. & Grunwald, N.J. (2017) Metacoder: An R package for visualization
514 and manipulation of community taxonomic diversity data. *Plos Computational Biology,* **13**
515 (2)**,** e1005404. doi:10.1371/journal.pcbi.1005404.

516 Frøslev, T.G., Kjøller, R., Bruun, H.H., Ejrnaes, R., Brunbjerg, A.K., Pietroni, C. & Hansen, A.J.
517 (2017) Algorithm for post-clustering curation of DNA amplicon data yields reliable
518 biodiversity estimates. *Nature Communications,* **8** (1)**,** 1188. doi:10.1038/s41467-017-
519 01312-x.

520 Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G.A.C. & Baird, D.J. (2011) Environmental
521 Barcoding: A Next-Generation Sequencing Approach for Biomonitoring Applications
522 Using River Benthos. *PLoS One,* **6** (4)**,** e17497. doi:10.1371/journal.pone.0017497.

523 Harper, L.R., Lawson Handley, L., Hahn, C., Boonham, N., Rees, H.C., Gough, K.C., . . .
524 Hanfling, B. (2018) Needle in a haystack? A comparison of eDNA metabarcoding and
525 targeted qPCR for detection of the great crested newt (Triturus cristatus). *Ecol Evol,* **8** (12)**,**
526 6330-6341. doi:10.1002/ece3.4013.

527 Hering, D., Borja, A., Jones, J.I., Pont, D., Boets, P., Bouchez, A., . . . Kelly, M. (2018)
528 Implementation options for DNA-based identification into ecological status assessment
529 under the European Water Framework Directive. *Water Research,* **138** 192-205.
530 doi:10.1016/j.watres.2018.03.003.

531 Hoshino, T. & Inagaki, F. (2017) Application of Stochastic Labeling with Random-Sequence
532 Barcodes for Simultaneous Quantification and Sequencing of Environmental 16S rRNA
533 Genes. *PLoS One,* **12** (1)**,** e0169431. doi:10.1371/journal.pone.0169431.

534 Ji, Y., Ashton, L., Pedley, S.M., Edwards, D.P., Tang, Y., Nakamura, A., . . . Yu, D.W. (2013)
535 Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology*
536 *Letters,* **16** (10)**,** 1245-1257. doi:10.1111/ele.12162.

537 Ji, Y., Huotari, T., Roslin, T., Schmidt, N.M., Wang, J., Yu, D.W. & Ovaskainen, O. (2020)
538 SPIKEPIPE: A metagenomic pipeline for the accurate quantification of eukaryotic species
539 occurrences and intraspecific abundance change using DNA barcodes or mitogenomes.
540 *Molecular Ecology Resources,* **20** (1)**,** 256-267. doi: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. doi: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. doi:10.1111/mec.13761.

Leray, M., Ho, S.L., Lin, I.J. & Machida, R.J. (2018) MIDORI server: a webserver for taxonomic assignment of unknown metazoan mitochondrial-encoded sequences using a curated database. *Bioinformatics,* **34** (21)**,** 3753-3754. doi: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. doi:10.1073/pnas.1424997112.

Leray, M., Yang, J.Y., Meyer, C.P., Mills, S.C., Agudelo, N., Ranwez, V., . . . 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. doi: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. doi: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. doi:10.1371/journal.pone.0124671.

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. doi:10.1371/journal.pone.0148698.

Oksanen, J., Blanchet, F.G., Friendly, F., Kindt, R., Legendre, P., McGlinn, D., . . . Wagner, H. (2017) vegan: Community Ecology Package. Ordination methods, diversity analysis and other functions for community and vegetation ecologists. Version 2.4-5. . https://CRAN.R-project.org/package=vegan

Peel, N., Dicks, L.V., Heavens, D., Percival-Alwyn, L., Cooper, C., Clark, M.D., . . . 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. doi: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. doi:10.1111/1755-0998.12180.

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. doi:10.1093/gigascience/giz092.

R Core Team (2018) R: A Language and Environment for Statistical Computing, Vienna, Austria. *R Foundation for Statistical Computing*. 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. doi: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. doi:10.1111/1755-0998.12402.

Smets, W.L., 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. doi:doi: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. (2012a) Environmental DNA. *Molecular Ecology,* **21** (8)**,** 1789-1793. doi:10.1111/j.1365-294X.2012.05542.x.

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

599 Thomsen, P.F., Kielgast, J., Iversen, L.L., Wiuf, C., Rasmussen, M., Gilbert, M.T., . . . Willerslev,
600 E. (2012) Monitoring endangered freshwater biodiversity using environmental DNA.
601 *Molecular Ecology,* **21** (11)**,** 2565-2573. doi:10.1111/j.1365-294X.2011.05418.x.

602 Tkacz, A., Hortala, M. & Poole, P.S. (2018) Absolute quantitation of microbiota abundance in
603 environmental samples. *Microbiome,* **6** (1)**,** 110. doi:10.1186/s40168-018-0491-7.

604 Wang, X., Hua, F., Wang, L., Wilcove, D.S., Yu, D.W. & Burridge, C. (2019) The biodiversity
605 benefit of native forests and mixed-species plantations over monoculture plantations.
606 *Diversity and Distributions,* **25** (11)**,** 1721-1735. doi:10.1111/ddi.12972.

607 Williamson, B.D., Hughes, J.P. & Willis, A.D. (2019) A multi-view model for relative and
608 absolute microbial abundances. *BioRxiv*. doi:doi: http://dx.doi.org/10.1101/761486.

609 Yoccoz, N.G. (2012) The future of environmental DNA in ecology. *Molecular Ecology,* **21** (8)**,**
610 2031-2038. doi:10.1111/j.1365-294X.2012.05505.x.

611 Yu, D.W., Ji, Y., Emerson, B.C., Wang, X., Ye, C., Yang, C. & Ding, Z. (2012) Biodiversity soup:
612 metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring.
613 *Methods in Ecology and Evolution,* **3** (4)**,** 613-623. doi:10.1111/j.2041-210X.2012.00198.x.

614 Yu, D.W., Yang, C., Bohmann, K., Wang, X., Cai, W., Wales, N., . . . Gopalakrishnan, S. (2021a)
615 Biodiversity Soup II. *Dryad, Dataset*. https://doi.org/10.5061/dryad.ncjsxksrc

616 Yu, D.W., Yang, C., Bohmann, K., Wang, X., Cai, W., Wales, N., . . . Gopalakrishnan, S. (2021b)
617 Biodiversity Soup II *Zenodo, Scripts*. https://zenodo.org/badge/latestdoi/250890697

618 Zepeda-Mendoza, M.L., Bohmann, K., Carmona Baez, A. & Gilbert, M.T. (2016) DAMe: a
619 toolkit for the initial processing of datasets with PCR replicates of double-tagged
620 amplicons for DNA metabarcoding analyses. *BMC Research Notes,* **9** 255.
621 doi:10.1186/s13104-016-2064-9.

622 Zhang, K., Lin, S., Ji, Y., Yang, C., Wang, X., Yang, C., . . . Yu, D.W. (2016) Plant diversity
623 accurately predicts insect diversity in two tropical landscapes. *Molecular Ecology,* **25** (17)**,**
624 4407-4419. doi:10.1111/mec.13770.

625 Zizka, V.M.A., Elbrecht, V., Macher, J.N. & Leese, F. (2019) Assessing the influence of sample
626 tagging and library preparation on DNA metabarcoding. *Molecular Ecology Resources,* **19**
627 (4)**,** 893-899. doi:10.1111/1755-0998.13018.

Table 1. Four classes of metabarcoding errors and their causes. Not included are software bugs, general laboratory and field errors like mislabeling, 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 lab	Champlot *et al.* 2010; 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, Lejzerowicz & Pawlowski 2015; Schnell, Bohmann & Gilbert 2015
	Intraspecific variability across the marker leading to multiple OTUs from the same species	Virgilio *et al.* 2010; 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 ('Drop-outs,' 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; Pinol *et al.* 2015; Alberdi *et al.* 2018
	PCR inhibition	Murray, Coghlan & Bunce 2015
	PCR stochasticity	Pinol *et al.* 2015
	PCR runaway (loss of diversity caused by some sequences outcompeting others during PCR)	Polz & Cavanaugh 1998

	Predator and collector DNA dominating the PCR product and causing target taxa (e.g. diet items) to fail to amplify	Deagle, Kirkwood & Jarman 2009; Shehzad *et al.* 2012
	Too many PCR cycles and/or too high annealing temperature, leading to loss of sequences with low starting DNA	Pinol *et al.* 2015
Poor quantification of target species abundances or biomasses	PCR stochasticity	Deagle *et al.* 2014
	Primer bias	Pinol *et al.* 2015; Pinol, Senar & Symondson 2019
	Polymerase bias	Nichols *et al.* 2018
	PCR inhibition	Murray, Coghlan & Bunce 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)	Intra-specific variability across the marker leading to multiple OTUs with different taxonomic assignments	Clarke *et al.* 2014
	Incomplete reference databases	

Fig. 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 setup in B was repeated eight times for the eight PCR experiments (A-H), which thus generated (3 X 8 =) 24 sequencing libraries. **C**. Key steps of the *Begum* bioinformatic pipeline. For clarity, primers and heterogeneity spacers not shown. The complete PCR setup schematic, including positive and negative controls, is in S09.

(see uploaded PDF)

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 50-200 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 (*H*)	high (*h*)	medium (*m*)	low (*l*)		
		50-200 ng/µl	10-48 ng/µl	1-8 ng/µl	0.001-0.1 ng/µl		
Body	*Hhml*	50	75	62	61	248	4.56
	hhhl	0	187	0	61	248	5.17
	hlll	0	61	0	187	248	4.08
	mmmm	0	0	247	1	248	5.39
	DNA concentration evenness	High (*H*)	high (*h*)	medium (*m*)	low (*l*)	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		
Legs	*Hhml*	69	63	63	53	248	4.21
	hhhl	0	195	0	53	248	5.04

	hlll	0	71	0	177	248	4.13
	mmmm	0	0	238	10	248	5.32

Fig. 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 ≥2 PCRs, ≥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).

(see uploaded PDF)

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 ≥97% similarity. False negatives (dropouts) are defined as reference species that fail to be matched by any OTU at ≥97% similarity. False-positive sequences are defined as OTUs that fail to match any reference species at ≥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 darker-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 *Effects of PCR condition and* Begum *filtering* for more details. Table **S05** shows the same information for all twelve *Begum* stringency levels.

Begum filtering parameters	Optimum T_a + optimum cycle number (45.5°C, 25)		Optimum T_a + low cycle number (45.5°C, 21)	High T_a + optimum cycle number (51.5°C, 25)		Optimum T_a + high cycle number (45.5°C, 30)	Touchdown PCR (62-46°C, 16+20 cycles)	
Present in ≥1 PCR replicate with ≥1 copies per PCR (i.e. *no filtering*)	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%**

Fig. 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 ± 0.007 SD, n = 3. Mean of different-tag correlations: 0.98 ± 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).

(see uploaded PDF)

Fig. 4. Taxonomic amplification bias of non-optimal PCR conditions. Pairwise-comparison heat trees of PCRs E, C, F, & 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 (sub-optimal) 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 sub-optimal PCR conditions. Shown here are the *mmmmbody* soups, at *Begum* filtering stringency ≥2 PCRs, ≥4 copies per PCR.

(see uploaded PDF)

Fig. 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 setup in B was repeated eight times for the eight PCR experiments (A-H), which thus generated (3 X 8 =) 24 sequencing libraries. **C**. Key steps of the *Begum* bioinformatic pipeline. For clarity, primers and heterogeneity spacers not shown. The complete PCR setup schematic, including positive and negative controls, is in S09.

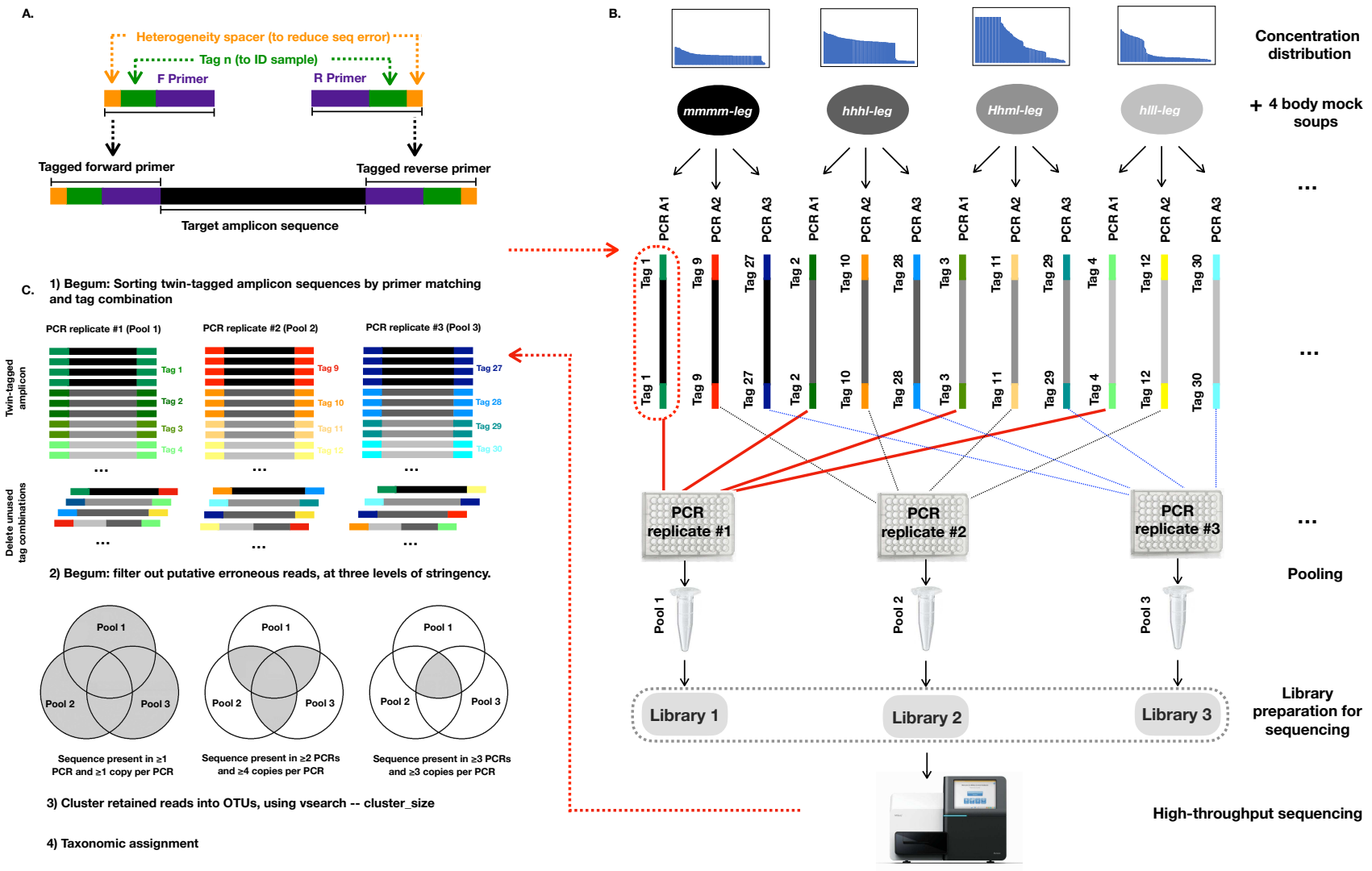


Fig. 2. Non-metric multidimensional scaling (NMDS) ordination of eight mock soups, which differ in absolute (Body/Leg) and relative (*Hhml*, *hhhl*, *hlll*, and *mmmm*) DNA concentrations of the input species (Table 2). Shown here is the output from the PCR A condition: optimum annealing temperature T_a (45.5 C) and cycle number (25), at *Begum* filtering stringency ≥2 PCRs, ≥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).

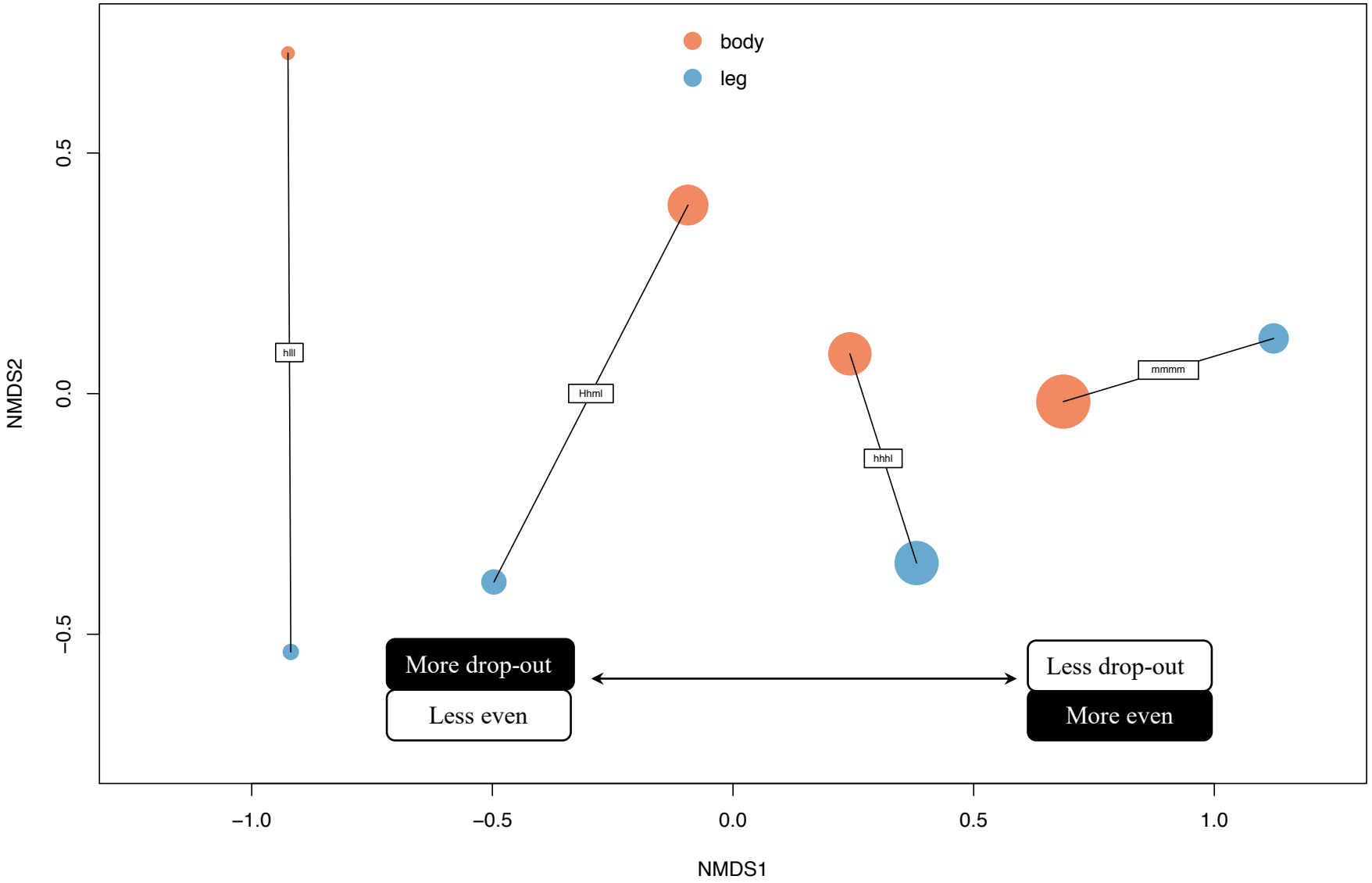
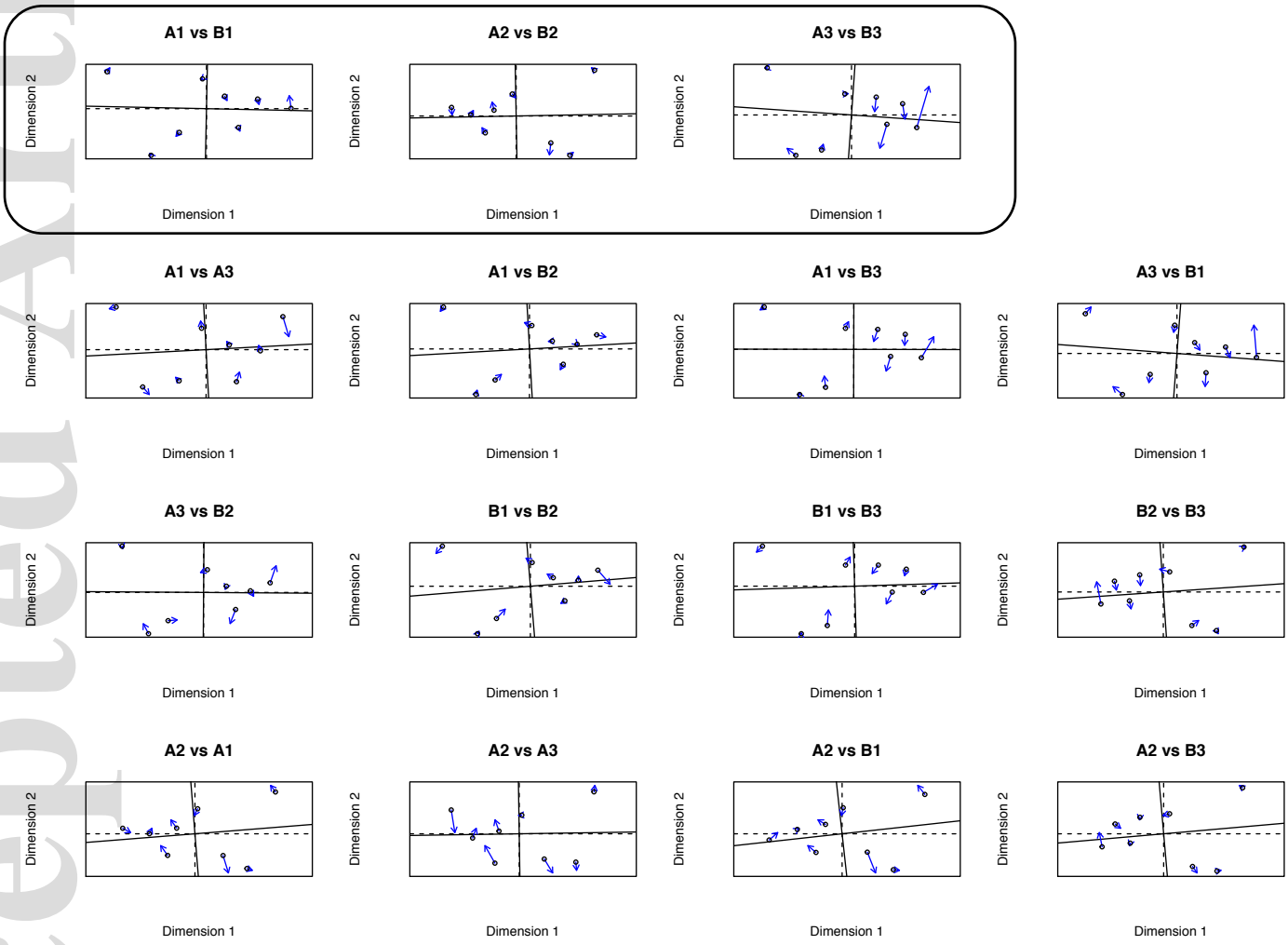


Fig. 3. Test for tag bias in the mock soups amplified at optimum annealing temperature T_a (45.5 °C) and optimum cycle number (25) (PCRs A and B). All pairwise Procrustes correlations of PCRs A and B. The top row (box) displays the three same-tag pairwise correlations. The other rows display the 12 different-tag 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 ± 0.007 SD, n = 3. Mean of different-tag correlations: 0.98 ± 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).



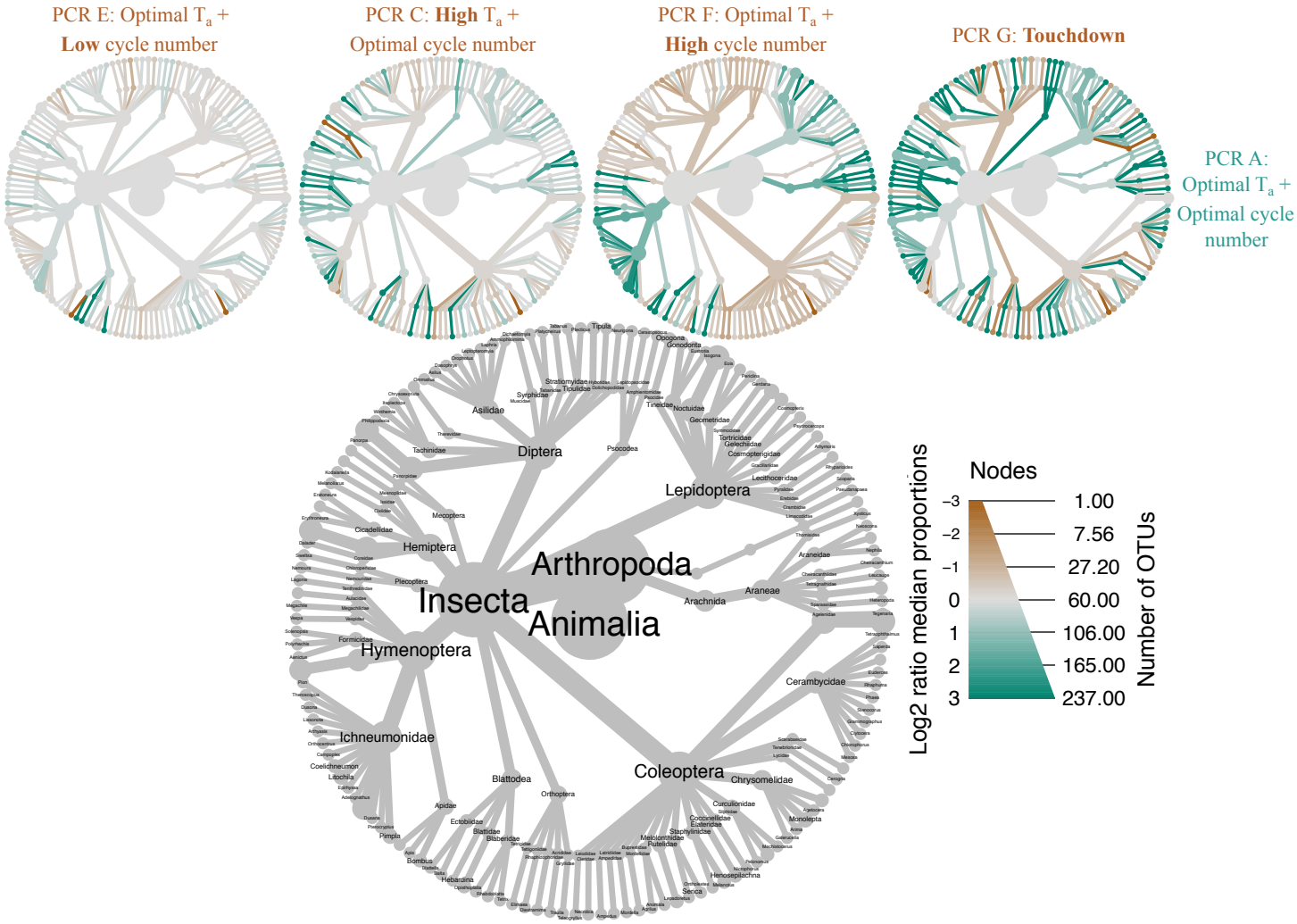


Fig. 4. Taxonomic amplification bias of non-optimal PCR conditions. Pairwise-comparison heat trees of PCRs E, C, F, & G versus PCR A (Table 3). Green branches indicate that PCR A (right side) produced relatively larger OTUs in those taxa. Brown branches indicate that PCR A produced smaller OTUs. Grey 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 (sub-optimal) 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 sub-optimal PCR conditions. Shown here are the *mmmmbody* soups, at *Begum* filtering stringency ≥2 PCRs, ≥4 copies per PCR.