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], Zhaoli Ding[3,4], 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] Kunming Biological Diversity Regional Center of Instruments, Chinese Academy of Sciences, Kunming 650223, China

[4] Public Technical Service Center, Kunming Institute of Zoology, Chinese Academy of Science, 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 ('drop-out') 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 at above a minimum copy number.

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-Fol-Degen-Rev 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 drop-outs. We replicate published findings that uneven size distribution of input biomasses leads to greater drop-out frequency and that OTU size is a poor predictor of species input biomass. Finally, we find no evidence that different primer tags bias PCR amplification ('tag bias').

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

Word count: 5844 (Intro to Refs) + 869 (Legends) = 6713 (limit 7000)
Abstract: 339 (limit 350)

Introduction

DNA metabarcoding enables rapid and cost-effective identification of eukaryotic 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 Piper *et al.* (2019) and Taberlet *et al.* (2018). 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

65 databases are incomplete or lacking. However, this is also a disadvantage, because we must
66 first spend effort to design efficient metabarcoding pipelines.

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

79 Here we consider one of the two main sample types used in metabarcoding: bulk-sample
80 DNA (the other type being environmental DNA, Bohmann et al., 2014) (Bohmann *et al.*
81 2014). Bulk-sample metabarcoding, such as mass-collected invertebrates, is being studied as
82 a way to generate multi-taxon indicators of environmental quality (Lanzén *et al.* 2016;
83 Hering *et al.* 2018), to track ecological restoration (Cole *et al.* 2016; Fernandes *et al.* 2018;
84 Barsoum *et al.* 2019; Wang *et al.* 2019), to detect pest species (Piper *et al.* 2019), and to
85 understand the drivers of species diversity gradients (Zhang *et al.* 2016).

86 We present a co-designed wet-lab and bioinformatic pipeline that uses qPCR-optimized PCR
87 conditions, three independent PCR replicates per sample, twin-tagging, and negative and
88 positive controls to: (i) remove sequence-to-sample misassignment due to tag-jumping, (ii)
89 reduce drop-out frequency and taxonomic bias in amplification, and (iii) reduce drop-in
90 frequency.

91 As part of the pipeline, we introduce a new version of the *DAMe* software package (Zepeda-
92 Mendoza *et al.* 2016), renamed *Begum* (Hindi for 'lady'), to demutiplex samples, remove tag-
93 jumped sequences, and filter out erroneous sequences (Alberdi *et al.* 2018). Regarding the
94 latter, the *DAMe*/*Begum* logic is that true sequences are more likely to appear in multiple,
95 independent PCR replicates and in multiple copies than are erroneous sequences (indels,
96 substitutions, chimeras). Thus, erroneous sequences can be filtered out by keeping only

Table 1. Four classes of metabarcoding errors and their causes. Not included are software bugs, general laboratory and field errors like mislabeling, DNA degradation, and sampling biases or inadequate effort.

Main Errors	Possible causes	References
False positives ('Drop-ins,' 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	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 the nature of the error)	Intra-specific variability across the marker leading to multiple OTUs with different taxonomic assignments	Clarke *et al.* 2014
	Incomplete reference databases	

sequences that appear in more than one (or a low number of) PCR replicate(s) at above a 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 (http://github.com/shyamsg/Begum).

To test our pipeline, we created eight mock arthropod soups, all containing the same 248 arthropod taxa but 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 ('drop-outs', i.e. unrecovered input species),

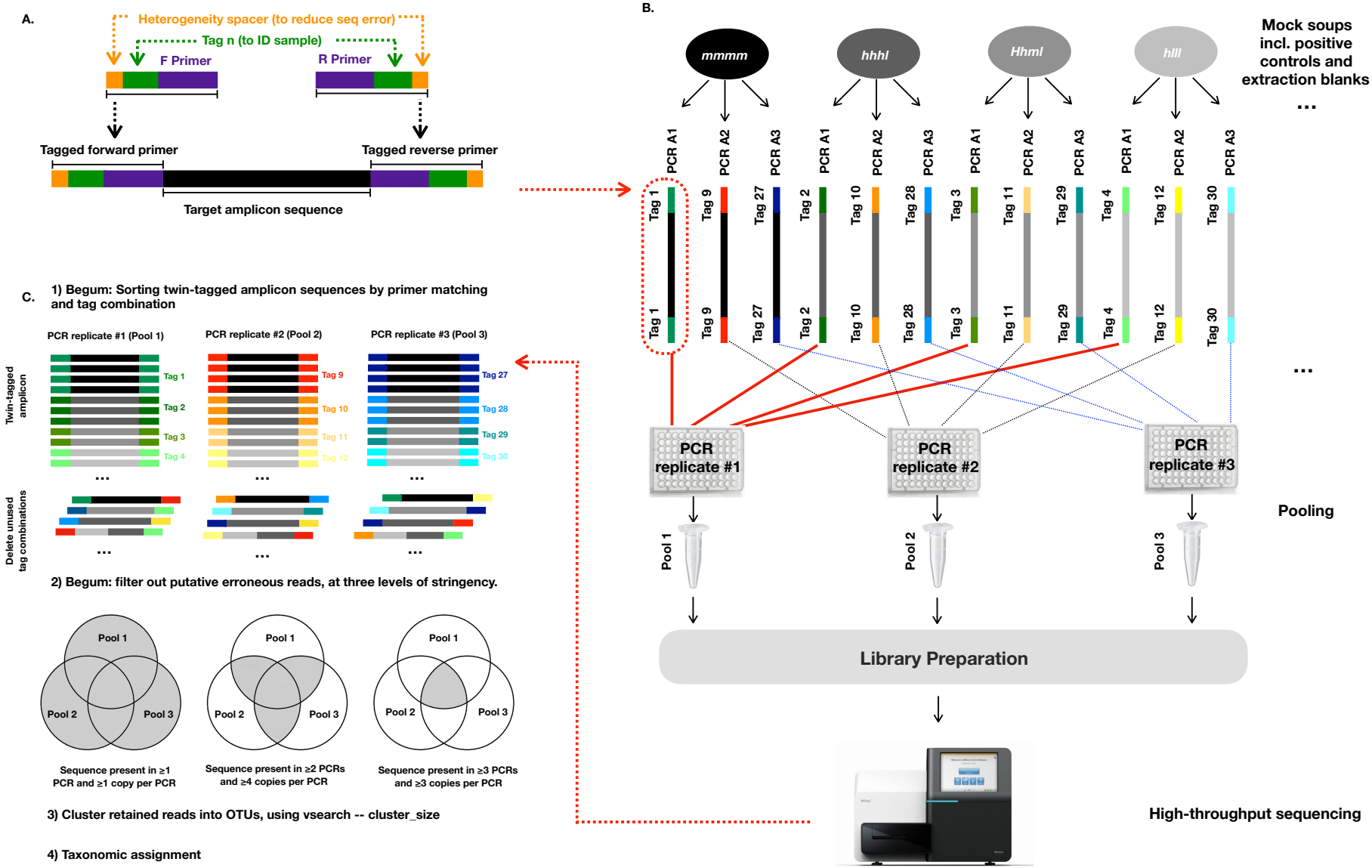
(2) the frequency of false-positive OTUs ('drop-ins'),

(3) the recovery of frequency information (does OTU size per species predict input DNA amount per species?), and

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

The highest efficiency would be achieved by recovering *all* and *only* the input species, in their original frequencies. We show that 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 (drop-ins), at the cost of only a small absolute rise in drop-out frequency, that greater species evenness and higher absolute concentrations reduce drop-outs (replicating Elbrecht, Peinert & Leese 2017), and that OTU sizes are not reliable estimators of species-biomass frequencies. We also find at most only small effect sizes 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 (e.g. Berry *et al.* 2011; O'Donnell *et al.* 2016), exacerbating taxonomic bias in PCR. All these results have important implications for using metabarcoding as a biomonitoring tool.

Fig. 1. Schematic of study. A. Twin-tagged primers with heterogeneity spacers (above) and final amplicon structure (below). **B.** Each mock soup (a 'sample') is PCR-amplified three times (#1, #2, #3), each with a different tag, following the Begum strategy. There are eight mock soups (*mmmm/hhhl/Hhml/hlll* X body/leg), and PCR replicates #1 from each of the eight mock soups are pooled into the first amplicon pool, PCR replicates #2 are pooled into the second amplicon pool, and so on. The setup in B. was itself replicated 8 times for the 8 PCR conditions (A-H), which generated (3 X 8 =) 24 pools in total. **C.** Key steps of the Begum bioinformatic pipeline. For clarity, primers and heterogeneity spacers not shown.



Methods

In Supp. Info. S1_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 barcoded 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 (Supp. Info. S2_MTBFAS.fasta).

DNA and COI quantification. – To create eight mock soups with different concentration evennesses of the 248 input species, we quantified their COI gene concentrations using qPCR with the Leray-Geller primers mlCOIintF and jgHCO2198 (Geller *et al.* 2013; Leray *et al.* 2013). To test the prediction value of OTU size (read number), we measured genomic DNA concentrations of the 248 input species.

Creation of mock-soups. – We used the leg and whole-body DNA extracts of the 248 input species to create eight mock soups (Supp. Info. S3_env_MTB). Each soup was pooled at different levels of COI-marker-concentration evenness: *Hhml*, *hhhl*, *hlll*, and *mmmm*, where *H*, *h*, *m*, and *l* represent four different COI marker concentration levels (Table 2). For instance, in the *Hhml* soup, one-fourth of the input species were included at the highest concentration (*H*), one-fourth at the next lower concentration (*h*), and so on. These soups thus represent eight bulk-samples with different absolute (leg vs. body) and relative biomass concentrations (*Hhml_leg*, *Hhml_body*, *hhhl_leg*, …), and we compare how these different concentrations affect species recovery. We also made a four-species positive control and three extraction negative controls. All eight mock soups, positive control and extraction negative controls are used in the DNA prescreening using qPCR step for PCR success monitoring, guides for Begum stringency selection during bioinformatic analysis, and contamination detection.

Primer tag design

For DNA metabarcoding, we also used the Leray-Geller 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) (Supp. Info. S4_Table 1) to design 100 unique tags of 7 nucleotides in length in which no nucleotide is repeated more

155 than twice, all tag pairs differ by at least 3 nucleotides, no more than 3 G and C nucleotides
156 are present, and none ends in either G or TT (to avoid homopolymers of GGG or TTT when
157 concatenated to the Leray-Geller primers). We increased diversity by adding one or two
158 'heterogeneity spacer' nucleotides to the 5' end of the forward and reverse tagged primers
159 (De Barba *et al.* 2014; Fadrosh *et al.* 2014). The total amplicon length including spacers,
160 tags, primers, and markers was expected to be ~382 bp. The primer sequences are listed in
161 Supp. Info. S4_Table 1.

162 *PCR optimization*

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

167 We followed Murray, Coghlan and Bunce (2015) (see also Bohmann *et al.* 2018) and first ran
168 the eight mock soups through qPCR to detect PCR inhibition and establish the appropriate
169 template amount, to assess extraction-negative controls, and to estimate the minimum cycle
170 number needed to amplify the target fragment across samples. qPCR amplification curves
171 indicated that 10X dilution of sample extracts was optimal to avoid PCR inhibition, and we
172 observed that the end of the exponential phase was achieved at 25 cycles, which we define
173 here as the 'optimal' cycle number. To test the effect of PCR cycle number on species
174 recovery, we also tested a 'low' cycle number of 21 (i.e. stopping amplification during the
175 exponential phase), and a 'high' cycle number of 30 (i.e. amplifying into the plateau phase).

176 *PCR amplifications of mock soups*

177 We metabarcoded the mock soups under 5 different PCR conditions:

178 **A, B**. Optimal T_a (45.5°C) and optimal PCR cycle number (25). A and B are technical
179 replicates.

180 **C, D**. High T_a (51.5°C) and optimal PCR cycle number (25). C and D are technical
181 replicates.

182 **E**. Optimal T_a (45.5°C) and low PCR cycle number (21).

183 **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 using a different tag sequence. The same tag sequence was added to the forward and reverse primers, which we call 'twin-tagging' (e.g. F1-R1, F2-R2, F3-R3, …), to ensure that tag-jumps would not result in false assignments of sequences to samples (Schnell, Bohmann & Gilbert 2015). In each PCR plate, we also included the positive control, three extraction-negative controls, and a row of PCR negative controls. Plate and tag setups are in Supp. Info. S5_PCRSETUP.xlsx.

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 (3 PCR replicates × 8 PCR replicates (A-H) =) 24 sequencing libraries (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 six libraries from PCR sets 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* 0.20.0 (Chen *et al.* 2018). We then 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 28 March 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 erroneous sequences. We filtered at three levels of stringency: 1) no filtering, 2) retaining sequences that appeared in minimum 2 of 3 PCR replicates with minimum 4 copies per PCR, and 3) retaining sequences that appeared in minimum 3 of 3 PCR replicates with minimum 3 copies per PCR.

We used *vsearch* 2.14.1 (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 used {lulu} 0.1.0 (Frøslev *et al.* 2017) to combine likely 'parent' and 'child' OTUs that had failed to cluster. For the remaining OTUs, we assigned taxonomies using *vsearch* (`--sintax`) and the MIDORI COI database (Leray *et al.* 2018). Only OTUs assigned to Arthropoda with probability ≥ 0.80 were retained.

In *R* 3.5.2 (R Core Team, 2018), we used {phyloseq} 1.19.1 (McMurdie & Holmes 2013) to choose minimum OTU sizes of 10 and 5 for the Begum-unfiltered and -filtered OTUs, respectively, and we filtered out these small OTUs and also set all cells with one read to 0. Lastly, we visually inspected the OTU tables and confirmed that none of the positive-control OTUs was present in any of the soup samples. After Begum-filtering, the OTU tables contained no sequences in the extraction-negative and PCR-negative controls.

Metabarcoding efficiency

Drop-out and Drop-in frequencies. – For each of the eight mock-soups (Table 2), eight PCRs (A-H), and three Begum filtering stringencies, we used *vsearch* (`--usearch_global`) to match the soup's OTU sequences against the 248 input species and the 4 positive-control species (Supp. Info. S2_MTBFAS.fasta). Drop-outs are defined as input species that failed to be matched by any OTU at ≥ 97% similarity, and drop-ins are defined as OTUs that matched to no input or positive-control species at ≥97% similarity. For clarity, we display results from the *mmmm_body* soups; all results can be accessed in the provided script.

Input DNA concentration and evenness and PCR conditions. – We used non-metric multidimensional scaling (NMDS) (`metaMDS(distance="bray", binary=FALSE)`) in {vegan} 2.4-5 (Oksanen *et al.* 2017) to visualise differences in OTU composition across the eight mock-soups per PCR condition (Table 2). We evaluated the information content of OTU size (number of reads) by regressing input DNA amount on OTU size, and we evaluated the effects of DNA concentration, evenness, and PCR condition on species recovery by regressing the number of recovered input species on each mock soup's Shannon diversity (Table 2).

Taxonomic bias. – To visualize the effects of PCR conditions on taxonomic amplification bias, we used {metacoder} 0.2.0 (Foster, Sharpton & Grunwald 2017) to pairwise-compare the compositions of the *mmmm_body* soup under different PCR conditions.

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*). The evenness of DNA concentrations in each mock soup is summarized by the Shannon index, in which higher values indicate a more even distribution of input DNA concentrations.

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	*hlll*	0	61	0	187	248	4.08
	Hhml	50	75	62	61	248	4.56
	hhhl	0	187	0	61	248	5.17
	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	*hlll*	0	71	0	177	248	4.13
	Hhml	69	63	63	53	248	4.21
	hhhl	0	195	0	53	248	5.04
	mmmm	0	0	238	10	248	5.32

Tag-bias test

We took advantage of the technical replicates in PCR sets A&B, C&D, and G&H (Table 3). For instance, the PCR sets A&B used the same eight tags in A1 and B1, and thus this pair should show no differences caused by tag bias. In contrast, non-matching pairs (e.g. A1 and B2, A2 and B1) used different tags, and if there is tag bias, those pairs should have produced differing communities. For each PCR set (A&B, C&D, G&H), we used `protest` in {vegan} to calculate Procrustes correlation coefficients on NMDS ordinations for pairs of PCRs that used the same tags (n = 3) and different tags (n = 12).

Results

The 18 libraries containing PCR sets A-F yielded 9,223,624 paired-end reads, and the 6 libraries of PCR sets G&H yielded 7,004,709 paired-end reads.

Effects of Begum filtering and PCR condition

The PCR conditions that resulted in the lowest drop-out frequencies were optimal T_a + optimal or low cycle numbers (PCRs A, B, E, Table 3). High T_a, high cycle number, or a Touchdown protocol all increased drop-out frequencies (PCRs C, D, F, G, H). Without any *Begum* filtering (reads present in ≥1 PCR replicate & ≥1 copy per PCR), erroneous OTUs (drop-ins) constituted ~27-46% of all OTUs (Table 3). Applying *Begum* filtering at either of the two stringency levels reduced drop-in frequencies by ~20-40 times. Importantly, the application of Begum filtering also caused drop-out frequencies to increase, but only by a small absolute amount (from 3-4% to 6-9%) *as long as* the PCR conditions were also optimal (A, B, E). In short, it is possible to combine wet-lab and bioinformatic protocols to simultaneously reduce false-negative and false-positive errors.

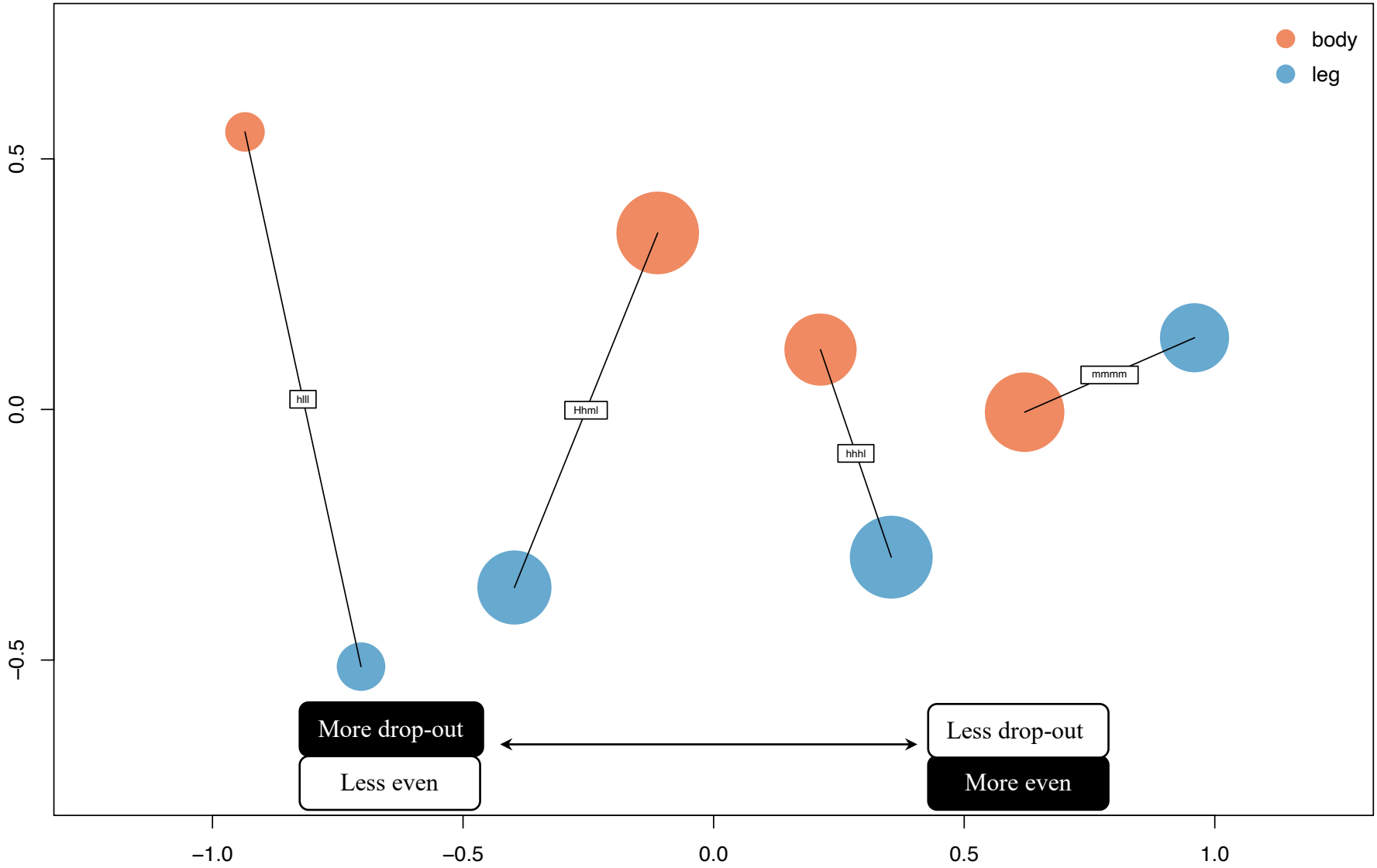
Effects of input-DNA absolute and relative concentrations on OTU recovery

Altering either the absolute (body/leg) or relative (*Hhml, hhhl, hlll,* and *mmmm*) input DNA concentrations created quantitative compositional differences in the OTU tables (Fig. 2). Soup *hlll*, with the most uneven distribution of input DNA concentrations, recovered the fewest OTUs (Fig. 2). The same effect can be seen by regressing the number of recovered OTUs on species evenness (Fig. S1). As expected, OTU size does a poor job of recovering information on input DNA amount per species (Fig. S2). Although there are positive relationships between OTU size and COI copy number (measured by qPCR), there is essentially no relationship between OTU size and genomic-DNA concentrations (a proxy of

Table 3. Species-recovery success by Begum filtering stringency level and PCR condition, using the *mmmm_body* soup. Recovered species are OTUs that match one of the 248 reference species at ≥97% similarity. Drop-outs are false-negatives and are defined as reference species that fail to be matched by an OTU at ≥97% similarity. Drop-ins are false-positives and are defined as OTUs that fail to match any reference species at ≥97% similarity. There are three levels of Begum filtering, starting with no filtering in the top section (accepting reads that are present in ≥1 PCR and ≥1 copy). Begum filtering reduces Drop-in frequencies by ~20-40 times (dark to light orange cells), at the cost of a small absolute rise in Drop-out frequency (grey to blue cells), provided that PCR conditions are optimal (PCRs A, B, E). With non-optimal PCR conditions (PCRs C, D, F, G, H: High T_a or cycle number, Touchdown), there is a trade-off: filtering to reduce false positives also increases false negatives (blue cells get darker on the lower right). See *Effects of Begum filtering and PCR condition* for details. Table **S_DROPS** shows the number of OTUs remaining after each major step of our pipeline. The dark box indicates best-performing combinations of Begum filtering and PCR conditions. Optimal T_a is 45.5°C, and high is 51.5°C. Optimal cycle number is 25, low is 21, and high is 30.

Begum filtering stringency	**PCR conditions**							
	Optimal Ta + Optimal cycle number		Optimal Ta + Low cycle number	High Ta + Optimal cycle number		Optimal Ta + High cycle number	Touchdown PCR	
Present in ≥1 PCR replicate & ≥1 read per PCR	**A**	**B**	**E**	**C**	**D**	**F**	**G**	**H**
Recovered species, Matched to Ref. (≥97% sim.)	240	239	238	239	237	237	228	231
Drop-outs	8	9	10	9	11	11	20	17
% Drop-out	**3%**	**4%**	**4%**	**4%**	**4%**	**4%**	**8%**	**7%**
Drop-ins	95	89	109	115	80	107	67	77
% Drop-in	**38%**	**36%**	**44%**	**46%**	**32%**	**43%**	**27%**	**31%**
Present in ≥2 PCR replicates & ≥4 reads per PCR	**A**	**B**	**E**	**C**	**D**	**F**	**G**	**H**
Recovered species, Matched to Ref. (≥97% sim.)	233	229	231	214	203	198	154	170
Drop-outs	15	19	17	34	45	50	94	78
% Drop-out	**6%**	**8%**	**7%**	**14%**	**18%**	**20%**	**38%**	**32%**
Drop-ins	5	5	7	3	2	3	3	3
% Drop-in	**2%**	**2%**	**3%**	**1%**	**1%**	**1%**	**1%**	**1%**
Present in ≥3 PCR replicates & ≥3 reads per PCR	A	B	E	C	D	F	G	H
Recovered species, Matched to Ref. (≥97% sim.)	225	226	232	195	190	180	126	137
Drop-outs	23	22	16	53	58	68	122	111
% Drop-out	**9%**	**9%**	**7%**	**21%**	**23%**	**27%**	**49%**	**45%**
Drop-ins	4	4	6	2	2	3	2	1
% Drop-in	**2%**	**2%**	**2%**	**1%**	**1%**	**1%**	**1%**	**0%**

Fig. 2. Non-metric multidimensional scaling (NMDS) ordination of the 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 (Table 3): optimum annealing temperature T_a (45.5 C) and optimum cycle number (25). Point size is scaled to the number of recovered OTUs. Species recovery is reduced in samples characterized by more uneven species frequencies (e.g. *hlll*) and, to a much lesser extent, lower absolute DNA input (leg).



biomass per species), which reflects the action of multiple species-specific biases along the metabarcoding pipeline (McLaren, Willis & Callahan 2019).

Taxonomic amplification bias

Optimal PCR conditions (T_a + cycle number) produce larger OTUs than do non-optimal PCR conditions (high T_a, high cycle number, and Touchdown), especially for Hymenoptera, Araneae, and Hemiptera (Fig. 3). These taxa are therefore at higher risk of failing to be detected by the Leray-Geller primers under sub-optimal PCR conditions.

Tag-bias test

We found no evidence for tag bias in PCR amplification. For instance, in PCRs A & B, pairs using the same tags (A1/B1, A2/B2, A3/B3) and pairs using different tags (e.g. A1/B2, A2/B1) both produced high Procrustes correlation coefficient values, and they differed by only 0.013 (=0.993-0.980, p=0.059, Fig. 4). At higher annealing temperatures T_a (PCRs C&D, G &H), where it is more plausible that tag sequences might aid primer annealing to a subset of template DNA, we still found no evidence for tag bias (Figs S3-4).

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, Fig. 2). We metabarcoded the soups under five different PCR conditions that varied annealing temperatures (T_a) and PCR cycles (Table 3), and we filtered the OTUs under different stringencies (Fig. 1, Table 3). We define high efficiency in metabarcoding as recovering most of sample's compositional and quantitative information, which in turn means that both drop-out and drop-in frequencies are low, that OTU size predicts input DNA amount per species, and that any drop-outs are spread evenly over the taxonomic range of the target taxon (here, Arthropoda).

Our results show that metabarcoding efficiency is higher 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 reduced when species DNA frequencies are uneven (Fig. 2, Fig. S1), and we found that OTU sizes are a poor predictor of input genomic DNA, which confirms that OTU size is a poor predictor of species frequencies (Fig. S2) (McLaren, Willis & Callahan 2019). Finally, we found no evidence for tag bias in PCR (Fig. 3, Figs S3-4).

Fig. 3. 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. 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 conditions) with A (optimal conditions), 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-Geller primers under sub-optimal PCR conditions. Shown here are the *mmmmbody* soups, at Begum filtering stringency ≥2 PCRs, ≥4 copies per PCR.

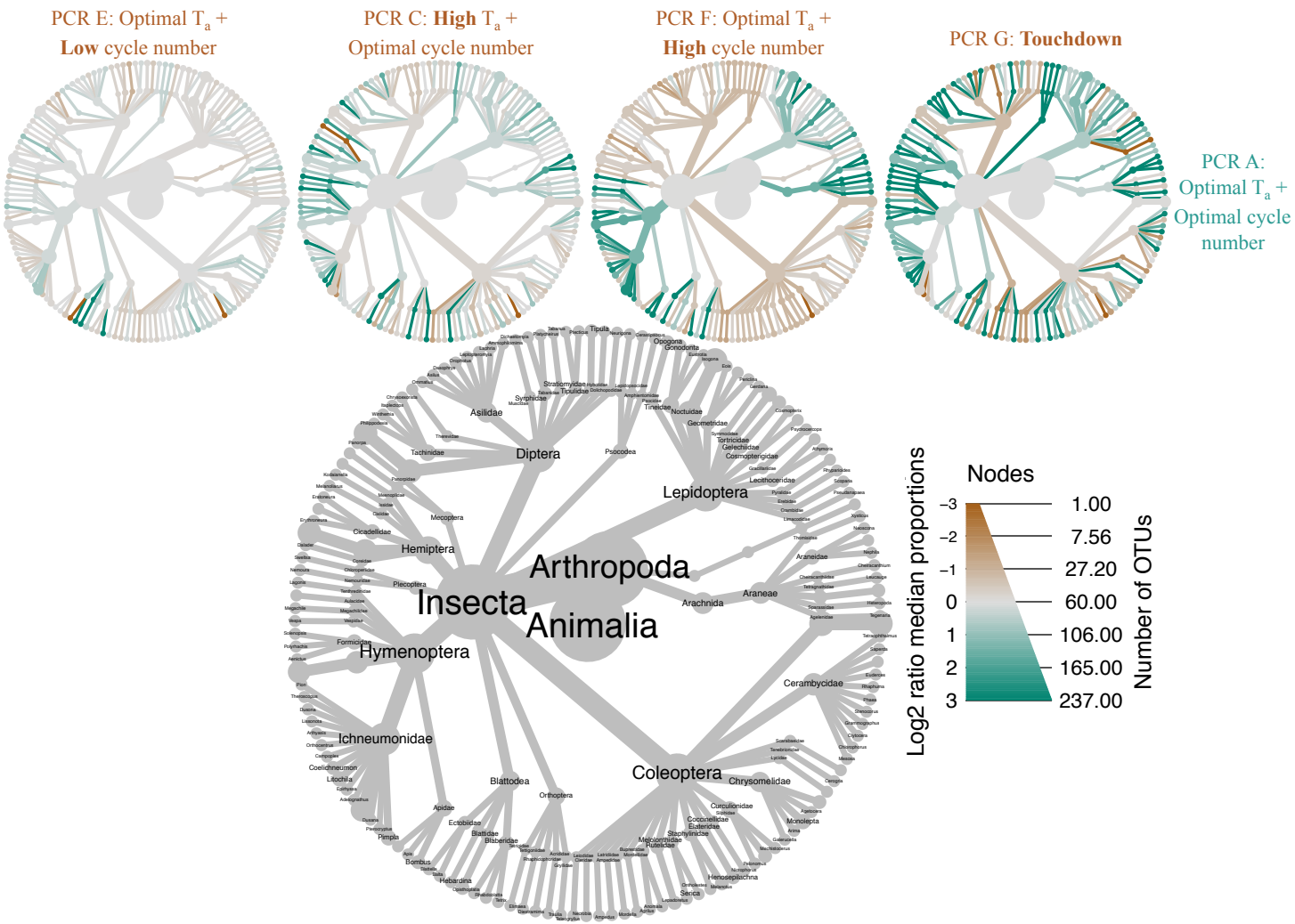
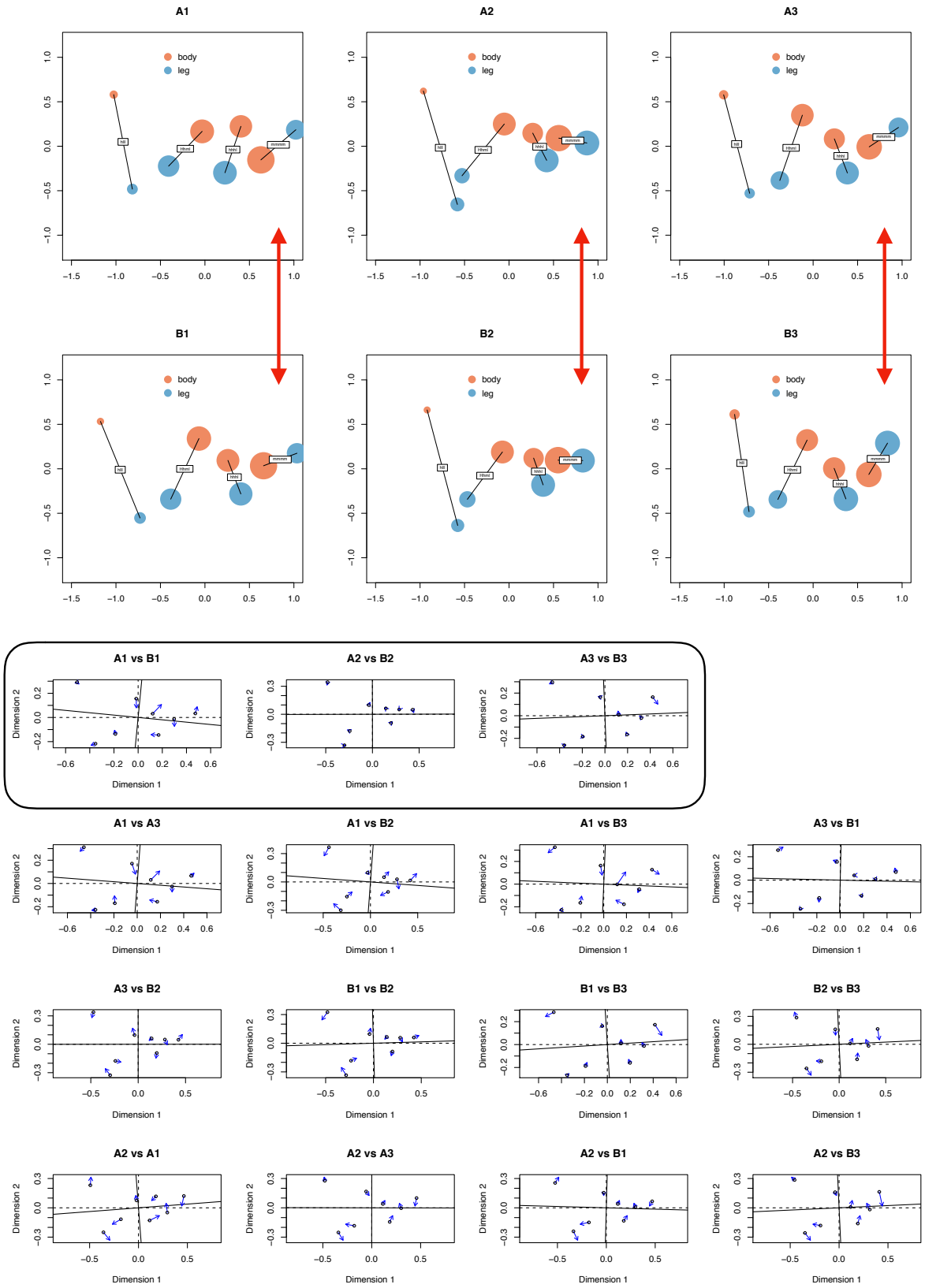


Fig. 4. Test for tag bias in the mock soups amplified at optimum annealing temperature T_a (45.5 °C) and optimum cycle number (25) (PCRs A and B). Top. PCRs A1 & B1, A2 & B2, and A3 & B3 are three replicate pairs (linked by red lines), and within each pair, the same eight tag sequences were used in PCR, and thus each pair should recover the same communities. **Bottom.** All pairwise Procrustes correlations of PCRs A and B. The top row (box) displays the three same-tag pairwise correlations. The other rows display all the different-tag pairwise correlations. If there is tag bias during PCR, the top row should show a greater degree of similarity. Mean correlations are not significantly different between same-tag and different-tag ordinations (Mean of same-tag Procrustes correlations: 0.993 ± 0.007 SD, n = 3. Mean of different-tag correlations: 0.980 ± 0.008 SD, n = 12. p=0.059). In Supplementary Information, we show the results for the high T_a treatment (PCRs C & D) and the Touchdown PCR treatment (PCRs G & H). Circle size is scaled to number of OTUs, and colors indicate tissue source (leg or body). Mock soups with the same concentration evenness (e.g. Hhml) are linked by lines.


A1
A2
A3
B1
B2
B3
body
leg
A1 vs B1
A2 vs B2
A3 vs B3
A1 vs A3
A1 vs B2
A1 vs B3
A3 vs B1
A3 vs B2
B1 vs B2
B1 vs B3
B2 vs B3
A2 vs A1
A2 vs A3
A2 vs B1
A2 vs B3
Dimension 1
Dimension 2

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

The *Begum* workflow co-designs the wet-lab and bioinformatic components (Fig. 1) to remove multiple sources of error (Table 1). Twin-tagging allows removal of tag jumps, which result in sequence-to-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 (Table 3). qPCR optimization reduces false negatives caused by PCR runaway, PCR inhibition, and annealing failure (Table 3, Fig. 4). Moderate dilution appears to be a better solution for inhibition than is increasing cycle number, since the latter increases drop-outs (Table 3). qPCR also allows extraction blanks to be screened for contamination. Size sorting (Elbrecht, Peinert & Leese 2017) should also reduce false negatives, but our finding that leg-only soups have more missing OTUs (Fig. 2) suggests that large insects should be replaced by heads not legs.

Begum filtering and complex positive controls

Increasing the stringency of *Begum* filtering reduces false-positive 'drop-ins' at the cost of increasing false-negative 'drop-outs', although fortunately, this trade-off is weakened under optimal PCR conditions (Table 3). To decide on a filtering stringency level, we therefore recommend using complex positive controls and to take into account one's study's aims. If the aim is to detect a particular taxon, like an invasive pest, it is better to set stringency low to reduce false-negatives, whereas if the aim is to generate data for an occupancy model, it is better to set stringency high to remove 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.

Future work

Our co-designed wet-lab and bioinformatic pipeline identifies and removes erroneous sequences while simultaneously reducing drop-out caused by PCR amplification bias. This contrasts with solutions, such as DADA2 (Callahan *et al.* 2016) and UNOISE2 (Edgar 2016), that use only sequence 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 and differences in total DNA mass across samples. Such corrections allow tracking of *within-species change across samples*, which means tracking how individual species change abundances along a time series or environmental gradient. However, spike-ins and UMIs cannot remove species biases in DNA-extraction and primer-binding efficiencies, and such biases make it highly error-prone to estimate *across-species frequencies within samples*, such as when we want to identify dominant diet components (e.g. Bell *et al.* 2019). Fortunately, methods for estimating within-sample species frequencies are being developed in metagenomic pipelines (Lang *et al.* 2019; Peel *et al.* 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).

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.

Data Availability

Supplementary Information includes a tutorial with a reduced dataset and simplified scripts (PCR B only, 253 MB). We have also archived the full dataset, with reference files, folder structure, output files, and scripts (to run the scripts, remove the output files listed in the DataDryad readme, ~13 GB). Both archives are are available for reviewer download at: https://datadryad.org/stash/share/_XecP4iDVnM2uFmFyCcEpV3bD58swl8lIuzc-2ZDMNc

The complete scripts are also kept updated on GitHub at https://github.com/dougwyu/BiodiversitySoupII.

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.

Bell, K.L., Burgess, K.S., Botsch, J.C., Dobbs, E.K., Read, T.D. & Brosi, B.J. (2019) Quantitative and qualitative assessment of pollen DNA metabarcoding using constructed species mixtures. *Molecular Ecology,* **28** (2)**,** 431-455. doi:10.1111/mec.14840.

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., Evans, A., Gilbert, M.T.P., Carvalho, G.R., Creer, S., Knapp, M., . . . de Bruyn, M. (2014) Environmental DNA for wildlife biology and biodiversity

monitoring. *Trends in Ecology & Evolution,* **29** (6)**,** 358-367. doi:10.1016/j.tree.2014.04.003.

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 Alex, Z.C., Wendy Monk, Teresita M.Porter, Royce Steeves, Erik Emilson, Nellie Gagne, Mehrdad Hajibabaei, Mélanie Roy, Donald J. Baird (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.

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,** e13042. doi: 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. 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.

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.

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. doi:10.1371/journal.pcbi.1005404.

Frøslev, T.G., Kjøller, 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. doi:10.1038/s41467-017-01312-x.

Geller, J., Meyer, C., Parker, M. & Hawk, H. (2013) Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. *Molecular Ecology Resources,* **13** (5)**,** 851-861. doi:10.1111/1755-0998.12138.

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. doi:10.1371/journal.pone.0017497.

Hering, D., Borja, A., Jones, J.I., Pont, D., Boets, P., Bouchez, 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. doi: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. doi:10.1371/journal.pone.0169431.

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

492 Ji, Y., Huotari, T., Roslin, T., Schmidt, N.M., Wang, J., Yu, D.W. & Ovaskainen, O. (2020)
493 SPIKEPIPE: A metagenomic pipeline for the accurate quantification of eukaryotic
494 species occurrences and intraspecific abundance change using DNA barcodes or
495 mitogenomes. *Molecular Ecology Resources,* **20** (1)**,** 256-267. doi:10.1111/1755-
496 0998.13057.

497 Lang, D., Tang, M., Hu, J. & Zhou, X. (2019) Genome-skimming provides accurate
498 quantification for pollen mixtures. *Molecular Ecology Resources,* **19** (6)**,** 1433-1446.
499 doi:10.1111/1755-0998.13061.

500 Lanzén, A., Lekang, K., Jonassen, I., Thompson, E.M. & Troedsson, C. (2016) High-
501 throughput metabarcoding of eukaryotic diversity for environmental monitoring of
502 offshore oil-drilling activities. *Molecular Ecology,* **25** (17)**,** 4392-4406.
503 doi:10.1111/mec.13761.

504 Leray, M., Ho, S.L., Lin, I.J. & Machida, R.J. (2018) MIDORI server: a webserver for
505 taxonomic assignment of unknown metazoan mitochondrial-encoded sequences using
506 a curated database. *Bioinformatics,* **34** (21)**,** 3753-3754.
507 doi:10.1093/bioinformatics/bty454.

508 Leray, M. & Knowlton, N. (2015) DNA barcoding and metabarcoding of standardized
509 samples reveal patterns of marine benthic diversity. *Proceedings of the National*
510 *Academy of Sciences of the United States of America,* **112** (7)**,** 2076-2081.
511 doi:10.1073/pnas.1424997112.

512 Leray, M., Yang, J.Y., Meyer, C.P., Mills, S.C., Agudelo, N., Ranwez, V., . . . Machida, R.J.
513 (2013) A new versatile primer set targeting a short fragment of the mitochondrial COI
514 region for metabarcoding metazoan diversity: application for characterizing coral reef
515 fish gut contents. *Frontiers in Zoology,* **10** 34. doi:10.1186/1742-9994-10-34.

516 McLaren, M.R., Willis, A.D. & Callahan, B.J. (2019) Consistent and correctable bias in
517 metagenomic sequencing experiments. *Elife,* **8** e46923. doi:10.7554/eLife.46923.

518 McMurdie, P.J. & Holmes, S. (2013) phyloseq: an R package for reproducible interactive
519 analysis and graphics of microbiome census data. *PLoS One,* **8** (4)**,** e61217.
520 doi:10.1371/journal.pone.0061217.

521 Murray, D.C., Coghlan, M.L. & Bunce, M. (2015) From benchtop to desktop: important
522 considerations when designing amplicon sequencing workflows. *PLoS One,* **10** (4)**,**
523 e0124671. doi:10.1371/journal.pone.0124671.

524 O'Donnell, J.L., Kelly, R.P., Lowell, N.C. & Port, J.A. (2016) Indexed PCR Primers Induce
525 Template-Specific Bias in Large-Scale DNA Sequencing Studies. *PLoS One,* **11** (3)**,**
526 e0148698. doi:10.1371/journal.pone.0148698.

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

531 Peel, N., Dicks, L.V., Heavens, D., Percival-Alwyn, L., Cooper, C., Clark, M.D., . . . Yu,
532 D.W. (2019) Semi-quantitative characterisation of mixed pollen samples using
533 MinION sequencing and Reverse Metagenomics (RevMet). *Methods in Ecology and*
534 *Evolution,* **10** (10)**,** 1690-1701. doi:10.1111/2041-210X.13265.

535 Piper, A.M., Batovska, J., Cogan, N.O.I., Weiss, J., Cunningham, J.P., Rodoni, B.C. &
536 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.

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.

Team, R.C. (2018) R: A Language and Environment for Statistical Computing, Vienna, Austria. *R Foundation for Statistical Computing*. https://www.R-project.org/

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

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

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. doi:10.1111/ddi.12972.

Yoccoz, N.G. (2012) The future of environmental DNA in ecology. *Molecular Ecology,* **21** (8)**,** 2031-2038. doi: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. doi:10.1111/j.2041-210X.2012.00198.x.

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. doi:10.1186/s13104-016-2064-9.

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

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

582

Table 1. Four classes of metabarcoding errors and their causes. Not included are software bugs, general laboratory and field errors like mislabeling, DNA degradation, and sampling biases or inadequate effort.

Main Errors	Possible causes	References
False positives ('Drop-ins,' 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	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 the nature of the error)	Intra-specific variability across the marker leading to multiple OTUs with different taxonomic assignments	Clarke *et al.* 2014
	Incomplete reference databases	

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*). The evenness of DNA concentrations in each mock soup is summarized by the Shannon index, in which higher values indicate a more even distribution of input DNA concentrations.

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	*hlll*	0	61	0	187	248	4.08
	Hhml	50	75	62	61	248	4.56
	hhhl	0	187	0	61	248	5.17
	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	*hlll*	0	71	0	177	248	4.13
	Hhml	69	63	63	53	248	4.21
	hhhl	0	195	0	53	248	5.04
	mmmm	0	0	238	10	248	5.32

Table 3. Species-recovery success by Begum filtering stringency level and PCR condition, using the *mmmm_body* soup. Recovered species are OTUs that match one of the 248 reference species at ≥97% similarity. Drop-outs are false-negatives and are defined as reference species that fail to be matched by an OTU at ≥97% similarity. Drop-ins are false-positives and are defined as OTUs that fail to match any reference species at ≥97% similarity. There are three levels of Begum filtering, starting with no filtering in the top section (accepting reads that are present in ≥1 PCR and ≥1 copy). Begum filtering reduces Drop-in frequencies by ~20-40 times (dark to light orange cells), at the cost of a small absolute rise in Drop-out frequency (grey to blue cells), provided that PCR conditions are optimal (PCRs A, B, E). With non-optimal PCR conditions (PCRs C, D, F, G, H: High T_a or cycle number, Touchdown), there is a trade-off: filtering to reduce false positives also increases false negatives (blue cells get darker on the lower right). See *Effects of Begum filtering and PCR condition* for details. Table **S_DROPS** shows the number of OTUs remaining after each major step of our pipeline. The dark box indicates best-performing combinations of Begum filtering and PCR conditions. Optimal T_a is 45.5°C, and high is 51.5°C. Optimal cycle number is 25, low is 21, and high is 30.

Begum filtering stringency	**PCR conditions**							
	Optimal Ta + Optimal cycle number		Optimal Ta + Low cycle number	High Ta + Optimal cycle number		Optimal Ta + High cycle number	Touchdown PCR	
Present in ≥1 PCR replicate & ≥1 read per PCR	**A**	**B**	**E**	**C**	**D**	**F**	**G**	**H**
Recovered species, Matched to Ref. (≥97% sim.)	240	239	238	239	237	237	228	231
Drop-outs	8	9	10	9	11	11	20	17
% Drop-out	**3%**	**4%**	**4%**	**4%**	**4%**	**4%**	**8%**	**7%**
Drop-ins	95	89	109	115	80	107	67	77
% Drop-in	**38%**	**36%**	**44%**	**46%**	**32%**	**43%**	**27%**	**31%**
Present in ≥2 PCR replicates & ≥4 reads per PCR	**A**	**B**	**E**	**C**	**D**	**F**	**G**	**H**
Recovered species, Matched to Ref. (≥97% sim.)	233	229	231	214	203	198	154	170
Drop-outs	15	19	17	34	45	50	94	78
% Drop-out	**6%**	**8%**	**7%**	**14%**	**18%**	**20%**	**38%**	**32%**
Drop-ins	5	5	7	3	2	3	3	3
% Drop-in	**2%**	**2%**	**3%**	**1%**	**1%**	**1%**	**1%**	**1%**
Present in ≥3 PCR replicates & ≥3 reads per PCR	A	B	E	C	D	F	G	H
Recovered species, Matched to Ref. (≥97% sim.)	225	226	232	195	190	180	126	137
Drop-outs	23	22	16	53	58	68	122	111
% Drop-out	**9%**	**9%**	**7%**	**21%**	**23%**	**27%**	**49%**	**45%**
Drop-ins	4	4	6	2	2	3	2	1
% Drop-in	**2%**	**2%**	**2%**	**1%**	**1%**	**1%**	**1%**	**0%**

Fig. 1. Schematic of study. A. Twin-tagged primers with heterogeneity spacers (above) and final amplicon structure (below). **B.** Each mock soup (a 'sample') is PCR-amplified three times (#1, #2, #3), each with a different tag, following the Begum strategy. There are eight mock soups (*mmmm/hhhl/Hhml/hlll* X body/leg), and PCR replicates #1 from each of the eight mock soups are pooled into the first amplicon pool, PCR replicates #2 are pooled into the second amplicon pool, and so on. The setup in B. was itself replicated 8 times for the 8 PCR conditions (A-H), which generated (3 X 8 =) 24 pools in total. **C.** Key steps of the Begum bioinformatic pipeline. For clarity, primers and heterogeneity spacers not shown.

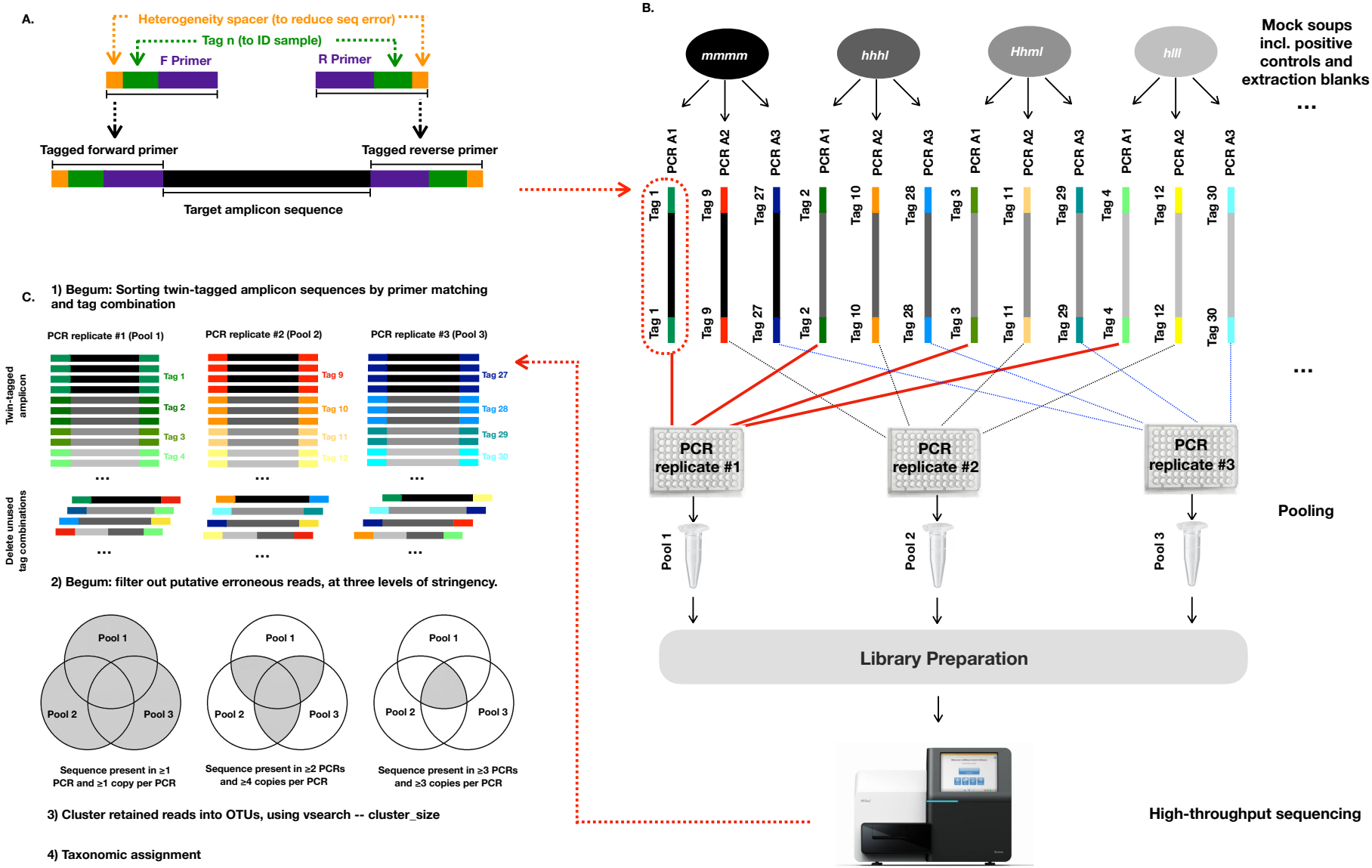


Fig. 2. Non-metric multidimensional scaling (NMDS) ordination of the eight mock soups, which differ in absolute (Body/Leg) and relative (*Hhml*, *hhhl*, *hlll*, and *mmmm*) DNA concentrations of the input species (Table 2). Shown here is the output from the PCR A condition (Table 3): optimum annealing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

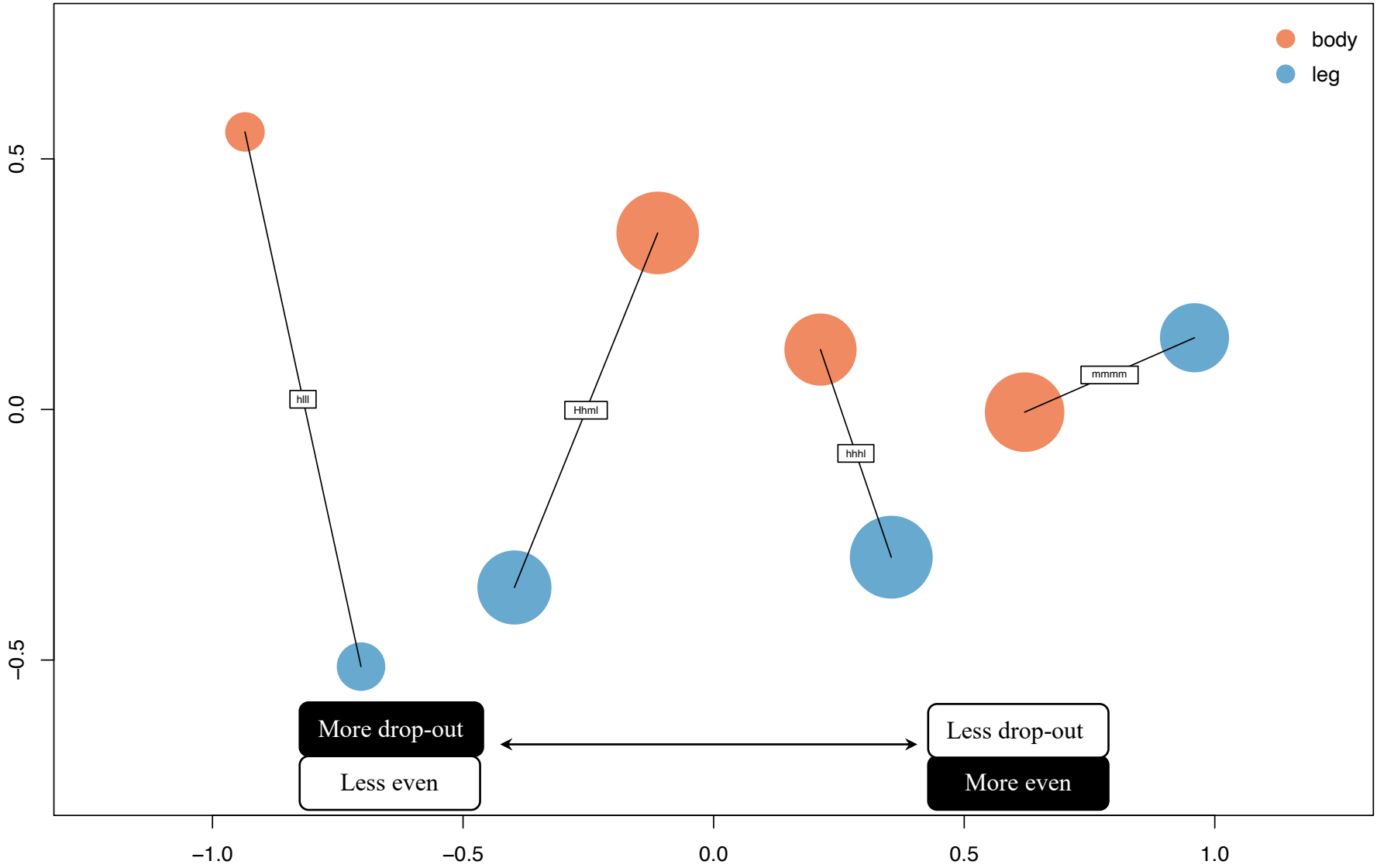


Fig. 3. Test for tag bias in the mock soups amplified at optimum annealing temperature T_a (45.5 °C) and optimum cycle number (25) (PCRs A and B). Top. PCRs A1 & B1, A2 & B2, and A3 & B3 are three replicate pairs (linked by red lines), and within each pair, the same eight tag sequences were used in PCR, and thus each pair should recover the same communities. **Bottom.** All pairwise Procrustes correlations of PCRs A and B. The top row (box) displays the three same-tag pairwise correlations. The other rows display all the different-tag pairwise correlations. If there is tag bias during PCR, the top row should show a greater degree of similarity. Mean correlations are not significantly different between same-tag and different-tag ordinations (Mean of same-tag Procrustes correlations: 0.993 ± 0.007 SD, n = 3. Mean of different-tag correlations: 0.980 ± 0.008 SD, n = 12. p=0.059). In Supplementary Information, we show the results for the high T_a treatment (PCRs C & D) and the Touchdown PCR treatment (PCRs G & H). Circle size is scaled to number of OTUs, and colors indicate tissue source (leg or body). Mock soups with the same concentration evenness (e.g. Hhml) are linked by lines.

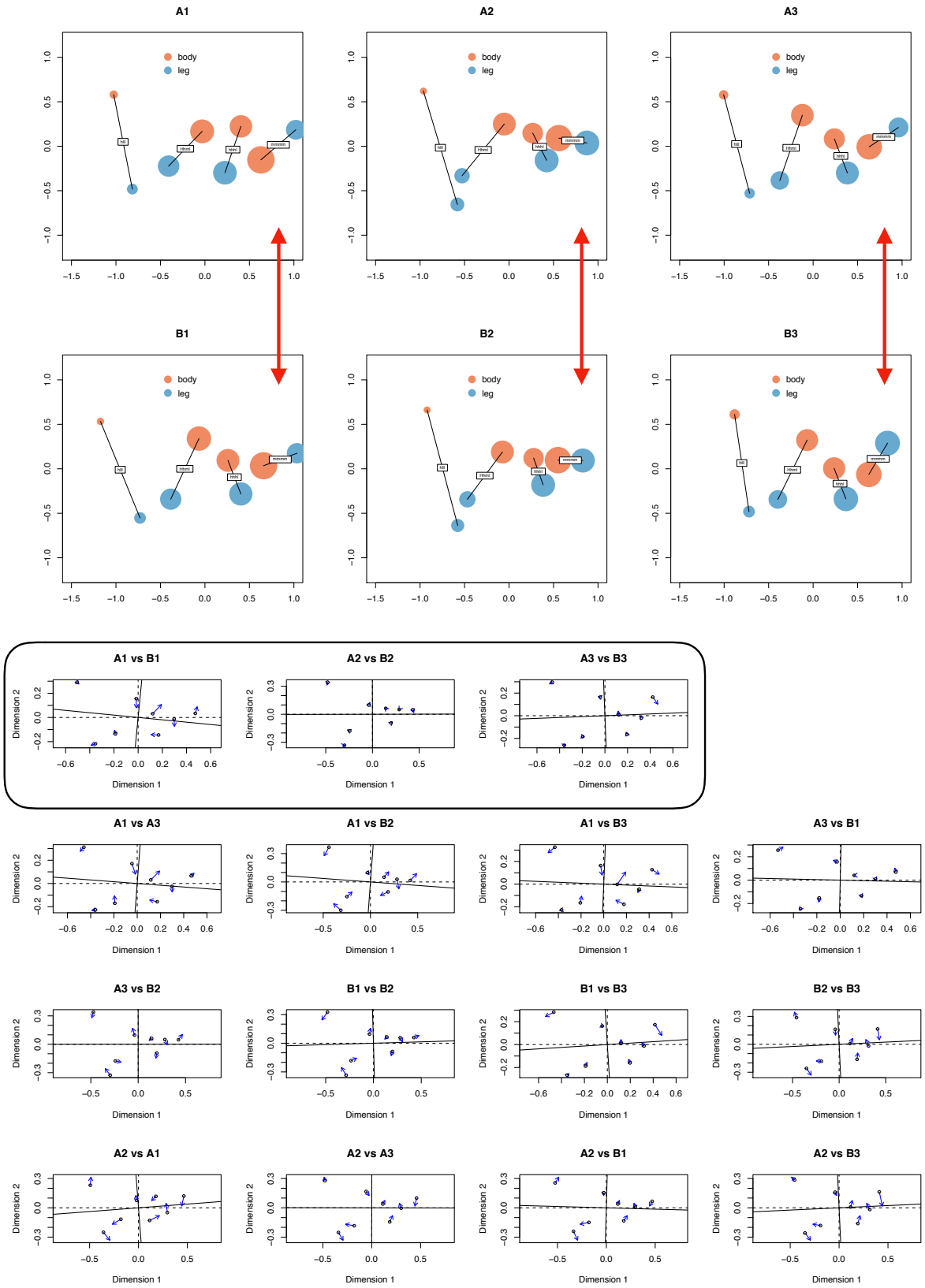

A1
A2
A3
B1
B2
B3
body
leg
A1 vs B1
A2 vs B2
A3 vs B3
A1 vs A3
A1 vs B2
A1 vs B3
A3 vs B1
A3 vs B2
B1 vs B2
B1 vs B3
B2 vs B3
A2 vs A1
A2 vs A3
A2 vs B1
A2 vs B3
Dimension 1
Dimension 2

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. 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 conditions) with A (optimal conditions), 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-Geller primers under sub-optimal PCR conditions. Shown here are the *mmmmbody* soups, at Begum filtering stringency ≥2 PCRs, ≥4 copies per PCR.

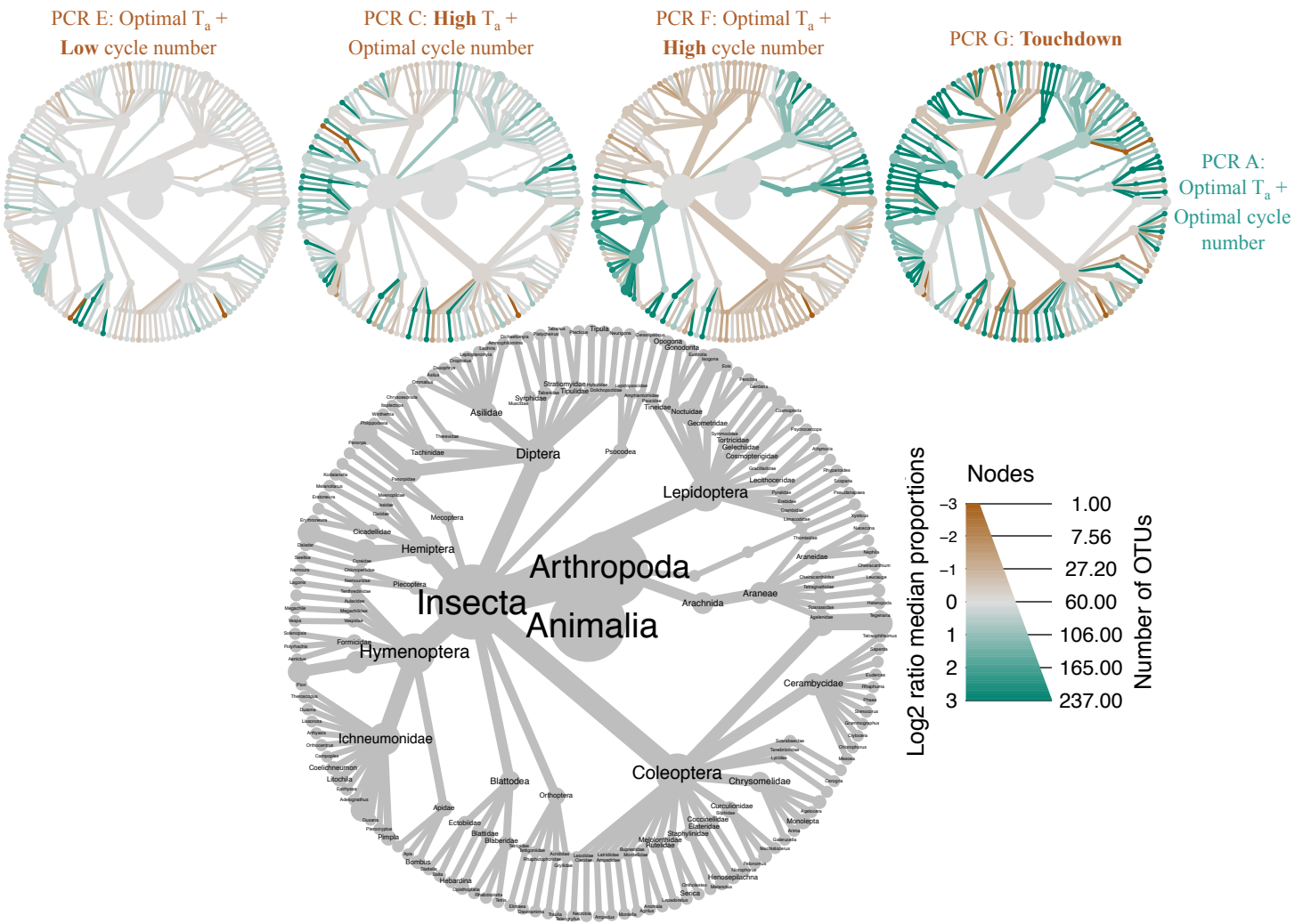


Fig. S1. Effect of species evenness on OTU recovery. More input species are recovered with a more even distribution of input DNA (Table 2). Shown here are the results from the Begum filtering stringency of ≥2 PCR replicates & ≥4 reads. Summary statistics from mixed-effects model `lme4::lmer(OTUs~Evenness+(1|PCR)`. R^2 reflects the fixed-effect contribution only (`MuMin::r.squaredGLMM`).

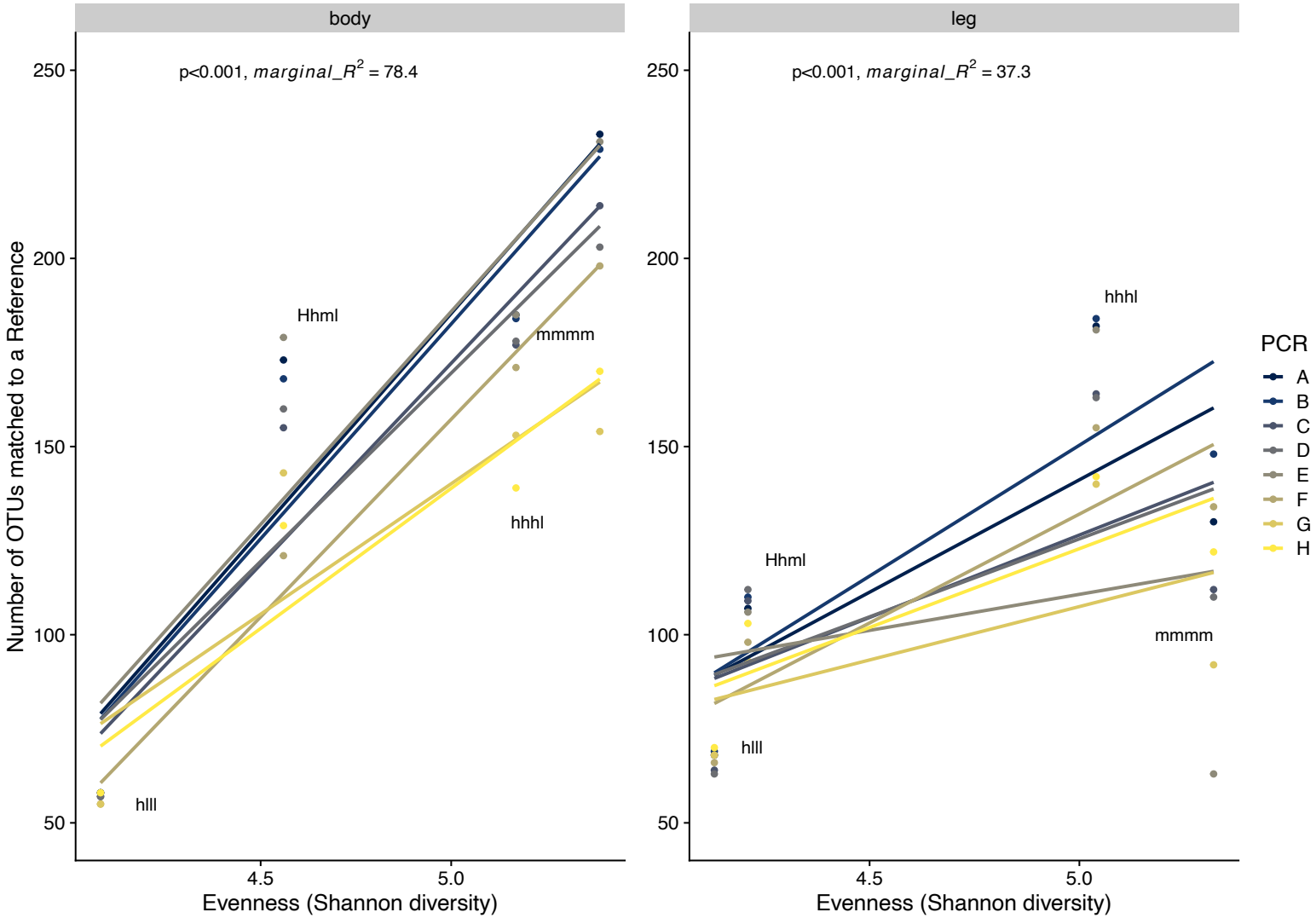


Fig. S2. Recovery of across-species abundance from read number. Each data point is one input species; fitted lines are linear regressions of input DNA quantity (**left**: mitochondrial COI; **right**: genomic DNA) on OTU size (number of reads). OTU size predicts input mitochondrial COI concentrations per species (left) somewhat better than it predicts input genomic DNA concentrations, but importantly, the relationship varies with species frequency distribution and absolute DNA amount (leg vs. body). Dataset from PCR A after Begum filtering stringency of ≥2 PCR replicates & ≥4 reads (Table 3). The *mmmm* soup was omitted because it has no variance.

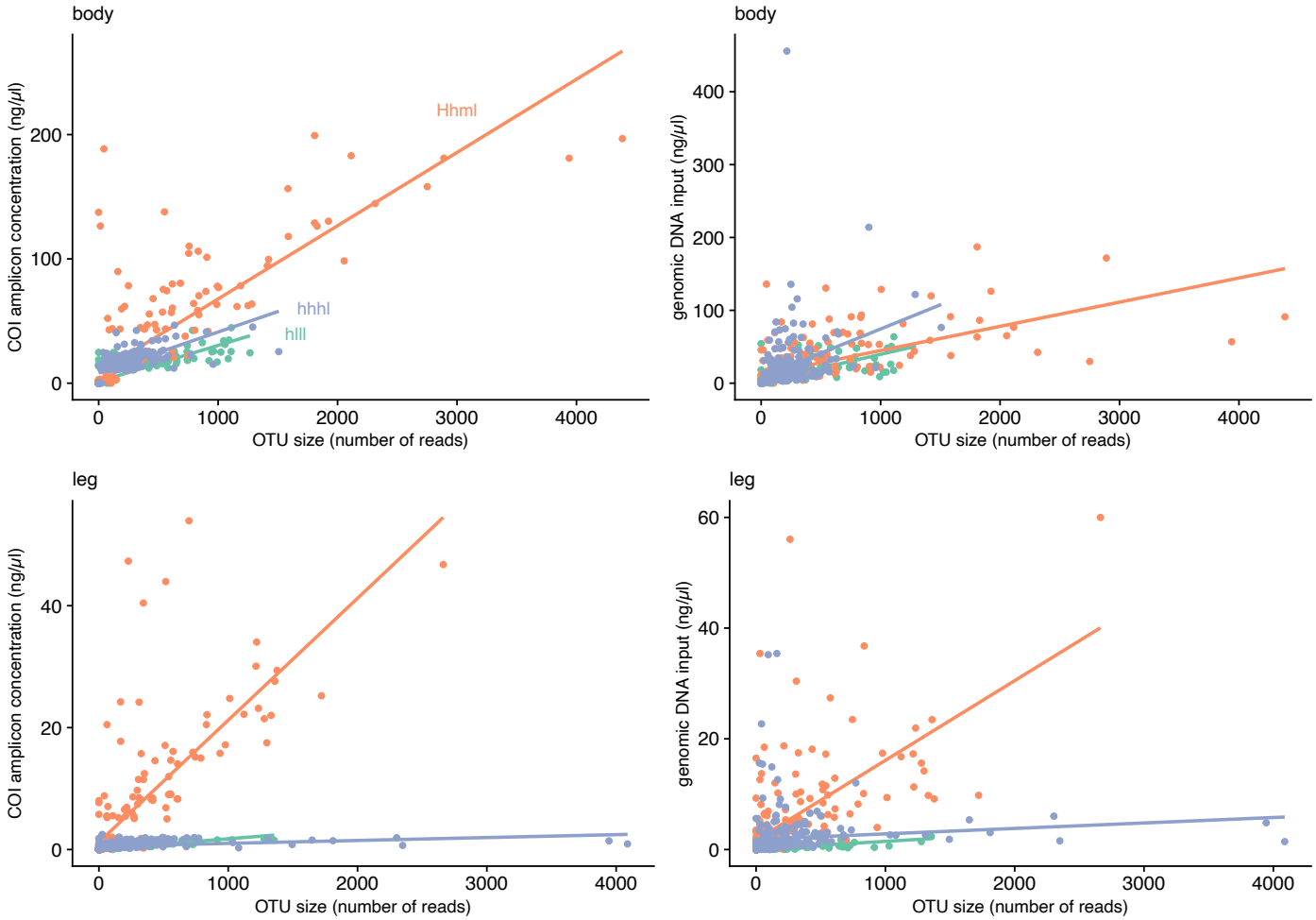


Fig. S3. Test for tag bias in the mock soups amplified at high annealing temperature T_a (51.5 °C) and optimum cycle number (25) (PCRs C and D). Top. PCRs C1 & D1, C2 & D2, and C3 & D3 are three replicate pairs (linked by red lines), and within each pair, the same eight tag sequences were used in PCR, and thus each pair should recover the same communities. **Bottom.** All pairwise Procrustes comparisons of PCRs C and D. The top row (box) displays the three same-tag pairwise correlations. The other rows display all the different-tag pairwise correlations. If there is tag bias during PCR, the top row should show a greater degree of similarity. Mean correlations are not significantly different between same-tag and different-tag ordinations (Mean of same-tag Procrustes correlations: 0.989 ± 0.010 SD, n = 3. Mean of different-tag correlations: 0.980 ± 0.012 SD, n = 12. $p<0.067$). Circle size is scaled to number of OTUs, and colors indicate tissue source (leg or body). Mock soups with the same concentration evenness (e.g. Hhml) are linked by lines.

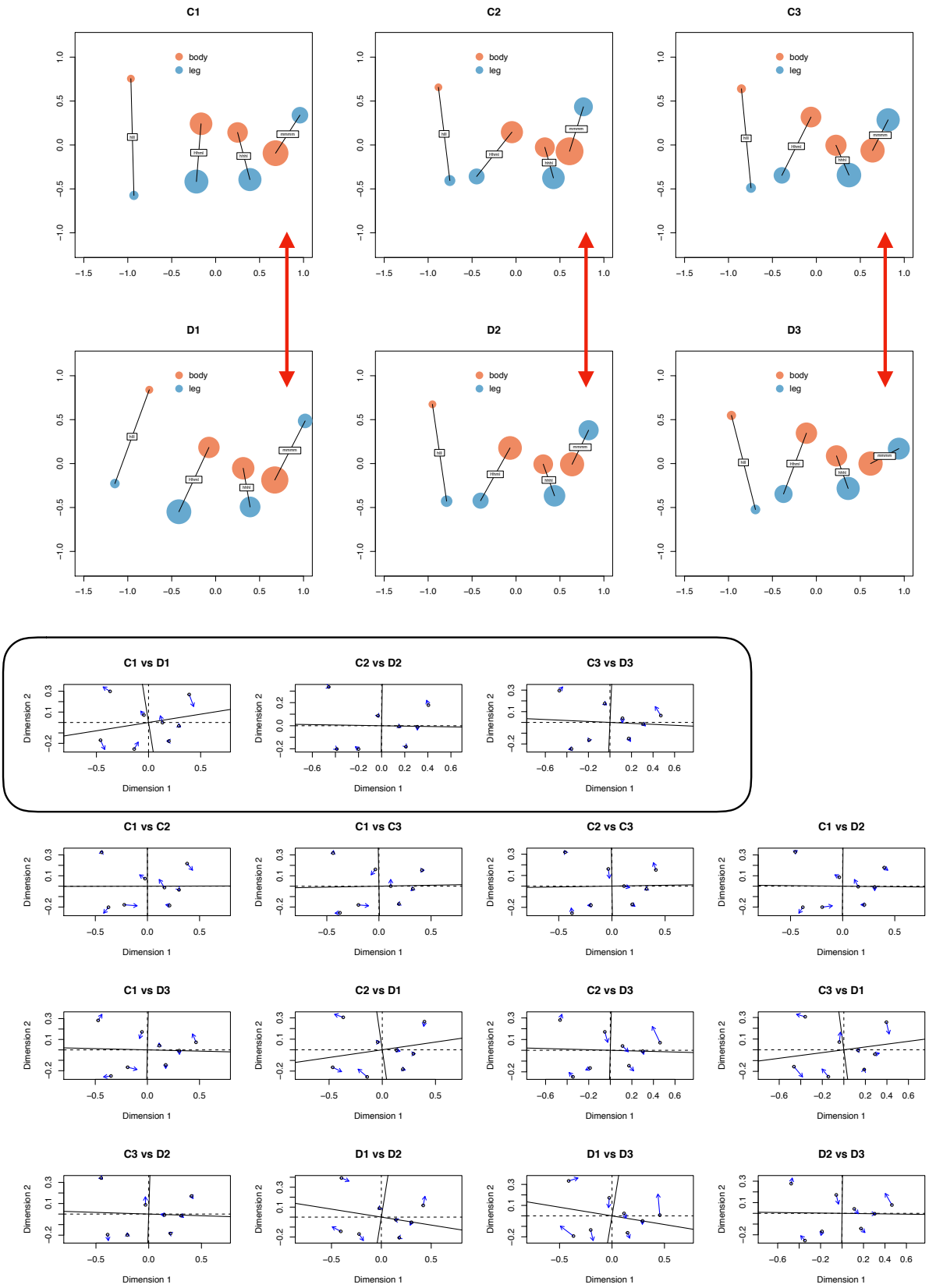
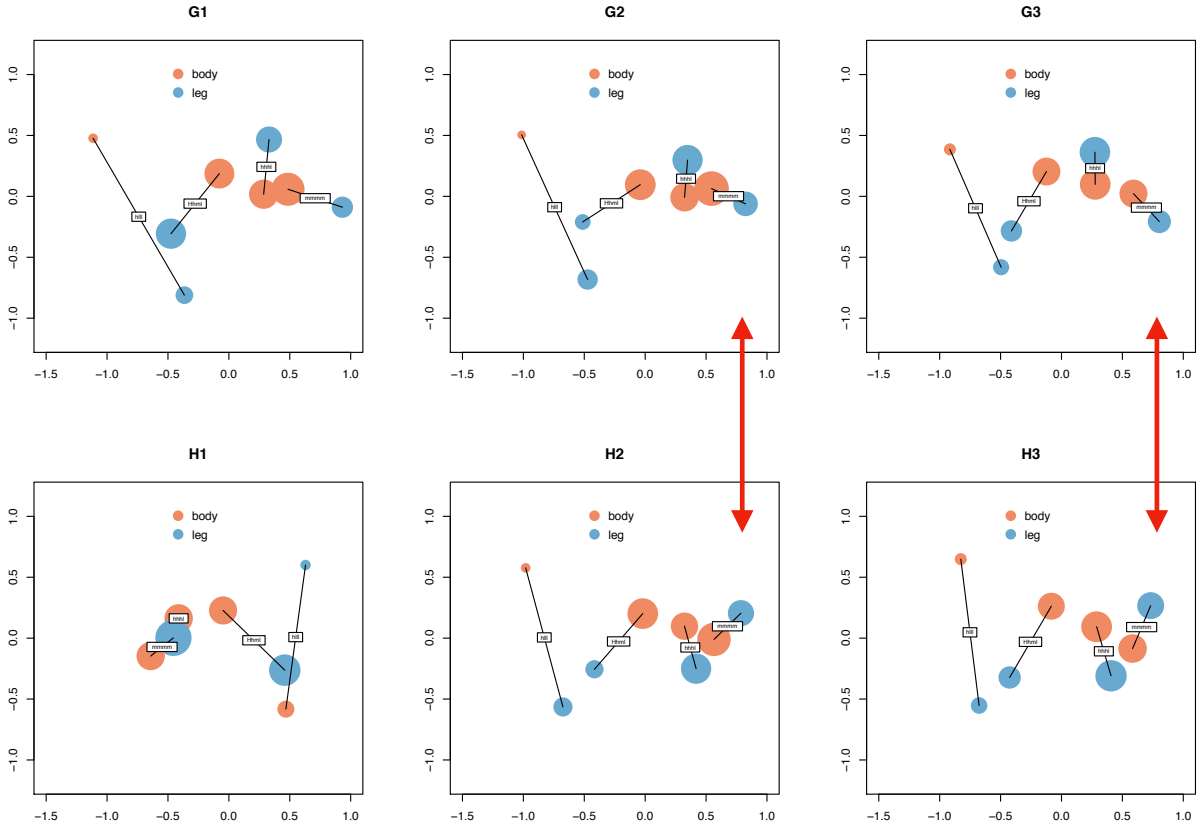
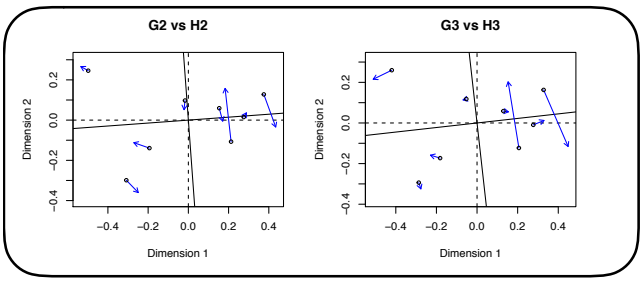
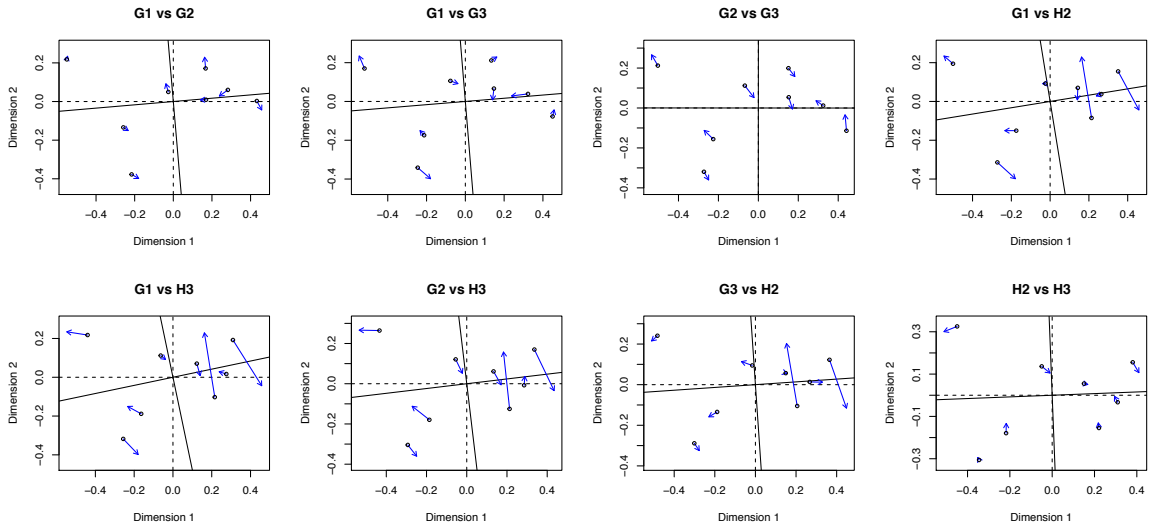

C1
C2
C3
D1
D2
D3
body
leg
C1 vs D1
C2 vs D2
C3 vs D3
C1 vs C2
C1 vs C3
C2 vs C3
C1 vs D2
C1 vs D3
C2 vs D1
C2 vs D3
C3 vs D1
C3 vs D2
D1 vs D2
D1 vs D3
D2 vs D3
Dimension 1
Dimension 2

Fig. S4. Test for tag bias in the mock soups amplified with the Touchdown PCR protocol (G and H). Top. PCRs G2 & H2, and G3 & H3 are two replicate pairs (linked by red lines), and within each pair, the same eight tag sequences were used in PCR, and thus each pair should recover the same communities. Pair G1 & H1 cannot be compared because H1 is missing one sample. **Bottom.** All pairwise Procrustes comparisons of PCRs G and H. The top row (box) displays the two pairwise comparisons between the technical replicates. The other rows display all the different-tag pairwise correlations. If there is tag bias during PCR, the top row should show a greater degree of similarity. Mean correlations are not significantly different between same-tag and different-tag ordinations, but we note that only two same-tag comparisons could be made (Mean of same-tag Procrustes correlations: 0.910 ± 0.038 SD, n = 2. mean correlation of different-tag correlations: 0.943 ± 0.050 SD, n = 6. p=0.419). Circle size is scaled to number of OTUs, and colors indicate tissue source (leg or body). Mock soups with the same concentration evenness (e.g. Hhml) are linked by lines.


G1
G2
G3
H1
H2
H3
body
leg


G2 vs H2
G3 vs H3
Dimension 1
Dimension 2


G1 vs G2
G1 vs G3
G2 vs G3
G1 vs H2
G1 vs H3
G2 vs H3
G3 vs H2
H2 vs H3
Dimension 1
Dimension 2