Extracting abundance information from   
DNA-based data

**Supplementary Information: Unabridged Methods**

Mingjie Luo1,2, Yinqiu Ji1, David Warton3,4, Douglas W. Yu1,5,6,\*

1 State Key Laboratory of Genetic Resources and Evolution and Yunnan Key Laboratory of Biodiversity and Ecological Security of Gaoligong Mountain, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, China 650223

2 Kunming College of Life Sciences, University of Chinese Academy of Sciences, Kunming, China

3 School of Mathematics and Statistics, UNSW Sydney, NSW 2052, Australia

4 Evolution and Ecology Research Centre, UNSW Sydney, NSW 2052, Australia

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

5 School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk, UK NR47TJ

\*Corresponding author: dougwyu@mac.com

## 1 Mock soup construction

*1.1 Input species*. – We used Malaise traps to collect 286 arthropods in Kunming, China (25°8’23” N, 102°44’17” E). DNA was extracted from each individual using the DNeasy Blood & Tissue Kit (Qiagen GmbH, Germany). Mean genomic DNA per species concentration was quantified from three replicates using PicoGreen fluorescent dye. We DNA-barcoded the individuals using the Folmer primer pair LCO1490 5’- GGTCAACAAATCATAAAGATATTGG -3’, and HC02198 (5’- TAAACTTCAGGGTGACCAAAAAATCA -3’) (Folmer et.al 1994) , with PCR parameters of initial denaturation at 95°C for 3 minutes, followed by 34 cycles of denaturation at 94°C for 30 seconds, annealing at 50°C for 30 seconds and extension at 72°C for 1 minute, a final extension at 72° C for 5 minutes, and a cooldown to 4°C. After the 658-bp COI sequences were cut to 313 bp based on our metabarcoding primers (see *4 Primer design*), 286 arthropods were clustered into 168 OTUs at 97% similarity. To create mock soups, we selected 52 individuals with genomic DNA >20 ng/µl, representing 52 OTUs.

*1.2 OTU quantification*. – We quantified COI concentrations of the 52 OTUs using the mean of three qPCRs with the *Leray-FolDegenRev* primer pair (Leray et al., 2013; Yu et al., 2012). First, to create a standard DNA curve, we used a purified and sequenced COI amplicon (Lepidoptera Bombycidae *Bombyx mori*) and quantified its concentration with a Qubit, from which we calculated COI copy number from the COI concentration (ng/µl), molar mass (233530.15 g/mol), and the Avogadro constant.

(Eq. 1)

This was 0.5X serially diluted into six concentration levels, as follows:

1. Start by diluting the above PCR amplicon to 100x by transferring 5 µl of amplicon to 495 µl of Ultragrade water. Vortex for 20 sec. Label as Tube 1.
2. Label 5 tubes from 2 to 6 and add 250 µl of water into each tube.
3. Transfer 250 µl from Tube 1 to Tube 2, and vortex for 20 sec.
4. Transfer 250 µl from Tube 2 to Tube 3, and vortex for 20 sec.
5. Continue the dilution series to Tube 6.
6. Use Tubes 1-6 as the standard DNA gradient for qPCR.

For each of the 52 OTUs, we used qPCR to quantify COI concentration (copy/µl). qPCRs were performed in a total volume of 20 µl with 10 µl of TAKARA SYBR premix Ex Taq (TaKaRa Biosystems, Dalian, China), 0.4 µl of Rox, 0.8 µl each of 10 µM of Leray and FolDegenRev primers (synthesized by Invitrogen, Shanghai, China), 2 µl of template DNA, and 6 µl of PCR-grade water. The qPCR conditions were 30 sec of initial denaturation at 95°C, followed by 35 cycles of denaturation at 95°C for 10 sec, 30 sec of annealing at 50°C, and 30 sec of extension at 72°C. Each OTU was quantified twice.

Chart, line chart

Description automatically generatedChart

Description automatically generated

**Figure S1**. Left. qPCR standard curve (CT ~ input-COI concentration). Right. qPCR amplification curves, with the 12 (= 2 replicates x 6 concentrations) red lines representing the standard-DNA curves. The other colours represent the 52 OTUs, each amplified twice. All OTU curves lie within the bounds of the standard-DNA curves, allowing interpolation of absolute COI concentrations.

*1.3 Creation of mock soups*. – Each mock soup was constructed with equal masses of purified genomic DNA from 52 OTUs. Input genomic DNA masses of each OTU from soup 1 to soup 7 decreased by 0.8X to create seven mock soups of differing absolute abundances. The individual-OTU genomic DNA masses for the seven mock soups were 61, 48.8, 39, 31.2, 25, 20, and 16 ng and were independently added to the mock soups. Each of the seven soups was made in triplicate (ntot = 21).

*2 Preparation of Malaise-trap samples*

2.1 *Malaise trap sampling*. – 244 Malaise-trap samples from 96 sites, using 99.9% ethanol as the trapping liquid, were collected in and near a 384 km2 forested landscape containing the H.J. Andrews Experimental Forest (44.2° N, 122.2° W), Oregon, United States in July 2018. Traps were left for 7 non-rainy days. To equalize biomass across individuals, we only kept the heads of large individuals (body lengths >2 cm) and then transferred the samples to fresh 99.9% ethanol to store at room temperature until extraction.

2.2 *DNA extraction*. – The samples were air dried individually on filter papers for less than an hour and then transferred to 50 ml tubes or 5 ml tubes according to sample volume. The samples were then weighed. DNA was non-destructively extracted by soaking the samples in lysis buffer, using the protocol from Ji et al. (2020) and Nielsen et al. (2019). In short, we added lysis buffer to each sample at a 5 ml:1 mg volume:air-dried-mass ratio and incubated the samples in an incubator at 56°C, shaking for 60 hrs. The recipe to make 50 ml of lysis buffer is 2 ml of 1M Tris-HCl buffer (PH 8.0), 1 ml of 5M sodium chloride, 10 ml of 10% Sodium Dodecyl Sulfate, 150 µl of 1M calcium chloride, 34.225 ml of PCR-grade water, 2 ml of 1M Dithiothreitol, and 625 µl of Proteinase K solution (20 mg/ml). After extraction, visual checks confirmed that the extracted insects largely retained their body shapes, although they became darker. Note that non-destructive DNA extraction increases the effects of species biases since species with hard exoskeletons likely release less tissue.

For this study, we selected seven samples spread over the study area. After completion of lysis, we serially diluted the 7 samples by using 0.7X lysis buffer volume (500 µl, 350 µl, 245 µl, 171.5 µl, 120 µl and 84 µl) to create six soups per sample (ntot = 42). We used QIAquick PCR purification kit (Qiagen GmbH, Germany) by following the manufacturer instructions to purify lysis buffer on one spin column per soup.

## 3 Adding spike-in DNA

*3.1 Spike-in DNA*. – For our spike-ins, we used three insect species from China, none of which is expected to appear in the Oregon Malaise-trap samples. (An alternative to our protocol is to use one or more synthetic, random DNA sequences (Tkacz et al., 2018)). Each of our three spike-ins is represented by a 658-bp COI fragment (Tables S1, S2) with primer binding sites that match the Folmer primers HCO2198 and LCO1490. For long-term storage, we inserted the COI fragments as plasmids into monoclonal bacteria. Plasmids were extracted using TIANprep Mini Plasmid Kit (Beijing, China) following manufacturer’s instructions.

qPCR was used to quantify the three spike-in DNA COI concentrations (ng/µl) using TAKARA SYBR Premix Ex Taq (TaKaRa Bio, Japan). COI copy numbers of the three spike-in DNA were calculated using equation 1.

In our plan, we will mix the three spike-ins in a ratio of 1:2:4 (Bombycidae:Elateridae:Mordellidae), to check for error and degradation; the reads from the three spike-ins should be found in the ratio of 1:2:4, but if, for instance, one of the spike-ins is not at the correct ratio with the other two, we can omit that species’ reads and recalculate the spike-in correction.

|  |  |
| --- | --- |
| species | full COI barcode sequences |
| Lepidoptera Bombycidae *Bombyx mori* | AACATTATATTTTATTTTTGGTATTTGATCAGGAATAATTGGAACATCTTTAAGACTTTTAATTCGAGCTGAATTAGGAAATCCAGGATCATTAATTGGAGATGATCAAATTTATAATACTATTGTAACAGCACATGCTTTTATTATAATTTTTTTTATAGTTATACCTATTATAATTGGAGGATTTGGAAATTGATTAGTTCCTCTTATACTAGGAGCACCAGATATAGCATTCCCACGAATAAATAATATAAGATTTTGACTCCTACCCCCCTCCCTTATATTATTAATTTCAAGAAGAATTGTAGAAAATGGTGCAGGAACAGGATGAACAGTTTACCCCCCACTTTCATCTAATATCGCACATAGAGGAAGATCCGTAGATCTTGCTATTTTTTCACTACATTTAGCAGGTATTTCATCAATTATAGGAGCAATTAATTTTATTACAACAATAATTAATATACGATTAAATAATATATCATTTGATCAATTACCCTTATTTGTATGAGCTGTAGGGATTACAGCATTTTTATTATTATTATCACTACCTGTTTTAGCTGGAGCTATTACAATATTATTAACAGATCGAAACTTAAATACATCATTTTTTGATCCTGCTGGAGGAGGAGACCCAATTTTATATCAACATTTATTT |
| Coleoptera Elateridae sp. | AACTCTATACTTCATTTTTGGTGCCTGATCAGGAATACTAGGAACATCCTTAAGATTACTAATCCGTGCAGAACTAGGAAACCCAGGGGCACTAATTGGAAATGACCAAATTTACAATGTTGTTGTTACAGCCCATGCATTTATCATAATTTTCTTCATAGTTATACCTATCATAATTGGTGGATTCGGAAATTGGTTAGTCCCTTTAATGCTAGGGGCACCAGATATAGCTTTCCCTCGAATAAATAACATAAGATTCTGACTTCTCCCACCTTCCCTCAGATTGCTTTTAATGAGTAGAATTGTTGAAAGAGGGGCTGGGACTGGATGAACTGTTTATCCACCTTTAGCAGCAAACATTGCCCACAGAGGATCTTCTGTAGATTTAGCTATTTTTAGTCTCCATTTAGCAGGAATTTCATCAATTTTAGGCGCTGTAAATTTCATCTCAACTGTAATCAACATGCGATCCACAGGGATAACTTTTGACCGTATACCTCTATTTGTTTGAGCAGTTGCCATTACAGCCCTCCTTCTTCTTTTGTCTCTTCCTGTCCTTGCAGGAGCAATTACAATACTTTTAACAGACCGAAACTTAAACACCTCATTTTTTGACCCAGCAGGAGGAGGGGACCCAATTCTTTACCAACACTTATTC |
| Coleoptera Mordellidae sp. | AACACTTTATTTTATTTTTGGGGCCTGGGCAGGCATATTGGGAACGTCTCTAAGATTATTAATTCGATCAGAATTAGGAACCCCAGGAACTCTGATTGGAGATGACCAAATTTACAACGTAATTGTTACAGCTCATGCTTTCGTAATAATTTTTTTTATAGTTATGCCTGTAATAGTAGGGGGATTCGGCAATTGATTAGTTCCTCTTATATTAGGAGCACCTGACATAGCATTCCCCCGACTCAATAATATAAGATTTTGATTATTGCCCCCCTCTCTTTCTTTATTACTAATGAGAAGATTAGTCGAAAACGGTGCCGGAACAGGATGTACAGTATACCCCCCTTTATCTGCAAATCTAGCCCACGGAGGAGCTTCTGTAGATCTAGCCATCTTCAGTCTACATCTAGCAGGAATCTCATCAATTTTAGGAGCAATTAATTTTATCTCAACTATACTTAATATACGTCCAATAGGAATAACATTAGATCGACTACCTTTATTTGTTTGAGCTATTGTAATTACAGCAGTCCTATTGCTTTTATCCCTACCCGTTCTCGCAGGGGCTATTACCATATTATTAACAGATCGAAATCTAAATACCTCATTTTTTGACCCTGCAGGAGGGGGAGATCCTATTCTTTATCAACACTTATTC |

|  |  |  |
| --- | --- | --- |
| species | COI concentration (ng/µl) | COI copy number (copy/µl) |
| Lepidoptera Bombycidae *Bombyx mori* | 0.0002 | 5.17 \* 105 |
| Coleoptera Elateridae sp. | 0.3914 | 1.02 \* 109 |
| Coleoptera Mordellidae sp. | 0.05807 | 1.51 \* 108 |

Table S1. DNA-barcode sequences of the three spike-in species.

Table S2. COI concentrations of the three spike-in species

*3.2 Adding spike-in to the mock soups*. – Adding too much spike-in wastes sequencing data, while adding too little spike-in risks the loss of abundance information because in some samples, the number of spike-in reads will be too low to use as a reliable correction factor. Thus, we quantified the COI copy numbers of the mock soups by qPCR and chose a volume so that spike-in reads should make up 1% of the total number of COI copies in the lowest-concentration mock soups, balancing efficiency with reliability.

Because we mixed 52 individual OTUs to create the mock soups, we can sum the input COI copy numbers of all 52 OTUs to calculate the COI copy number of the lowest-concentration mock soup (5.25 \* 108 copies). We used 1% of this number to calculate the mass of total spike-in to add to all 21 soups (5.25 \* 106 copies). To mix spike-ins DNA conveniently, the Elateridae spike-in was diluted 1000-fold to 1.02 \* 106 copies and the Mordellidae spike-in was diluted 100-fold to 1.51 \* 106 copies before mixing. Then we added 1.45 µl of *Bombyx* spike-in(1.45 \* 5.17 \*105 = 0.75 \* 106), 1.5 µl of diluted Elateridae (1.5 \* 1.02 \* 106 = 1.5 \* 106) and 2 µl of diluted Mordellidae (2 \* 1.51 \* 106 = 3 \* 106) to each soup.

The sum of the three spike-ins was thus: , and their ratio was: .

Spike-in DNA was added directly to the mock soups’ DNA since they were already purified.

*3.3 Adding spike-in to the Malaise-trap samples*. – From the 244 Malaise-trap samples, we first extracted 17 Malaise-trap samples without adding spike-ins, and then in order to decide how much spike-in should be added, we used qPCR to quantify the COI concentrations of these 17 Malaise-trap samples, using the same protocol that we used for the mock-soup OTUs. The mean COI concentration per sample was 0.006 ng/µl, for each of which we added 200 µl volume. Mean COI mass per aliquot was thus 1.2 ng (0.006 ng/µl x 200 µl), and 1% is 0.012 ng.

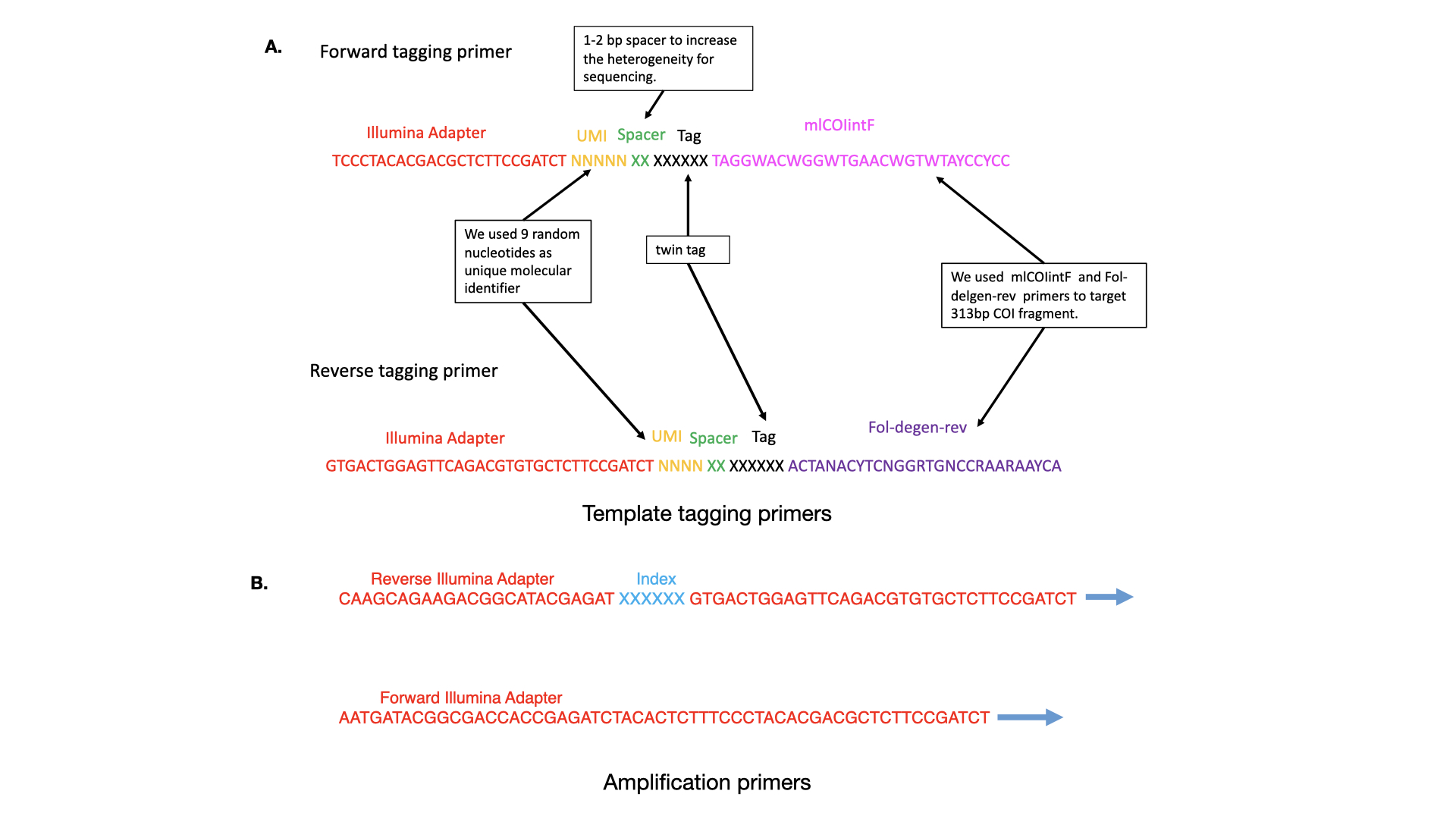
Because the Bombycid DNA spike-in had degraded after we arrived in University of Oregon where we extracted DNA from the Malaise-trap samples, we used only the other two spike-in species for the Malaise trap samples, at a ratio of 1:9 (Mordellidae:Elateridae). Mordellidae and Elateridae were re-quantified by qPCR and were diluted 100-fold. We added 2.7 µl of diluted Elateridae (2.7 µl \* 0.004 ng/µl = 0.0108 ng) and 2 µl of diluted Mordellidae (2 µl \* 0.0006 ng/µl = 0.0012 ng) to make spike-ins’ mixture for each Malaise-trap sample.

The sum of the two spike-ins: 0.0012 ng + 0.0108 ng = 0.012 ng, and their ratio was: .

We used the same spike-in mass for all the Malaise-trap samples except for the 17 samples that had already been extracted. For the selected 7 samples used in this study, lysis buffer (500 µl, 350 µl, 245 µl, 171.5 µl, 120 µl and 84 µl) from each sample was transferred into a clean 1.5 ml tube, and the spike-in DNA was added and vortexed for 10 sec. We then extracted DNA with the Qiagen QIAquick PCR purification kit, following the manufacturer instructions. DNA was eluted with 200 µl of elution buffer. In this way, the spike-in DNA was co-purified, co-amplified, and co-sequenced along with the sample DNA (Figure 5). We also recorded the total lysis buffer volume of each sample, for downstream correction.

## 4 Primer design

For this study, we simultaneously tested two methods for extracting abundance information: spike-ins and UMIs (Unique Molecular Identifiers). UMI tagging requires a two-step PCR procedure, first using tagging primers and then using amplification primers (Figure S2). The tagging primers include (1) the Leray-FolDegenRev primer pair to amplify the 313-bp COI amplicon of interest, (2) a 1- or 2-nucleotide heterogeneity spacer on both the forward and reverse primers to increase sequence entropy for the Illumina sequencer, (3) the same 6-nucleotide sequence on both the forward and reverse primers to ‘twin-tag’ the samples for downstream demultiplexing, (4) a 5N random sequence on the forward primer and a 4N random sequence on the reverse primer (9N total) to act as the UMI tags for across-species abundance estimation, (5) and parts of Illumina universal adapter sequences to anneal to the 3’ ends of the forward and reverse primer regions for the second PCR. By splitting the 9N UMI into 5N + 4N over the forward and reverse primers, we avoid primer dimers. The amplification primers include (1) the annealing primer pair to bind to the amplicons of the first PCR, (2) the full length of the Illumina adapter sequences. For further explanation of the design of the tagging primers (except for the UMI sequences), see Yang et al. (2021).



**Figure S2**. Two-step PCR protocol. A. In the first step, we amplified a 313bp COI amplicon using the Leray-FolDegenRev primers. To Both the forward and reverse primers, we added a 1-2 nucleotide heterogeneity spacer, the same 6-nucleotide sample tag (twin tags), a 9-random-nucleotide UMI (unique molecular identifier, split as 5 Ns on the forward and 4 Ns on the reverse), and a part of Illumina adapter sequence to anneal to the PCR primers in the second step. B. In the second step, we concatenate a library index sequence and the forward and reverse Illumina adapter sequences. Additional details on primer design (other than the UMI portion) can be found in Yang et al. (2021)

## 5 PCR and the Begum pipeline

We performed a two-step PCR (Lundberg et al., 2013).

The first PCR amplifies COI and concatenates sample tags and UMIs and runs for only two cycles using KAPA 2G Robust HS PCR Kit (Basel, Roche KAPA Biosystems). We used the mlCOIintF-FolDegenRev primer pair (Leray et al., 2013; Yu et al., 2012), which amplifies a 313-bp fragment of the COI barcode; and we followed the *Begum* protocol (Zepeda-Mendoza et al. 2018, Yang et al. 2021), which is a wet-lab and bioinformatic pipeline that combines multiple independent PCR replicates per sample, twin-tagging and false positive controls to remove tag-jumping and reduce erroneous sequences. Twin-tagging means using the same tag sequence on both the forward and reverse primer in a PCR. We performed 3 PCR replicates per sample, which means we used 3 different twin-tags to distinguish the 3 independent PCR replicates. *Begum* removes erroneous sequences by filtering out the sequences that appear in a low number of PCR replicates (e.g. one) at a low number of copies per PCR (e.g. 4 copies), because true sequences are more likely to appear in multiple PCRs and with high copy numbers. The 20 µl reaction mix included 4 µl Enhancer, 4 µl Buffer A, 0.4 µl dNTP (10 mM), 0.8 µl of 10 mM forward primer, 0.8 µl of 10 mM reverse primer, 0.08 µl KAPA 2G HotStart DNA polymerase (Basel, Roche KAPA Biosystems), 5 µl template DNA and 5 µl water. PCR conditions were initial denaturation at 95°C for 3 minutes, followed by two cycles of denaturation at 95°C for 1 minute, annealing at 50°C for 90 seconds, and extension at 72°C for 2 minutes. Then the products were purified with 14 µl of KAPA pure beads (Roche KAPA Biosystems, Switzerland) to remove the primers and PCR reagents and were eluted into 16 µl of water.

The second PCR amplifies the tagged templates for building the libraries which can be sequenced directly on Illumina platform. The 50 µl reaction mix included 5 µl TAKARA buffer, 4 µl dNTP (10 mM), 1.2 µl of 10mM forward primer, 1.2 µl of 10mM reverse primer, 0.25 µl TAKARA Taq DNA polymerase, 15 µl DNA product from the first PCR, and 23.35 µl water. PCR conditions were initial denaturation at 95°C for 3 minutes, 5 cycles of denaturation at 95°C for 30 seconds, annealing at 59°C for 30 seconds (-1 °C per cycle), extension at 72°C for 30 seconds, followed by 25 cycles of denaturation at 95°C for 30 seconds, annealing at 55°C for 30 seconds, extension at 72°C for 30 seconds; a final extension at 72°C for 5 minutes, and cooldown to 4°C.

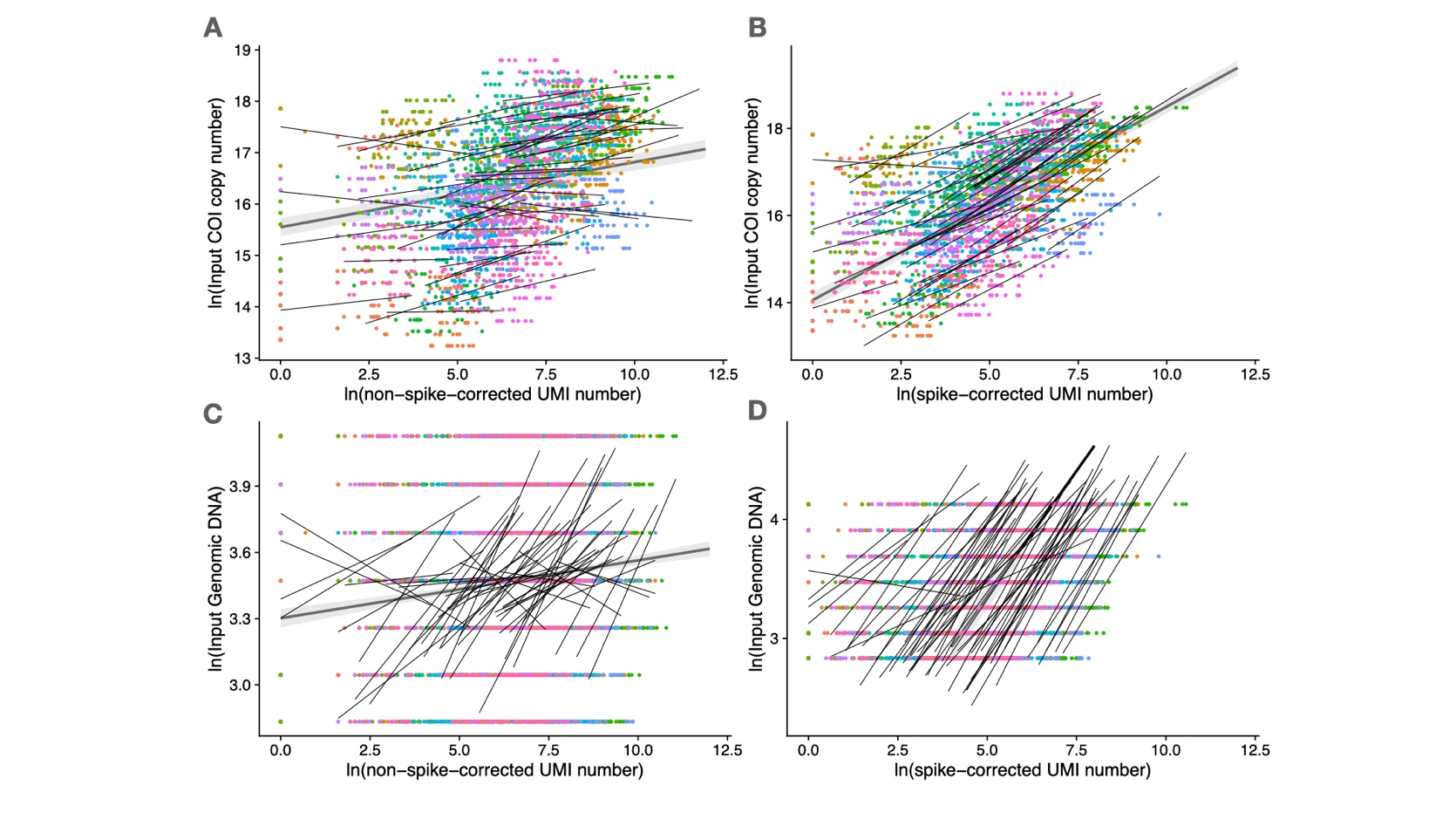
From all second PCR products, 2 µl was roughly quantified on 2% agarose gel with Image Lab 2.0 (Bio-Rad, USA). For each set of PCR reactions with the same index, amplicons were mixed at equimolar ratios to make a pooled library. One PCR negative control were set for each library. We sent our samples to Novogene (Tianjin, China) to do PE250 sequencing on Illumina NovaSeq 6000, requiring a 0.8 GB raw data from each PCR reaction.

## 6 Bioinformatic processing

*AdapterRemoval* 2.1.7 was used to remove any remaining adapters from the raw data (Schubert et al., 2016). *Sickle* 1.33 was used to trim away low-quality bases at the 3’ends. *BFC* v181 was used to denoise the reads (Li, 2015). Read merging was performed using *Pandaseq* 2.11 (Masella et al., 2012). *Begum* was used to demultiplex the reads by sample tag and to filter out erroneous reads (<https://github.com/shyamsg/Begum>, accessed 07 Sep 2021). We allowed 2-bp primer mismatches to the twin-tags while demultiplexing, and we filtered at a stringency of accepting only reads that appeared in at least two PCRs at a minimum copy number of 4 reads per PCR, with minimum length of 300 bp. This stringency minimized the false positive reads in the negative PCR control. *vsearch* 2.14.1 (Rognes et al., 2016) was used to remove chimeras (--uchime\_denovo). *Sumaclust* 1.0.2 was used to cluster the sequences into 97% similarity OTUs. The python script *tabulateSumaclust.py* from the DAMe toolkit was used generate the OTU table. Finally, we applied the R package LULU 0.1.0 with default parameters to merge oversplit OTUs (Frøslev et al., 2017). We also removed any OTUs in which we found stop codons. The OTU table and OTU representative sequences were used for downstream analysis.

*Begum* removed UMIs while sorting tags. Because of our complicated primer structure, there is no software available for our data to count the number of UMIs per OTU. We wrote our own bash scripts to process data from the Pandaseq-merged files, which includes all the UMIs, tags, and primers. We first used *Begum*-filtered sequences as a reference to filter reads in each PCR set and put the UMI information on the read headers. Then we carried out reference-based OTU clustering for each PCR with QIIME 1.9.1 (pick\_otus.py -m uclust\_ref -s 0.99), using the OTU representative sequences as the reference, counted UMIs and reads for each OTU in each PCR set, and generated the UMI and READ OTU tables.

**UMI Results**



**Figure S3**. **Recovery of mock soup within-species abundance change with UMI in COI copy number and in genomic DNA concentration**. For visualisation, all data points are shown (including all soup and PCR replicates), each thin line is fit to one of the OTUs across the seven serially diluted mock-soup samples, and the thick line represents the fitted model in which OTUs were treated as a random factor. **A**. UMI number (per OTU per soup) poorly predicts within-species variation in input COI copy number (linear mixed-effects model, marginal R2 = 0.05, conditional R2 = 0.85). **B**. Spike-corrected UMI number successfully predicts within-species variation in input COI copy number (mixed-effects linear model, marginal R2 = 0.43, conditional R2 = 0.95), but species bias remains, as can be seen in the orders-of-magnitude variation in intercepts. **C**. Non-spike-corrected UMI number poorly predicts within-species variation in input genomic DNA concentration (linear mixed-effects model, marginal R2 = 0.01, conditional R2 = 0.03). **D**. Spike-corrected UMI number successfully predicts within-species variation in input genomic DNA concentration (linear mixed-effects model, marginal R2 = 0.52, conditional R2 = 0.94) despite species bias (Figure 1). Model syntax: lme4::lmer(log.input\_gDNA or log.inputCOI\_copynumber ~ log.UMInumber + (log.UMInumber | OTUID) +   
(1 |soupRep/pcrRep)) (Bates et al., 2015).Marginal R2 represents variance explained by the fixed effect, and conditional R2 represents variance explained by the whole model. **(222 words)**

# References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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

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

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., Boehm, J. T., & Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, *10*(1), 34. https://doi.org/10.1186/1742-9994-10-34

Li, H. (2015). BFC: Correcting Illumina sequencing errors. *Bioinformatics*, *31*(17), 2885–2887. https://doi.org/10.1093/bioinformatics/btv290

Lundberg, D. S., Yourstone, S., Mieczkowski, P., Jones, C. D., & Dangl, J. L. (2013). Practical innovations for high-throughput amplicon sequencing. *Nature Methods*, *10*(10), 999–1002. https://doi.org/10.1038/nmeth.2634

Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G., & Neufeld, J. D. (2012). PANDAseq: Paired-end assembler for illumina sequences. *BMC Bioinformatics*, *13*(1), 31. https://doi.org/10.1186/1471-2105-13-31

Nielsen, M., Gilbert, M. T. P., Pape, T., & Bohmann, K. (2019). A simplified DNA extraction protocol for unsorted bulk arthropod samples that maintains exoskeletal integrity. *Environmental DNA*, *1*(2), 144–154. https://doi.org/10.1002/edn3.16

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Schubert, M., Lindgreen, S., & Orlando, L. (2016). AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Research Notes*, *9*(1), 88. https://doi.org/10.1186/s13104-016-1900-2

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

Yang, C., Bohmann, K., Wang, X., Cai, W., Wales, N., Ding, Z., Gopalakrishnan, S., & Yu, D. W. (2021). Biodiversity Soup II: A bulk‐sample metabarcoding pipeline emphasizing error reduction. *Methods in Ecology and Evolution*, *12*(7), 1252–1264. https://doi.org/10.1111/2041-210X.13602

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