MBMG
Metabarcoding & Metagenomics

Metabarcoding and Metagenomics 7: 73–88 (2023)
DOI: 10.3897/mbmg.7.101266

Research Article

# Does phylogeny explain bias in quantitative DNA metabarcoding?

Mingxin Liu[1,2,3], Christopher P. Burridge[1], Laurence J. Clarke[4,5], Susan C. Baker[1,2], Gregory J. Jordan[1,2,6]

1 *Biological Sciences, University of Tasmania, Private Bag 55, Hobart, Tasmania 7001, Australia*
2 *ARC Centre for Forest Value, University of Tasmania, Hobart, Tasmania 7001, Australia*
3 *Present affiliation: Institute of Ecology, College of Urban and Environmental Sciences, Peking University, Beijing 100871, China*
4 *Australian Antarctic Division, Kingston, Tasmania, 7050, Australia*
5 *Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, Tasmania 7001, Australia*
6 *ARC Centre of Excellence in Plant Success in Nature and Agriculture, University of Tasmania, Hobart, Tasmania 7001, Australia*

*Corresponding author: Mingxin Liu (mingxinl@pku.edu.cn)*

## Abstract

Estimating species biomass or abundance from the number of high-throughput sequencing (HTS) reads is an aspirational goal for DNA metabarcoding, yet studies have found varied correlations. Performance varies depending on the gene marker and taxonomic group and, in part, may be related to primer-template mismatches, which are likely to exhibit phylogenetic signals. In this study, we compared commonly used fragments of two gene markers for beetles, the mitochondrial cytochrome *c* oxidase subunit I (COI) and 16S ribosomal RNA (16S), which have similar lengths, but different propensity for primer-template mismatches. We tested whether primer-template mismatches influence the relationship between species biomass and HTS read abundance and whether the effect of mismatches was explained by phylogeny. A significant correlation between species biomass and HTS read abundance existed for 16S, but not for COI, which had more primer-template mismatches. Models incorporating the effects of mismatch type or number improved the estimation of species biomass from HTS read abundance for COI and strong phylogenetic signals were identified. Researchers seeking to quantify biomass from metabarcoding studies should consider the effect of primer-template mismatches for the taxonomic group of interest and, for beetles, 16S appears a good candidate. Phylogenetic correction can also improve biomass estimation when using gene markers with higher primer mismatching.

**Key words:** Coleoptera, high-throughput sequencing, PCR amplification, phylogenetic signal, species biomass



Academic editor: Florian Leese
Received: 3 April 2023
Accepted: 22 May 2023
Published: 13 June 2023

Citation: Liu M, Burridge CP, Clarke LJ, Baker SC, Jordan GJ (2023) Does phylogeny explain bias in quantitative DNA metabarcoding? Metabarcoding and Metagenomics 7: e101266. https://doi.org/10.3897/mbmg.7.101266



## Introduction

Community composition is conventionally measured by identifying organisms using phenotypic features. However, this approach can be constrained by the availability of time and taxonomic expertise, especially for speciose groups, such as arthropods (Yu et al. 2012; Ji et al. 2013). A shift towards DNA metabarcoding – identifying multiple species from a mixed sample with high-throughput sequencing (HTS) of a DNA marker – has been shown to alleviate these constraints and expedite biodiversity surveys (Yu et al. 2012; Braukmann et al. 2019; Liu et al. 2020, 2021). However, one remaining key issue with DNA metabarcoding

is whether the abundance of taxa can be reliably obtained for downstream biodiversity modelling. Despite extensive interest, quantifying species abundance with metabarcoding has had limited success (Bista et al. 2018; Lamb et al. 2018; Nichols et al. 2018; Schenk et al. 2019). While the relative amount of HTS reads per species has been used as a proxy for relative species biomass (Yu et al. 2012; Piñol et al. 2015), this assumes that extracted DNA of different species in a sample is proportionally amplified and sequenced relative to biomass. In turn, species abundance can be approximated by modelling species biomass with body traits at species-level (Brady and Noske 2006; Wardhaugh 2013). Some studies found significant positive correlations between species biomass and HTS read abundance with mock invertebrate samples (e.g. Krehenwinkel et al. (2017); Bista et al. (2018); Schenk et al. (2019)). Other studies have applied metabarcoding for estimating species niche breadth, based on relative abundance between sampling sites. For example, by controlling specimens counts and size classes in metabarcoding, Lim et al. (2022) showed similar temperature niche conservatism for arthropod OTUs on two volcanoes of Hawaii Island. However, some other studies suggested the primer pairs used and sample evenness in mixtures were also influential (see table 1 in Piñol et al. (2019)).

The causes of quantitative biases in metabarcoding are diverse and complex. An important source of bias during quantitative DNA metabarcoding is the oligonucleotide primer mismatch to binding sites during PCR amplification (Deagle et al. 2014; Nichols et al. 2018; Piñol et al. 2019). PCR amplification is less efficient in the presence of primer-template mismatches and primers that have fewer mismatches often show better amplification efficiency and a stronger correlation between species biomass and HTS reads (Elbrecht et al. 2016; Piñol et al. 2019). While binding site polymorphisms amongst species can be accommodated with degenerate primers (a mixture of primers that differ subtly in sequences), primer-template mismatches might be unavoidable in phylogenetically diverse samples, where DNA metabarcoding is particularly valuable relative to morphological species identification (Braukmann et al. 2019). Additionally, amplification efficiency (and thus the number of HTS reads) is also influenced by the type and position of primer-template mismatches (Stadhouders et al. 2010; Elbrecht et al. 2017); for example, mismatches are more detrimental at the primer 3' end than at the 5' end (Bru et al. 2008; Boyle et al. 2009). Other sources of bias during quantitative DNA metabarcoding also include polymerase bias for templates with certain base composition (Nichols et al. 2018), GC content variations of the amplified template, length differences between gene marker fragments and variation in copy number. Nichols et al. (2018) showed that DNA polymerases have preference for certain GC content which can dramatically affect the relative abundance estimation. Unlike the cytochrome *c* oxidase subunit I (COI), the stem-loop structure in 16S is likely to cause considerable amplicon length differences between different taxa. This may also bias the amplification efficiency. Last, mitochondrial copy number can vary significantly between different species and even ontogenetic stages of individual species (Kembel et al. 2012). Such copy number variation between taxa can also influence read abundance.

Quantification correction factors have been proposed to address amplification bias. For example, order-specific correction factors were developed using mock arthropod samples and the correlations between corrected HTS read abundance and input DNA increased from 0.09 to 0.82 (Krehenwinkel et

al. 2017). A similar approach of generating correction factors was also tested with a library of DNA barcoded species for seal prey and then validated with real-world prey samples (Thomas et al. 2016). While correction factors developed from known samples can account for multiple sources of bias, it represents substantial additional workload and, in samples of unknown species composition, it is difficult to apply a certain correction.

Given that the evolution of primer site mismatches amongst species likely followed similar phylogenetic trajectories to the species themselves, knowledge of phylogenetic relationships has potential value for improving biomass estimation from HTS reads. In other words, closely-related taxa are more likely to share primer-template mismatches and deviation from an overall biomass-HTS read abundance relationship. Such phylogenetic signal has been identified in other sources of quantitative metabarcoding bias, such as gene copy number (Kembel et al. 2012), suggesting phylogenetic approaches could correct other sources of bias, such as those created by primer site mismatches. Furthermore, many DNA metabarcoding studies exploit reference DNA sequences for species identification (e.g. the COI barcoding region) and, hence, these data can also be used to estimate phylogenetic relationships and to quantify primer-template mismatches when smaller internal fragments are amplified for metabarcoding.

In this study, we tested whether primer-template mismatches influenced the relationship between biomass and HTS reads and whether the relationship could be explained by phylogeny. We used DNA fragments from two mitochondrial gene markers, the COI and 16S ribosomal RNA (16S) genes. These fragments had similar lengths (157 bp and 124–165 bp, respectively), but COI had more primer-template mismatches than 16S (Deagle et al. 2014). Our aims were to: (1) examine how variation in the number of primer-template mismatches impacts estimation of species biomass in PCR-based DNA metabarcoding and (2) quantify the phylogenetic signal in primer-template mismatches and deviation from a species biomass-HTS read abundance relationship. The magnitude of phylogenetic signal indicates the potential for phylogenetic-based biomass correction. We analysed DNA metabarcoding data derived from a study assessing recovery of beetle communities following forest harvesting (Liu et al. 2020), as analyses of mock samples are typically based on unrealistically simple samples, which may reduce the applicability of their inferences.

## Methods

We re-analysed the high-throughput sequencing datasets in Liu et al. (2020). These were derived using primer sets for COI (ZBJ-ArtF1c/ZBJ-ArtR2c; Zeale et al. (2011)) and 16S (modified Ins16S_1 short set; forward: 5'-AGACGAGAAGACCCTATAGA-3'; reverse: 5'-TACGCTGTTATCCCTAAGGTA-3'; Clarke et al. (2014)) to amplify similar length DNA fragments (157 bp for COI and 124–165 bp for 16S) from field-collected beetle samples. The samples represented pitfall-trapping collections from 12 regeneration forests and 12 neighbouring mature forests, containing 173 morphologically identified beetle species. The bioinformatic pipeline was the same as that of Liu et al. (2020) unless otherwise stated. A full description of the bioinformatic pipeline is provided in the Suppl. material 1. In brief, our bioinformatic pipeline included raw sequence demultiplexing, quality control and clustering of zero-radius

OTUs (ZOTUs) and taxonomic assignment. First, raw sequences were demultiplexed by sample identifiers on the Miseq. Fastq reads were then merged and trimmed. Reads that had unused identifiers or contained any mismatch of sample-specific identifiers were discarded. Second, retained reads were quality filtered to remove low quality reads and the UNOISE algorithm was used to cluster unique reads into ZOTUs. Third, taxonomic assignment was performed for ZOTUs with BLASTn using an e-value cut-off of $1e^{-4}$ for both gene markers. We used Sanger sequencing to generate a local reference library representing 150 COI sequences and 181 16S sequences for 92 and 96 beetle species, respectively, from pilot sampling along a forest chronosequence. Reference sequences were also downloaded from the Barcode of Life Database (BOLD; Ratnasingham and Hebert 2007) and NCBI (NCBI Resource Coordinators 2018). Particularly, we curated COI ZOTUs, based on sequence similarity and co-occurrence using the LULU algorithm to reduce the inflated number of ZOTUs that had same taxonomic assignment (Froslev et al. 2017). For this study, we only included ZOTUs that had ≥ 99% sequence identity to a species in the reference database, from which the number of primer-template mismatches could be calculated. In total, 22 and 25 ZOTUs that had ≥ 99% sequence match identity to reference species were identified for COI and 16S, respectively. Our downstream analyses were limited to these species to enable a comparison of biomass and HTS read abundance (Suppl. material 2).

Primer-template mismatches were scored using existing reference sequences. The total number of primer-template mismatches is likely to oversimplify models estimating species biomass with HTS read abundance (Elbrecht et al. 2017). Therefore, we accommodated the effects of different types and positions of each primer-template mismatch (see Table 1, Suppl. material 3) by weighting with different scores as per Elbrecht et al. (2017), providing continuous variables that scale according to predicted impact on primer annealing. This scoring system was based on Stadhouders et al. (2010), which quantitatively investigated the effects of primer-template mismatches within the 3' end primer region on real-time PCR using different commercially available 5-nuclease assay master mixes. The mismatch position was scored because those near the 3' end of primer-template binding region would have a greater impact

**Table 1.** Variables used in modelling HTS read abundance, based on species biomass and primer-template mismatches.

| Term | Description |
|---|---|
| ***Dependent variable*** | |
| HTS read abundance | Average sum of HTS reads for each species |
| ***Independent variables*** | |
| Species biomass | Average weight (in grams; to 4 decimal places) of specimens for a species across samples |
| Mismatch number | Count of primer-template mismatches |
| Mismatch position | Sum of mismatch scores based on their positions relative to the 3' primer end† |
| Mismatch type | Sum of mismatch scores based on their type† |
| Deviation of predicted HTS read abundance | Residual in the linear model $\log_{10}$(HTS read abundance) ~ $\log_{10}$(species biomass) |

†: Schemes derived by Elbrecht et al. (2017) and summarised in Suppl. material 3.

on PCR amplification efficiency than at the 5' end (Stadhouders et al. 2010; Piñol et al. 2015). Likewise, the type of mismatch was recorded and scored because some varieties (e.g. A-A, G-A and C-C) more severely impact PCR amplification efficiency than others (e.g. A-C and T-G) (Stadhouders et al. 2010; Elbrecht et al. 2017).

We retrieved original species biomass and HTS read abundance from each of the 24 samples. Then, our analysis was based on these data to model the relationship between them along with the information of primer-template mismatches. For example, we would have multiple data for a given species if they were present in multiple samples. Data were normalised and expressed as relative proportions by dividing the total HTS read abundance and total sample biomass by the original HTS read abundance and species biomass for each of the 24 samples. We have provided the original species abundance and their HTS read abundance across 24 samples in Suppl. material 2. We used linear regression models ('lm' function) to test the relationship between HTS read abundance (response variable) and species biomass and primer-template mismatches as predictor variables (continuous fixed effects; Table 1). For each of COI and 16S, we developed eight alternative models of HTS read abundance (Table 2). The models were compared using Akaike's Information Criterion corrected for small sample sizes (AICc) and the predictor variable in a model of $\Delta AICc < 2$ was considered to have significant explanatory power (Arnold 2010). The residuals in the models of relative proportion of HTS read abundance ~ relative proportion of species biomass represented the deviation of HTS read abundance from that predicted with species biomass alone and were considered a proxy measure of the effect of primer-template mismatch.

In order to examine the phylogenetic effect of primer-template mismatches, we constructed a phylogenetic tree with reference DNA sequences of partial COI and 16S genes. Two species from the order Neuroptera (*Apochrysa matsumurae*, GenBank Accession: NC_015095 and *Ascaloptynx appendiculatus*, GenBank Accession: NC_011277) were selected as outgroups, based on Misof et al. (2015). A Maximum Likelihood tree was built under the GTR + Γ + I nucleotide substitution model with raxmlGUI version 2.0 (Stamatakis 2014; Edler et al. 2019). We used the "contMap" function in the phytools package (Revell 2012) to create and visualise ancestral state reconstruction under Brownian motion evolution of quantitative traits. For each gene marker, we applied this practice to the number of primer-template mismatches and the residuals of relative proportion of HTS read abundance ~ relative proportion of species biomass. To test for phylogenetic signal, phylogenetic mixed models were performed with the brm package (Bürkner 2017). Phylogenetic signal (λ; Pagel (1997, 1999)) was separately calculated for relative proportion of species biomass, relative proportion of HTS read abundance, number of mismatches and the residuals described above, using the "phylosig" function in phytools. High values of λ (approaching 1) indicate that closely-related species have very similar trait values and $\lambda = 0$ indicates that trait values are randomly distributed across the phylogeny. The substitution rate is known to saturate quickly for the third codon position in protein-coding genes and, thus, contain homoplastic information (Shao et al. 2003; Lin and Danforth 2004). Therefore, the phylogenetic signal was particularly tested for the first and second codons of the COI gene. All statistical analyses were performed using R version 4.1.1 (R Core Team 2020).

**Table 2.** Alternative linear models of relative proportion of HTS read abundance, based on the relative proportion of species biomass and/or primer-template mismatches for COI and 16S datasets. The best model as determined with lowest AICc value is highlighted in bold.

| Marker | Model | Variable | Adj. $R^2$ | *P* | ΔAICc |
|---|---|---|---|---|---|
| COI | **#1** | **Relative proportion of species biomass** | **0.30** | **0.55** | **0** |
| | | **Mismatch position** | | **0.04*** | |
| | | **Mismatch type** | | **< 0.001***** | |
| | #2 | Mismatch type | 0.24 | < 0.001*** | 1.92 |
| | #3 | Relative proportion of species biomass | 0.25 | 0.19 | 2.36 |
| | | Mismatch type | | < 0.001*** | |
| | #4 | Relative proportion of species biomass | 0.18 | 0.12 | 7.85 |
| | | Mismatch number | | < 0.01** | |
| | #5 | Mismatch number | 0.16 | < 0.01** | 8.04 |
| | #6 | Relative proportion of species biomass | 0.02 | 0.16 | 17.05 |
| | #7 | Relative proportion of species biomass | 0.01 | 0.12 | 18.52 |
| | | Mismatch position | | 0.37 | |
| | #8 | Mismatch position | -0.01 | 0.59 | 18.83 |
| 16S | **#1** | **Relative proportion of species biomass** | **0.42** | **< 0.001***** | **0** |
| | | **Mismatch position** | | **0.11** | |
| | | **Mismatch type** | | **0.02*** | |
| | #2 | Relative proportion of species biomass | 0.41 | < 0.001*** | 0.46 |
| | | Mismatch type | | 0.08 | |
| | #3 | Relative proportion of species biomass | 0.39 | < 0.001*** | 1.46 |
| | #4 | Relative proportion of species biomass | 0.39 | < 0.001*** | 3.41 |
| | | Mismatch number | | 0.65 | |
| | #5 | Relative proportion of species biomass | 0.39 | < 0.001*** | 3.61 |
| | | Mismatch position | | 0.91 | |
| | #6 | Mismatch type | 0.02 | 0.08 | 51.07 |
| | #7 | Mismatch position | -0.01 | 0.68 | 54.13 |
| | #8 | Mismatch number | -0.01 | 0.88 | 54.29 |

Significance codes: *: 0.01 ≤ *P* < 0.05; **: 0.001 ≤ *P* < 0.01; *** *P* < 0.001.

## Results

The total number of mismatches in the forward and reverse COI primer ranged from 0–7 across their 22 ZOTUs (mean ± SD = 3.1 ± 2.1), whereas there were only 0–3 mismatches for the 16S primer pair across their 25 ZOTUs (0.8 ± 1.2, Fig. 1). For COI, the best model for HTS read abundance incorporated the effects of species biomass and primer-template mismatch position, where mismatches at the 3' end of a primer affected PCR amplification efficiency more severely than at the 5' end (Model 1, Table 2). However, HTS read abundance was not significantly predicted by species biomass alone for COI (Model 6, Table 2; Fig. 1), nor by mismatch position alone (Model 8, Table 2). Other models that significantly predicted HTS read abundance had either mismatch type

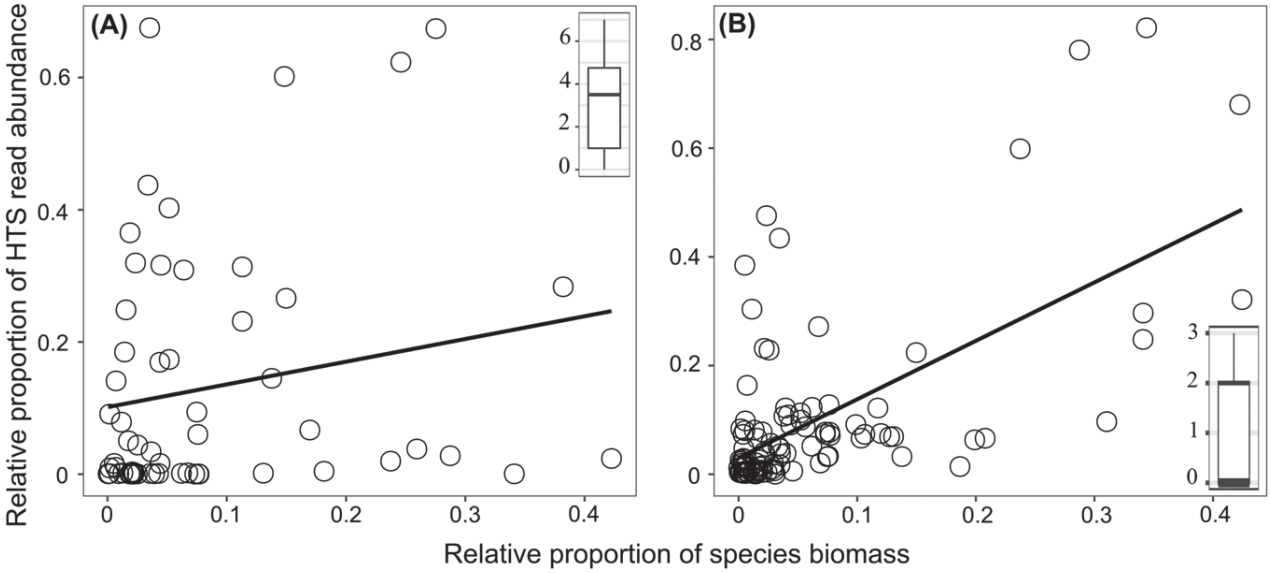


**Figure 1.** Linear regression models showing the direct relationship between relative proportion of HTS read abundance ~ relative proportion of species biomass for (**A**) COI and (**B**) 16S. The model fit was improved in each case by including terms relating to primer-template mismatches (Table 2). Insert boxplots showing the number of template-primer mismatches for each species in the COI and 16S dataset, respectively.

(Model 2) or mismatch number (Model 5) as single predictors or in combination with species biomass (Model 4).

For 16S, HTS read abundance was significantly predicted by species biomass alone (Model 3, $R^2$ = 0.39, $P$ < 0.001, Table 2), although there was slightly better model fit (ΔAICc < 2) for the best model which included mismatch position and mismatch type (Model 1, Table 2). Models including mismatch type (Model 2), mismatch number (Model 4) or mismatch position (Model 5) in combination with species biomass were not significant. In isolation, mismatch type, mismatch position or mismatch number also did not significantly predict HTS read abundance (Models 6, 7 and 8, Table 2).

The phylogenetic mixed models showed a strong signal for the COI dataset (posterior mean and 95% credible intervals for $\hat{h}^2$ = 0.64 (0.31–0.87)) and, hence, the tendency for similar correlation between relative HTS read abundance and relative species biomass amongst closely-related species. However, such phylogenetic signal was not observed for the 16S dataset ($\hat{h}^2$ = 0.37 (0–0.89)). The mismatch type and mismatch number both showed strong and significant phylogenetic signals for COI (all codons: λ = 0.79 and 0.72, P < 0.001; first and second codons together: λ = 0.85 and 0.87, P < 0.05;), but not for 16S (Table 3). While there was no statistically significant phylogenetic signal in the COI model residuals of relative proportion of HTS read abundance ~ relative proportion of species biomass, visualisation of ancestral state reconstruction showed that higher number of primer-template mismatches were mostly corresponded with under-represented HTS read abundance (i.e. negative residuals) (Fig. 2A). This is also suggested by the significant phylogenetic signal of mismatch number (Table 3). For example, the Carabidae COI clade had fewer primer-template mismatches and higher residuals of relative proportion of HTS read abundance ~ relative proportion of species biomass, while Staphylinidae, Nitidulidae and Curculinoidae had lower residuals of relative proportion of HTS read abundance ~

**Table 3.** Phylogenetic signal (Pagel's λ) of predictor variables for the COI dataset. "Residuals" are derived from the model of relative proportion of HTS read abundance ~ relative proportion of species biomass. Variables ranked by Pagel's λ from highest to lowest. A higher Pagel's λ suggests stronger phylogenetic signal (more similar traits in more closely related taxa).

| Variable | λ | *P* |
|---|---|---|
| Mismatch number (first and second codons) | **0.87** | **< 0.05*** |
| Mismatch type (first and second codons) | **0.85** | **< 0.05*** |
| Mismatch position (first and second codons) | 0.00 | 1.00 |
| Mismatch number (all codons) | **0.72** | **< 0.001***** |
| Mismatch type (all codons) | **0.79** | **< 0.001***** |
| Mismatch position (all codons) | 0.04 | 0.74 |
| Residuals | 0.17 | 0.25 |
| Relative proportion of HTS read abundance | 0.14 | 0.33 |
| Relative proportion of species biomass | 0.00 | 1.00 |

Significance codes: *: 0.01 ≤ *P* < 0.05; **: 0.001 ≤ *P* < 0.01; *** *P* < 0.001.

relative proportion of species biomass and more primer-template mismatches (Fig. 2A). However, there were a few disparities where the carabids *Pentagonica vittipennis* and *Percosoma carenoides* showed opposite patterns. In comparison, ancestral state reconstructions of 16S model residuals of relative proportion of HTS read abundance ~ relative proportion of species biomass bear little relationship to the pattern of primer-template mismatches (Fig. 2B), nor did it exhibit a phylogenetic signal (Table 3). For instance, both Carabidae and Staphylinidae had very low (mostly zero) primer-template mismatches for 16S, yet the Carabidae had mostly negative residuals, while the Staphylinidae had mostly positive residuals from the model of relative proportion of HTS read abundance ~ relative proportion of species biomass. It should be noted that the phylogenetic tree in our study is reconstructed with reference DNA sequences of COI and 16S and it is not ultimately resolved as monophyletic for Staphylinidae. Last, the magnitude of variation in residuals and mismatches is clearly much lower for 16S (Fig. 2).

## Discussion

To date, attempts to quantify species' biomass and abundance from HTS read abundance have had limited success (Krehenwinkel et al. 2017; Bista et al. 2018; Lamb et al. 2018; Schenk et al. 2019). Our real-world application of DNA metabarcoding showed that primer-template mismatches had an impact on the efficacy of COI in estimating species biomass and only little impact on that of 16S. In particular, our models supported the hypothesis that gene markers with fewer primer-template mismatches are likely to enable more reliable estimation of species biomass (Elbrecht et al. 2016; Piñol et al. 2019). However, marker selection often represents a compromise of multiple objectives (Elbrecht et al. 2016). While 16S was better at estimating species biomass in our study, taxonomic resolution using this marker may be lower due to less complete reference sequence databases compared to COI (Elbrecht et al. 2016), with downstream implications regarding study power (Liu et al. 2020). Likewise, if primer mismatches correlate with broader DNA fragment variation, biomass quantification accuracy may come at the expense of taxonomic resolution.

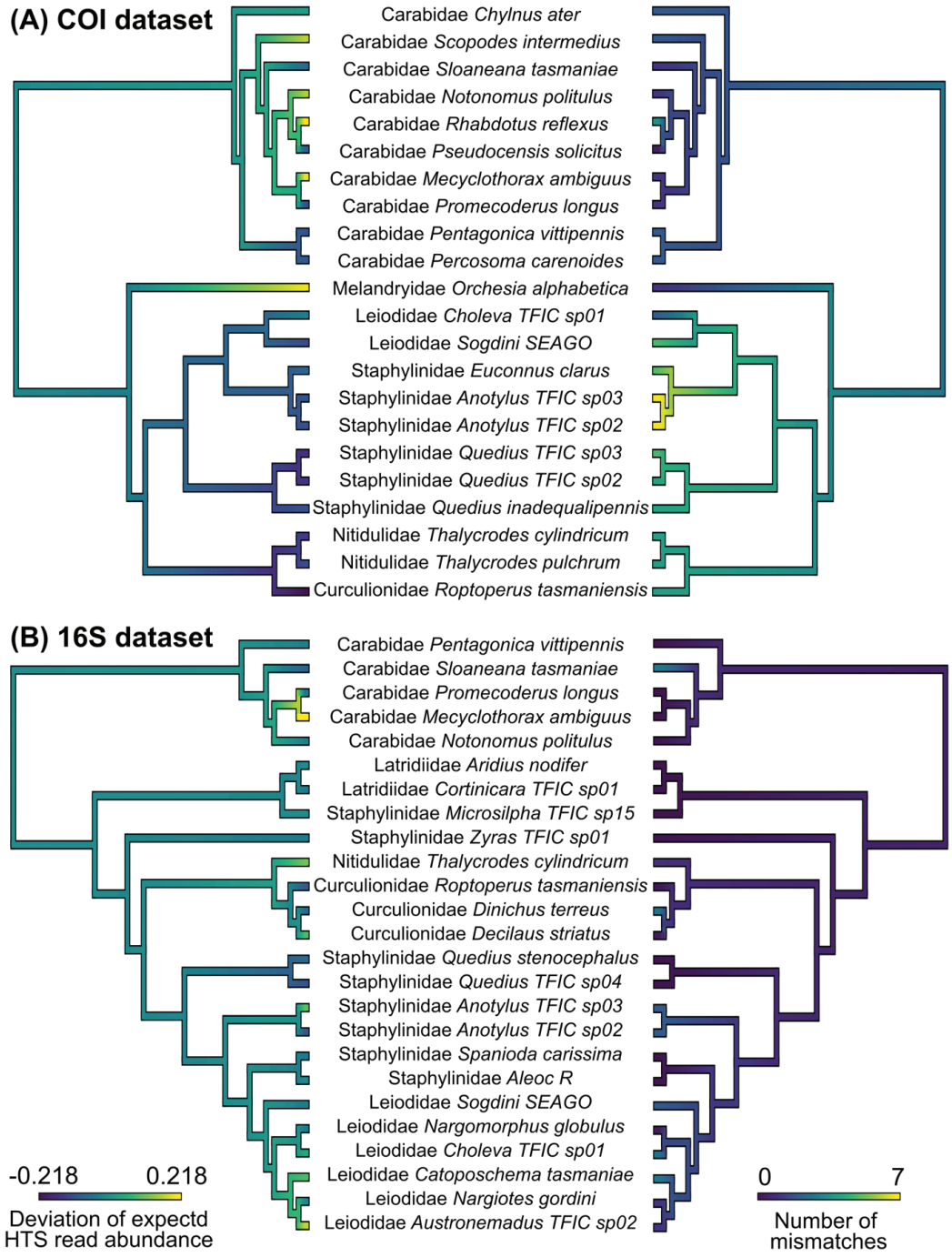


**Figure 2.** Correspondence between the model residuals of relative proportion of HTS read abundance ~ relative proportion of species biomass (left hand side) and the number of primer-template mismatches (right hand side) on phylogenetic trees for (**A**) COI and (**B**) 16S datasets. Species with lower number of primer-template mismatches are likely to have higher PCR amplification efficiency and produce relatively more HTS reads.

Our study showed the potential of phylogenetic approaches to improve biomass quantification with COI. The number and type of COI primer-template mismatches exhibited significant phylogenetic signals which may be stronger with a denser sampling of related taxa and improved reference sequence databases. The number of mismatches was also associated with the model residuals of relative proportion of HTS read abundance ~ relative proportion of species biomass. This raises the possibility of phylogenetically corrected biomass

estimation for gene markers like COI with more primer-template mismatches. The lack of a significant phylogenetic signal in either mismatch position or the model residuals of relative proportion of HTS read abundance ~ relative proportion of species biomass may be due to the relatively small number (n = 22 or 25) of species considered (a subset of data due to the quality of the reference sequence collection). We expect the phylogenetic signal to be stronger for larger datasets across distinct clades (Kembel et al. 2012; Krehenwinkel et al. 2017). However, for closely-related taxa, a smaller phylogenetic effect is expected. More complete reference databases will help inform researchers whether phylogenetic adjustment will be useful for their project.

It has been suggested that considering only the total number of primer-template mismatches oversimplifies their effects on amplification efficacy (Elbrecht et al. 2017) and, therefore, taking into account mismatch type and position might improve biomass estimation. We are aware of only one other study considering mismatch position relative to the 3' end of primers (Piñol et al. 2019) and their simulations showed similar quantification accuracy to those based on mismatch numbers. For our COI dataset, including mismatch position with species biomass had similar explanatory power for HTS read abundance as the analogous model with mismatch number. However, mismatch type had no significant impact, perhaps reflecting the small values derived from the mismatch type weighting scheme compared to mismatch number and mismatch position.

Our study did not explore other factors, such as annealing temperature and cycle number during PCR amplification and lower annealing temperatures, in particular, have improved biomass quantification during metabarcoding (Sipos et al. 2007). Factors other than primer-template mismatches also likely contribute to variation in the relationship between HTS read abundance and species biomass, such as DNA extraction efficiency, mitochondrial DNA copy number and amplicon GC content (Stadhouders et al. 2010; Kirse et al. 2023). Additionally, the different morphology amongst taxa, especially in complex communities (e.g. in exoskeletons or surface area to volume ratios), is an additional source of variation in DNA extraction efficiency. Further limitations are that we only used one primer pair for COI and 16S, respectively and our study was restricted to beetles. While more primers should be tested and on a wider range of species, we anticipate similar patterns to what we observed. Other primers for COI will consistently incur primer-template mismatches across more degenerate codon positions (i.e. third codon), while 16S primers are typically designed in more conserved ribosomal RNA coding regions (Deagle et al. 2014). In addition, our studied species were only a subset of reliably identified beetles from a single HTS run and, therefore, might not necessarily reflect their actual competition for amplification and sequencing that occurred amongst all species within the sample. Studying samples of known species that have reference sequences, but not mock samples of low diversity, could, therefore, help further clarify the relationships amongst mismatches, phylogeny and HTS read abundance.

Multiple mismatches are common in primers and COI primers with high degeneracy have become popular for arthropod metabarcoding (Elbrecht et al. 2019). The COI primers used in our study are non-degenerate and could have

an amplification bias for some arthropod lineages (Elbrecht et al. 2019). Amplification bias is less likely an issue for highly degenerate primers. In addition, Elbrecht and Leese (2017) have shown that highly degenerate primers have good efficacy in species identification while providing with a consistent and equal read abundance estimation across mock samples. However, cocktails of degenerate primers increase the chances of simultaneously amplifying a lot of non-target lineages (e.g. bacterial; Mioduchowska et al. (2018); Zafeiropoulos et al. (2021)) in an arthropod sample, depending on sample quality. Recent advances in environmental DNA metabarcoding have made it possible to monitor eDNA concentrations of multiple species. For example, by using internal standard DNAs, positive linear correlations are found between the sequence reads and the copy numbers of standard DNAs in marine fishes (Ushio et al. 2018; Sato et al. 2021).

Overall, our study demonstrates a phylogenetic basis of quantification biases from DNA metabarcoding and suggests an opportunity for correcting such biases based on phylogeny. Taxon-specific correction factors can be derived from mock samples by fitting a regression line for the correlation of input species biomass and recovered HTS read abundance while accounting for phylogeny (Krehenwinkel et al. 2017). An alternative approach is to focus on markers with few primer-template mismatches. Thus, the significant correlation between HTS read abundance and species biomass found in the 16S dataset was also supported in a similar study on this gene (Elbrecht et al. 2016). However, future studies on quantitative metabarcoding should navigate the trade-offs of using alternative DNA markers. The quantitative performance should be considered in line with the available reference sequence data or need to generate new reference sequences and the ease of bioinformatic filtering (Andujar et al. 2018). With improved taxonomic coverage of 16S in reference datasets, this marker may be a good candidate for estimating species abundance in PCR-based DNA metabarcoding studies.

## Acknowledgements

This project was funded by the Australian Research Council Linkage Grant (LP140100075) with partner organisations Sustainable Timber Tasmania and VicForests, a Holsworth Wildlife Research Endowment for fieldwork and a student research grant for laboratory work from the Forest Practices Authority. ML was also supported by the China Scholarship Council (CSC No. 201608360109). We would like to thank Lynne Forster (UTAS) for confirming beetle species identifications, Adam Smolenski and Sharee McCammon (UTAS) for helping with HTS preparation.

## Additional information

### Conflict of interest

No conflict of interest was declared.

### Ethical statement

No ethical statement was reported.

## Funding

This project was funded by the Australian Research Council Linkage Grant (LP140100075) with partner organisations Sustainable Timber Tasmania and VicForests, a Holsworth Wildlife Research Endowment for fieldwork and a student research grant for laboratory work from the Forest Practices Authority. ML was also supported by the China Scholarship Council (CSC No. 201608360109).

## Author contributions

Mingxin Liu: Conceptualisation, Formal analysis, Investigation, Writing - original draft & editing. Christopher P. Burridge: Funding acquisition, Formal analysis, Writing - review & editing. Laurence J. Clarke: Formal analysis, Writing - review & editing. Susan C. Baker: Funding acquisition, Formal analysis, Writing - review & editing. Gregory J. Jordan: Conceptualisation, Funding acquisition, Formal analysis, Writing - review & editing.

## Author ORCIDs

Mingxin Liu https://orcid.org/0000-0003-0436-4058
Christopher P. Burridge https://orcid.org/0000-0002-8185-6091
Laurence J. Clarke https://orcid.org/0000-0002-0844-4453
Susan C. Baker https://orcid.org/0000-0002-7593-0267
Gregory J. Jordan https://orcid.org/0000-0002-6033-2766

## Data availability

The demultiplexed paired-end HTS reads (fastq files) used for this study are available in the NCBI Short Sequence Archive (BioProjects: PRJNA612410 and PRJNA656915). All analysis code can be accessed via https://github.com/mingxinliu/Quantitative-Metabarcoding.

# References

Andujar C, Arribas P, Yu DW, Vogler AP, Emerson BC (2018) Why the COI barcode should be the community DNA metabarcode for the metazoa. Molecular Ecology 27(20): 3968–3975. https://doi.org/10.1111/mec.14844

Arnold TW (2010) Uninformative parameters and model selection using Akaike's information criterion. The Journal of Wildlife Management 74(6): 1175–1178. https://doi.org/10.1111/j.1937-2817.2010.tb01236.x

Bista I, Carvalho GR, Tang M, Walsh K, Zhou X, Hajibabaei M, Shokralla S, Seymour M, Bradley D, Liu S, Christmas M, Creer S (2018) Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. Molecular Ecology Resources 18(5): 1020–1034. https://doi.org/10.1111/1755-0998.12888

Boyle B, Dallaire N, MacKay J (2009) Evaluation of the impact of single nucleotide polymorphisms and primer mismatches on quantitative PCR. BMC Biotechnology 9(1): 75. https://doi.org/10.1186/1472-6750-9-75

Brady CJ, Noske RA (2006) Generalised regressions provide good estimates of insect and spider biomass in the monsoonal tropics of Australia. Australian Journal of Entomology 45(3): 187–191. https://doi.org/10.1111/j.1440-6055.2006.00533.x

Braukmann TWA, Ivanova NV, Prosser SWJ, Elbrecht V, Steinke D, Ratnasingham S, deWaard JR, Sones JE, Zakharov EV, Hebert PDN (2019) Metabarcoding a diverse arthropod mock community. Molecular Ecology Resources 19(3): 711–727. https://doi.org/10.1111/1755-0998.13008

Bru D, Martin-Laurent F, Philippot L (2008) Quantification of the detrimental effect of a single primer-template mismatch by real-time PCR using the 16S rRNA gene as an example. Applied and Environmental Microbiology 74(5): 1660–1663. https://doi.org/10.1128/AEM.02403-07

Bürkner P-C (2017) brms: An R Package for bayesian multilevel models using Stan. Journal of Statistical Software 80(1): 1–28. https://doi.org/10.18637/jss.v080.i01

Clarke LJ, Soubrier J, Weyrich LS, Cooper A (2014) Environmental metabarcodes for insects: In silico PCR reveals potential for taxonomic bias. Molecular Ecology Resources 14(6): 1160–1170. https://doi.org/10.1111/1755-0998.12265

Deagle BE, Jarman SN, Coissac E, Pompanon F, Taberlet P (2014) DNA metabarcoding and the cytochrome c oxidase subunit I marker: Not a perfect match. Biology Letters 10(9): 20140562. https://doi.org/10.1098/rsbl.2014.0562

Edler D, Klein J, Antonelli A, Silvestro D (2019) raxmlGUI 2.0 beta: a graphical interface and toolkit for phylogenetic analyses using RAxML. BioRxiv: 800912. https://doi.org/10.1101/800912

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

Elbrecht V, Taberlet P, Dejean T, Valentini A, Usseglio-Polatera P, Beisel J-N, Coissac E, Boyer F, Leese F (2016) Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects. PeerJ 4: e1966. https://doi.org/10.7717/peerj.1966

Elbrecht V, Leese F, Bunce M (2017) PrimerMiner: An R package for development and *in silico* validation of DNA metabarcoding primers. Methods in Ecology and Evolution 8(5): 622–626. https://doi.org/10.1111/2041-210X.12687

Elbrecht V, Braukmann TWA, Ivanova NV, Prosser SWJ, Hajibabaei M, Wright M, Zakharov EV, Hebert PDN, Steinke D (2019) Validation of COI metabarcoding primers for terrestrial arthropods. PeerJ 7: e7745. https://doi.org/10.7717/peerj.7745

Froslev TG, Kjoller R, Bruun HH, Ejrnaes R, Brunbjerg AK, Pietroni C, Hansen AJ (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, Ashton L, Pedley SM, Edwards DP, Tang Y, Nakamura A, Kitching R, Dolman PM, Woodcock P, Edwards FA, Larsen TH, Hsu WW, Benedick S, Hamer KC, Wilcove DS, Bruce C, Wang X, Levi T, Lott M, Emerson BC, Yu DW (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecology Letters 16(10): 1245–1257. https://doi.org/10.1111/ele.12162

Kembel SW, Wu M, Eisen JA, Green JL (2012) Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. PLoS Computational Biology 8(10): 16–18. https://doi.org/10.1371/journal.pcbi.1002743

Kirse A, Bourlat SJ, Langen K, Zapke B, Zizka VMA (2023) Comparison of destructive and nondestructive DNA extraction methods for the metabarcoding of arthropod bulk samples. Molecular Ecology Resources 23(1): 92–105. https://doi.org/10.1111/1755-0998.13694

Krehenwinkel H, Wolf M, Lim JY, Rominger AJ, Simison WB, Gillespie RG (2017) Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. Scientific Reports 7(1): 17668. https://doi.org/10.1038/s41598-017-17333-x

Lamb PD, Hunter E, Pinnegar JK, Creer S, Davies RG, Taylor MI (2018) How quantitative is metabarcoding: A meta-analytical approach. Molecular Ecology 28(2): 420–430. https://doi.org/10.1111/mec.14920

Lim JY, Patiño J, Noriyuki S, Cayetano L, Gillespie RG, Krehenwinkel H (2022) Semi-quantitative metabarcoding reveals how climate shapes arthropod community assembly along elevation gradients on Hawaii Island. Molecular Ecology 31(5): 1416–1429. https://doi.org/10.1111/mec.16323

Lin CP, Danforth BN (2004) How do insect nuclear and mitochondrial gene substitution patterns differ? Insights from Bayesian analyses of combined datasets. Molecular Phylogenetics and Evolution 30(3): 686–702. https://doi.org/10.1016/S1055-7903(03)00241-0

Liu M, Baker SC, Burridge CP, Jordan GJ, Clarke LJ (2020) DNA metabarcoding captures subtle differences in forest beetle communities following disturbance. Restoration Ecology 28(6): 1475–1484. https://doi.org/10.1111/rec.13236

Liu M, Jordan GJ, Burridge CP, Clarke LJ, Baker SC (2021) Metabarcoding reveals landscape drivers of beetle community composition approximately 50 years after timber harvesting. Forest Ecology and Management 488: 119020. https://doi.org/10.1016/j.foreco.2021.119020

Mioduchowska M, Czyż MJ, Gołdyn B, Kur J, Sell J (2018) Instances of erroneous DNA barcoding of metazoan invertebrates: Are universal cox1 gene primers too "universal"? PLoS ONE 13(6): e0199609. https://doi.org/10.1371/journal.pone.0199609

Misof B, Liu S, Meusemann K, Peters RS (2015) Phylogenomics resolves the timing and pattern of insect evolution. Science 349: 487. https://doi.org/10.1126/science.aaa5460

NCBI Resource Coordinators (2018) Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 46(D1): D8–D13. https://doi.org/10.1093/nar/gkx1095

Nichols RV, Vollmers C, Newsom LA, Wang Y, Heintzman PD, Leighton M, Green RE, Shapiro B (2018) Minimizing polymerase biases in metabarcoding. Molecular Ecology Resources 18(5): 927–939. https://doi.org/10.1111/1755-0998.12895

Pagel M (1997) Inferring evolutionary processes from phylogenies. Zoologica Scripta 26(4): 331–348. https://doi.org/10.1111/j.1463-6409.1997.tb00423.x

Pagel M (1999) Inferring the historical patterns of biological evolution. Nature 401(6756): 877–884. https://doi.org/10.1038/44766

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

Piñol J, Senar MA, Symondson WOC (2019) The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. Molecular Ecology 28(2): 407–419. https://doi.org/10.1111/mec.14776

Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). Molecular Ecology Notes 7(3): 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

R Core Team (2020) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Revell LJ (2012) phytools: An R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3(2): 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x

Sato M, Inoue N, Nambu R, Furuichi N, Imaizumi T, Ushio M (2021) Quantitative assessment of multiple fish species around artificial reefs combining environmental DNA metabarcoding and acoustic survey. Scientific Reports 11(1): 19477. https://doi.org/10.1038/s41598-021-98926-5

Schenk J, Geisen S, Kleinboelting N, Traunspurger W (2019) Metabarcoding data allow for reliable biomass estimates in the most abundant animals on earth. Metabarcoding and Metagenomics 3: 117–126. https://doi.org/10.3897/mbmg.3.46704

Shao R, Dowton M, Murrell A, Barker SC (2003) Rates of gene rearrangement and nucleotide substitution are correlated in the mitochondrial genomes of insects. Molecular Biology and Evolution 20(10): 1612–1619. https://doi.org/10.1093/molbev/msg176

Sipos R, Székely AJ, Palatinszky M, Révész S, Márialigeti K, Nikolausz M (2007) Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. FEMS Microbiology Ecology 60: 341–350. https://doi.org/10.1111/j.1574-6941.2007.00283.x

Stadhouders R, Pas SD, Anber J, Voermans J, Mes THM, Schutten M (2010) The effect of primer-template mismatches on the detection and quantification of nucleic acids using the 5' nuclease assay. The Journal of Molecular Diagnostics 12(1): 109–117. https://doi.org/10.2353/jmoldx.2010.090035

Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics (Oxford, England) 30(9): 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Thomas AC, Deagle BE, Eveson JP, Harsch CH, Trites AW (2016) Quantitative DNA metabarcoding: Improved estimates of species proportional biomass using correction factors derived from control material. Molecular Ecology Resources 16(3): 714–726. https://doi.org/10.1111/1755-0998.12490

Ushio M, Murakami H, Masuda R, Sado T, Miya M, Sakurai S, Yamanaka H, Minamoto T, Kondoh M (2018) Quantitative monitoring of multispecies fish environmental DNA using high-throughput sequencing. Metabarcoding and Metagenomics 2: e23297. https://doi.org/10.1101/113472

Wardhaugh CW (2013) Estimation of biomass from body length and width for tropical rainforest canopy invertebrates. Australian Journal of Entomology 52(4): 291–298. https://doi.org/10.1111/aen.12032

Yu DW, Ji Y, Emerson BC, 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

Zafeiropoulos H, Gargan L, Hintikka S, Pavloudi C, Carlsson J (2021) The Dark mAtteR iNvestigator (DARN) tool: Getting to know the known unknowns in COI amplicon data. Metabarcoding and Metagenomics 5: e69657. https://doi.org/10.3897/mbmg.5.69657

Zeale MR, Butlin RK, Barker GL, Lees DC, Jones G (2011) Taxon-specific PCR for DNA barcoding arthropod prey in bat faeces. Molecular Ecology Resources 11(2): 236–244. https://doi.org/10.1111/j.1755-0998.2010.02920.x

## Supplementary material 1

### The bioinformatic pipeline of processing metabarcoding data

Authors: Mingxin Liu, Christopher P. Burridge, Laurence J. Clarke, Susan C. Baker, Gregory J. Jordan

Data type: docx. file



Link: https://doi.org/10.3897/mbmg.7.101266.suppl1

## Supplementary material 2

### Composition and biomass of 24 studied samples, and HTS read abundance of studied species

Authors: Mingxin Liu, Christopher P. Burridge, Laurence J. Clarke, Susan C. Baker, Gregory J. Jordan

Data type: tables (xlsx. file)

Explanation note: table S1: The measurements of individual number, species mean biomass and total biomass across 24 samples; table S2: HTS read abundance of 22 and 25 species studied in COI and 16S dataset across 24 samples



Link: https://doi.org/10.3897/mbmg.7.101266.suppl2

## Supplementary material 3

### Scoring scheme for mismatch between DNA template and primers

Authors: Mingxin Liu, Christopher P. Burridge, Laurence J. Clarke, Susan C. Baker, Gregory J. Jordan

Data type: tables (docx. file)

Explanation note: table S3: Scoring scheme of mismatch types from Elbrecht et al. (2017); table S4: Scoring scheme of mismatch position from Elbrecht et al. (2017).



Link: https://doi.org/10.3897/mbmg.7.101266.suppl3