**Can amplicon sequence conservation and PCR cycle number reduce amplification bias and enable quantitative metabarcoding**

Henrik Krehenwinkel1 2, Madeline Wolf1, Jun Ying Lim1, Andrew J. Rominger1, Warren B. Simison2, Rosemary G. Gillespie1

Department of Environmental Sciences, Policy and Management

University of California Berkeley

Mulford Hall, Berkeley, California, USA

Center for Comparative Genomics

California Academy of Sciences

Music Concourse Drive, San Francisco, California, USA

[Krehenwinkel@berkeley.edu](mailto:Krehenwinkel@berkeley.edu)

+1-510-646-3409

Running head: Amplification bias in metabarcoding

**Abstract**

Amplicon based metabarcoding promises rapid and cost-efficient analyses of species composition but due to taxon specific PCR amplification biases, there is still caution over the use of relative abundance estimates derived from metabarcoding. PCR-free approaches have been suggested to mitigate this problem, but come with considerable increases in workload and cost. Here, we analyze multilocus datasets of diverse arthropod communities, to evaluate whether amplification bias can be countered by **1)** targeting loci with highly degenerate primers or conserved priming sites, **2)** increasing DNA template concentration during PCR, or **3)** by reducing the PCR cycle number or completely avoiding locus specific amplification by directly sequencing genomic DNA. Our results show that amplicon sequencing reliably recovers species compositions, but quantitative analyses suffer from read abundance biases between taxa. However, this amplification bias can be greatly reduced by choosing degenerate primers or targeting amplicons with conserved priming sites. Surprisingly, neither a reduction of PCR cycles, nor a complete exclusion of locus specific amplification have a strong effect on amplification bias between taxa. This suggests copy number variation of the target loci is a primary explanation for read abundance differences between taxa, which would affect amplicon based and PCR free methods alike. Nevertheless, as read abundance biases are taxon specific and highly predictable, we also show that the application of correction factors can allow reliable abundance estimates using metabarcoding approaches.

**Keyword:** amplification bias,amplicon sequencing,metabarcoding, abundance estimates,copy number variation, metagenomics, arthropods

**Introduction**

Next generation sequencing technology has ushered in a revolution in evolutionary biology and ecology. This revolution has spurred various studies in the field of molecular barcoding. Next generation sequencing-based barcoding comes with a small workload, is cost efficient1, and provides ecologists with a means to identify large numbers of taxa in a given community. The resulting leap in throughput has allowed large-scale metabarcoding of entire ecosystems2,3,4,5 and promises unprecedented insights into ecosystem function and assembly through the recovery of species richness, food web structure, cryptic species, and hidden diversity, such as internal parasitoids6,7,8,9. Nevertheless, a critical, but not yet sufficiently understood application, of metabarcoding approaches is the potential estimation of species abundances10.

The difficulty in inferring abundances of taxa stems largely from the numerous biases incurred through commonly used PCR approaches. The main reason for this difficulty is that sequence divergence in priming sites affects priming (and subsequently amplification) efficiency directly11. Furthermore, there are other factors inherent to the targeted sequence that can bias amplification as well. For example, short sequences are amplified preferentially in amplicon mixes of variable length (e.g. ribosomal DNA), and templates of very low or very high GC content amplify less well. Mitochondrial genes are known to integrate into the nuclear genome as nonfunctional pseudogenes, which often coamplify during PCR12. Another confounding factor in the recovery of abundance estimates from PCR approaches is copy number variation (CNV) of the target locus13. All these factors can lead to flawed abundance estimates from amplicon sequencing data, even with highly conserved priming sites14,15.

The many avenues through which biases can be introduced imply that only presence and absence of taxa can be scored reliably from community amplicon sequencing. But as most measures of alpha and beta diversity are dependent on the reliable recovery of taxon abundances16, the utility of metabarcoding for diversity assessments has been questioned. Consequently, several suggestions have been made to improve metagenomic assessments of diversity and make abundance estimates possible17,18. A short stretch of the mitochondrial Cytochrome Oxidase Subunit I (COI) gene is commonly used as a barcoding marker in animals19. While the high variability of COI makes it an ideal choice to identify species or even intraspecific variation, this variation will also amplify priming bias. Hence, other markers with more conserved priming sites have been suggested as potential substitutes for COI20,21,22,23. Such novel markers, however, usually provide less taxonomic resolution24 and do not have well-developed sequence reference databases25. Another solution is the use of degenerate COI primers, which mitigate PCR bias and allow amplification across a broader taxonomic range26, or the design of taxon specific primers. However, factors like GC content and amplicon length variation will affect amplification irrespective of primer sequence conservation. Thus, alternative approaches suggested to mitigate PCR bias include the increase of DNA template concentrations or reduction of cycle numbers during PCR27. As PCR exponentially amplifies DNA templates, amplification bias should significantly increase with the number of PCR cycles. Reducing the number of PCR cycles should mitigate bias and allow for a more accurate correlation of input DNA to recovered reads15. PCR-free approaches have also been suggested to exclude amplification bias. The direct sequencing of genomic DNA or sequence capture of barcodes does not require a PCR amplification stage and is hence assumed to provide more accurate predictions of abundance28,29,30. However, such PCR-free methods come with a considerable increase in workload and processing cost (e.g. for enrichment and library preparation), and while they mitigate amplification bias, they are also sensitive to CNV in the target loci.

Despite the evidence for strong PCR biases outlined above, we can capitalize on known elements of PCR predictability and accuracy, such as those shown through applications of quantitative PCR31. For example, the proportion of input DNA of a taxon in a community should be tightly correlated to the proportion of recovered reads of that taxon, and amplification bias or CNVs should only affect the slope of this correlation. Recent research has shown that read abundance correction could help in the prediction of species abundances from sequencing data13,17,32,33. Since PCR bias is induced by sequence composition, it should be similar in closely related taxonomic groups, as has been shown in bacteria13,34. Hence, similar correction factors could possibly be derived for closely related taxa, allowing for community level abundance estimates without the need to calibrate a correction model for every taxon in the community.

Considering the afore-mentioned issues, the current study examines the hypotheses that PCR bias in amplicon based metabarcoding can be countered by: **1)** Choosing appropriate barcode markers with high sequence conservation and/or high levels of primer degeneracy, **2)** reducing the PCR cycle number and increasing the template concentration during library preparation, or completely avoiding locus specific amplification and **3)** identifying and correcting for taxon-specific read abundance bias.

To test these hypotheses, we performed three experiments using DNA and tissue mock communities of taxonomically diverse sets of Hawaiian and Californian arthropods. **1)** Using eight primer pairs, we test for the effect of different factors on amplification bias as well as qualitative and quantitative community characterization. The targeted amplicons showed varying degrees of sequence conservation. In addition, we used primers of varying degrees of degeneracy. **2)** In a second experiment, we explored the effect of varying PCR cycle numbers and increasing DNA template concentration during library preparation. **3)** Finally, we compared the quantitative recovery of taxa by amplicon sequencing with that from metagenomic sequencing of genomic DNA, e.g. completely avoiding amplification with locus specific primers.

**Methods**

*Sample collection, mock community preparation, PCR and library preparation*

Arthropod samples were collected using beat sheets in native rainforests on the Hawaiian Islands of Maui and Hawaii and an oak woodland near the University of California Berkeley in the Spring of 2015 and 2016. Specimens were stored in 99-% ethanol, morphologically identified to order and then assigned to species (or morphotype when identity was uncertain). We extracted DNA from 43 taxa, representing 19 orders (in the Arachnida, Crustacea, Hexapoda & Myriapoda). DNA extractions were performed on whole bodies using the Qiagen Puregen Kit according to the manufacturer’s protocol (Qiagen, Hilden, Germany). The concentration of each extraction was determined using a Qubit Fluorometer (Thermo Scientific, Waltham, USA) and each sample diluted to a final concentration of 15 ng/µl. We prepared 23 mock communities by pooling randomized volumes of each of the 43 samples. Each pool contained all samples in randomized volumes from 0.7 to 5 µl per sample in increments of 0.1 µl.

We chose 8 primer combinations amplifying three mitochondrial and four nuclear markers (see Table 1). We had previously generated reference sequences for the specimens in the mock communities for these markers. The primers showed varying degrees of degeneracy and amplified sequences of varying degrees of conservation, from the highly conserved nuclear ribosomal DNA to more variable mitochondrial markers (See Table 2). All primer pairs amplified sequences shorter than 500 bp to achieve an overlap between paired 300 bp Illumina MiSeq reads.

**Table 1** Targeted genes, primer combinations and primer sequences used in this study.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Gene** | **Marker type** | **Forward** | **Sequence 5'-3'** | **Reverse** | **Sequence 5'-3'** |
| COI | Mitochondrial | ArF14 | GCNCCWGAYATRGCNTTYCCNCG | Fol-degen-rev26 | TANACYTCNGGRTGNCCRAARAAYCA |
| COI | Mitochondrial | mlCOIintF46 | GGWACWGGWTGAACWGTWTAYCCYCC | Fol-degen-rev26 | TANACYTCNGGRTGNCCRAARAAYCA |
| CytB | Mitochondrial | CB347 | GAGGAGCAACTGTAATTACTAA | CB447 | AAAAGAAARTATCATTCAGGTTGAAT |
| 12SrDNA | Mitochondrial | 12sai48 | AAACTAGGATTAGATACCCTATTAT | 12sbi48 | AAGAGCGACGGGCGATGTGT |
| 18SrDNA | Nuclear | SSU\_FO449 | GCTTGTCTCAAAGATTAAGCC | SSU\_R2249 | GCCTGCTGCCTTCCTTGGA |
| 18SrDNA | Nuclear | 18s\_2F50 | AACTTAAAGRAATTGACGGA | 18s\_4R50 | CKRAGGGCATYACWGACCTGTTAT |
| 28SrDNA | Nuclear | 28s\_3F50 | TTTTGGTAAGCAGAACTGGYG | 28s\_4R50 | ABTYGCTACTRCCACYRAGATC |
| Histone H3 | Nuclear | H3aF51 | ATGGCTCGTACCAAGCAGACVGC | H3aR51 | ATATCCTTRGGCATRATRGTGAC |

PCRs were run in 10 µl volumes using the Qiagen Multiplex PCR kit, with 1 µl of DNA and 0.5 µl of each 10 µM primer. An optimal annealing temperature of 55ºC for the nuclear and 46 ºC for the mitochondrial markers was identified by running gradient PCRs. PCR amplification was performed in two rounds. The first round consisted of 32 cycles using tailed primers, whereas a second indexing PCR was performed on these tails with 6 cycles, to introduce Illumina TruSeq adapters and dual indices. The basic PCR layout followed that described in Lange et al35. We amplified the mock communities for each of the 8 markers. After each round of PCR, the remaining primer sequences were cleaned from the product with 1X AMpure XP Beads (Beckman Coulter, Indianapolis, USA). The final libraries were quantified with a Qubit Fluorometer, then all samples pooled in equimolar amounts.

*PCR cycle reduction*

Additionally, we ran a series of PCRs with varying cycle numbers. All DNA mock communities were used for this experiment. 4 µl of template DNA (60 ng) were used in a 10 µl PCRs to allow an initial priming of as many template molecules as possible with few PCR rounds. Experiments with 4, 8, 16 and 32 first-round PCR cycles were run, followed by second-round indexing PCRs of 26, 22, 14 and 6 cycles. Assuming that primarily locus specific PCR priming bias leads to inaccurate species abundances in community samples, a low number of first-round PCR cycles should reduce this bias. As the indexing PCR is based on the same priming sites (5’-tails introduced in the first round PCR) on all samples, priming bias should be of minor concern (See Suppl. Figure 1 for concept visualization).

*Metagenomic gDNA sequencing*

In addition, we sequenced one of our mock community pools as a metagenomic library. The library was prepared from untreated gDNA using the Illumina TruSeq kit and only six cycles of indexing PCR. We completely avoided amplification with locus specific primers for the metagenomic library preparation. The six-cycle indexing PCR however, was the same for metagenomic and amplicon libraries. This allowed us to estimate the effect of primer sequences on recovery of different taxa in the communities. Also, the metagenomic data allowed to estimate the effect of PCR cycle number. With strong PCR amplification bias, the metagenomic pool would be expected to yield significantly more even sequence recovery across taxa, than PCR libraries.

*Tissue mock communities*

To test the applicability of our approach under real conditions, we used mock communities from tissue pools of different Hawaiian taxa. Specimens were identified to species (or morphotype) as described above and defined amounts of tissue of approximately 20 taxa were combined into 30 mock communities. Due to the limited number of samples, we were not able to make exact replicates for the same species for some taxa, but had to make pools with more distant relatives. Specimens were dried for 1 hour on Kimwipes at room temperature. Depending on specimen size, specimens were either added whole or cut into sections using a scalpel blade. Each tissue piece was weighed on a micro balance (Mettler-Toledo, Oakland, CA, USA). The respective body parts for each specimen and pool were noted. The final communities contained 5.25 – 24.12 mg (mean = 15.36 mg) of tissue. They were combined in 2 ml Eppendorf tubes, with a 5 mm stainless steel bead and disrupted by shaking for 2 min at 1,200 hz on a Genogrinder 2010 (OPS Diagnostics, Metuchen, NJ, USA). DNA was extracted from the lysate and the DNA quantified as described above. Mitochondrial COI was amplified from each sample using the primer pairs *mlCOIintF/Fol-degen-rev*.

*Sequencing and sequence analysis*

The final pools were sequenced on an Illumina MiSeq, using V3 chemistry and 2 x 300 bp reads according to the manufacturer’s protocol (Illumina, San Diego, USA). Reads were assembled using PEAR36 with a minimum overlap of 50 and a minimum quality of 30. The assemblies were quality filtered using the FastX Toolkit37 with a minimum of 90-% of bases ≥ Q30. Separate primer pair samples were demultiplexed by marker, using the forward and reverse primer sequences as indices and primer sequences trimmed with a custom UNIX script. Each of the previously generated alignments of reference specimens per marker was used to calculate average uncorrected pairwise genetic distances between all taxa in the reference library (as a measure of conservation of the amplicon) and to create BLAST databases. Using BLASTn against these databases, we quantified the abundance of reads for each of our target taxa and genes in the DNA mock communities. Only the best BLAST hit was retained per sequence. We did not generate separate reference sequences for the tissue mock communities. Instead, an OTU clustering of all concatenated COI sequences from the tissue pools was performed using USEARCH38 with a minimum similarity of 97%. The taxonomy of the resulting OTU centroid sequences was assigned using BLAST. Taxon recovery and read abundance to input tissue proportion were analyzed as described above for the DNA pools.

The metagenomic library was quality trimmed using PEAR. We blasted the reads against the previously generated reference libraries for all 8 PCR amplicons, to estimate abundances of sequences for the according genes and taxa.

*Qualitative and quantitative community analyses*

Using linear regression of the proportion of reads per specimen against its actual proportion in each mock community, we obtained the coefficient of determination (*R2*) and the slope of the associated regression line for each specimen and marker. We used *R2* as a measure of predictability of the amount of input abundance per taxon vs. the proportion of reads recovered. The slope, on the other hand, served as a measure of fold change between the input proportion of DNA in the mock community and the resulting number of reads. Specifically, we compared slopes to the 1:1 line (representing ideal prediction of recovered reads from input DNA) by taking the difference between the absolute value of the observed slope and 1.

As we did not have replicates of the gDNA library, we could not perform linear regression for this sample. Instead, a fold change was calculated between the proportion of input DNA for each taxon and the recovered sequences for all eight markers. This fold change was compared to a fold change for amplicon samples of the same genes and taxa.

We then compared alpha diversities between all actual specimen based and amplicon sequencing based communities. Alpha diversity (Simpson index & species richness) was calculated using the Vegan package52 in R54. Moreover, we estimated beta diversity between specimen based and sequence based communities using the Ecodist R package53. A low beta diversity indicated an accurate quantitative recovery of the whole community by sequencing. We calculated Jaccard distances as a predictor for qualitative similarity between specimen based and sequence based communities and Bray Curtis dissimilarities as a measure of quantitative similarity. Alpha and beta diversity were also calculated for the gDNA library. Replicates were generated by randomly resampling the OTU table 50 times.

We tested for an effect of our different experimental conditions on the above variables, e.g. primer degeneracy, amplicon sequence conservation, PCR cycle number and DNA template concentration. Amplicon sequence conservation and primer degeneracy were strongly associated. E.g. high primer degeneracy or high amplicon sequence conservation can interchangeably reduce amplification bias in PCR. We thus used the difference of pairwise genetic distance between taxa for marker and the primer degeneracy of that marker as an explanatory variable.

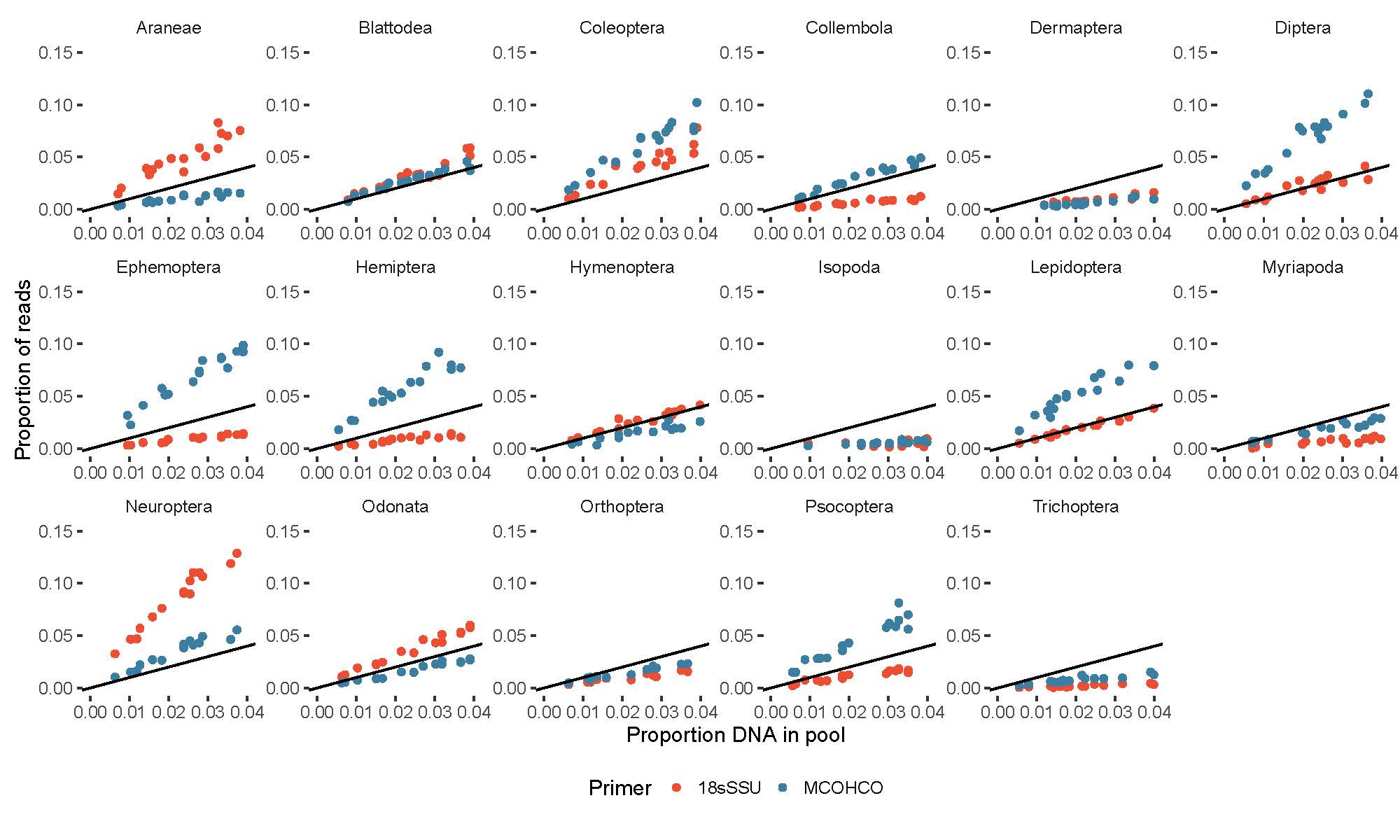
*Correcting abundance estimates*

We derived correction factors to estimate the relative abundance of taxa. Out of 16 total mock communities, we randomly chose 5 and 10 and fitted a regression line for the correlation of input DNA and recovered reads for each taxon in the community. The recovered slope of the regression was used to correct the estimated abundance of the respective taxon for the remaining six community samples. This was done by dividing the recovered proportion of reads per taxon by their corresponding taxon-specific slopes.

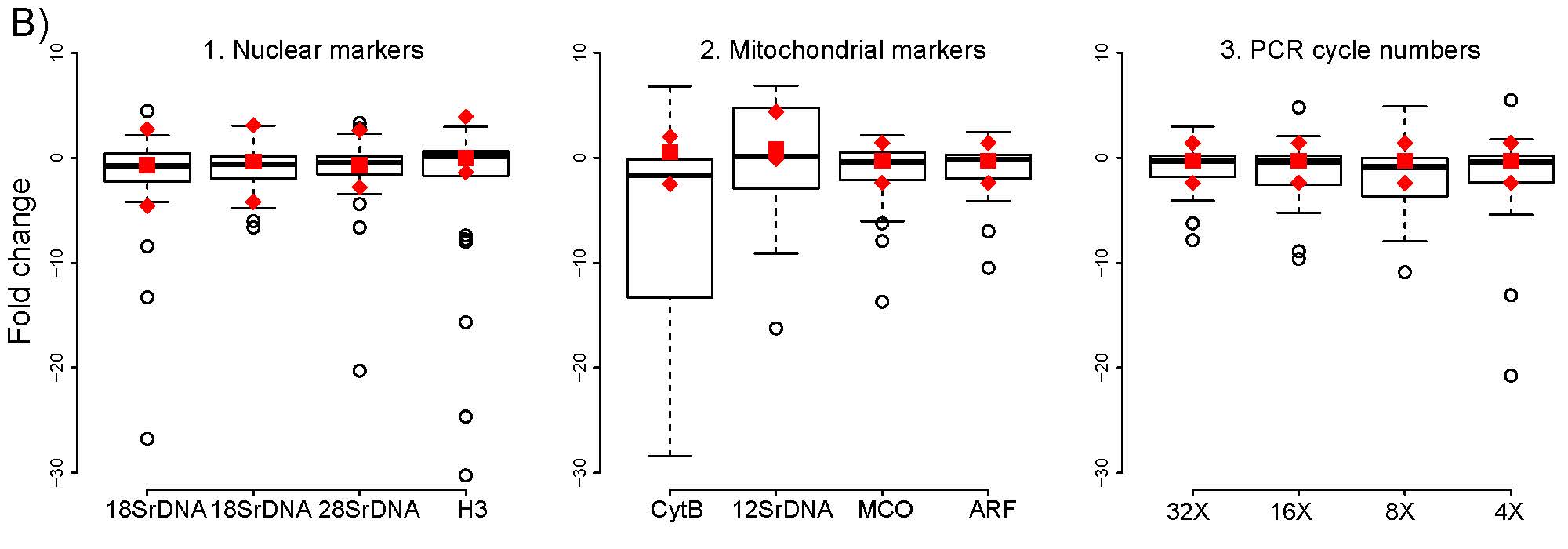
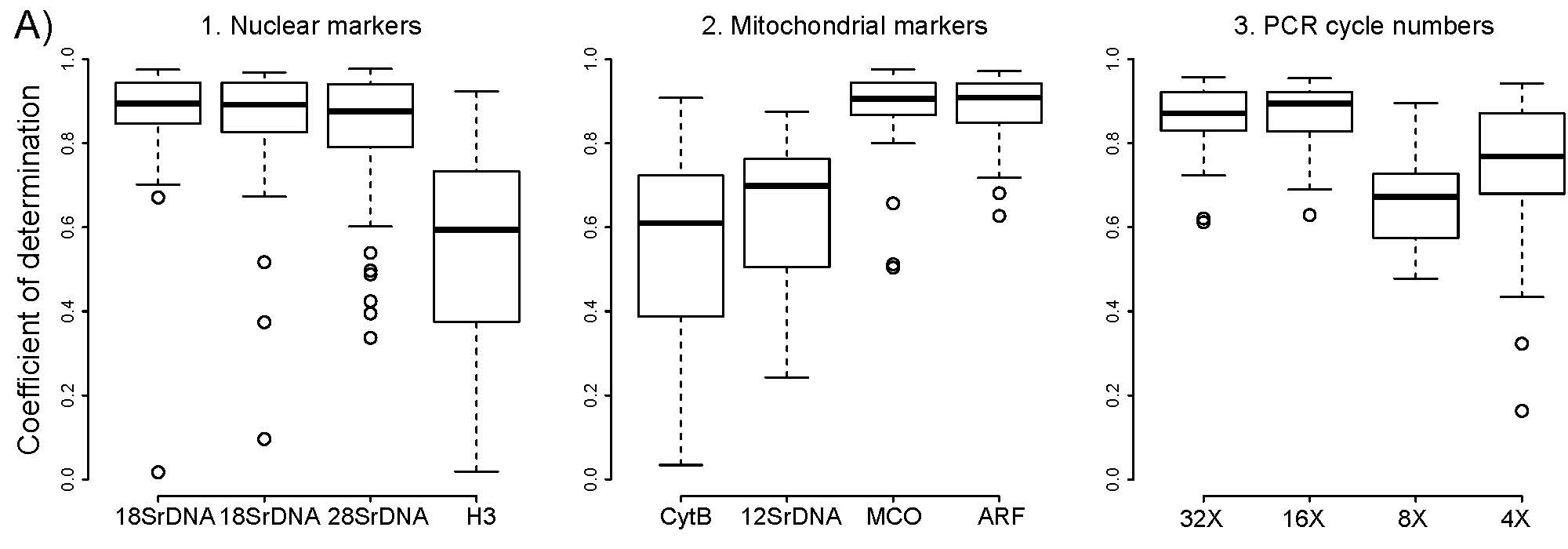
**Results**

Sequences for most samples were of high quality and coverage. After quality filtering, we recovered 8,889 reads per DNA mock community and 15,077 reads per tissue mock community on average. 2 of the 30 tissue community samples and 6 of the 220 DNA pools had to be excluded due to too low coverage. The metagenomic library yielded 835.87 x 106 bp in 3.85 x 106 sequences.

*Qualitative and quantitative community analyses*



***Figure 1*** *Proportion of input DNA against the proportion of recovered reads per taxon for various arthropod orders. Plots are based on DNA mock communities using mitochondrial COI (blue) and nuclear 18SrDNA (red) markers. 1:1 lines are in black.*



**Figure 2** **A)** Coefficient of determination (R2) of the linear association between input DNA and recovered read proportions for 43 arthropod taxa. The boxplots show R2 for nuclear and mitochondrial markers, as well as mitochondrial COI amplified with varying first round PCR cycle numbers and increased amount of DNA template during PCR. **B)** Fold change between input DNA and recovered read proportions for the same taxa and experimental conditions. Red squares indicate the median fold change for the same taxa and loci based on a gDNA library prepared without locus specific amplification. Red diamonds indicate the location of upper and lower whiskers for the boxplots of the same gDNA samples.

For most arthropod taxa, we found a positive linear association and a tight correlation (i.e., high R2) between recovered read counts and input DNA (Table 2 & Figure 1). This association was independent of the amount of the target taxon or other taxa in the mock community. The slope of the association varied across taxa and markers, for example the highly conserved nuclear ribosomal 18SrDNA, as well as the variable mitochondrial COI (Figure 1).

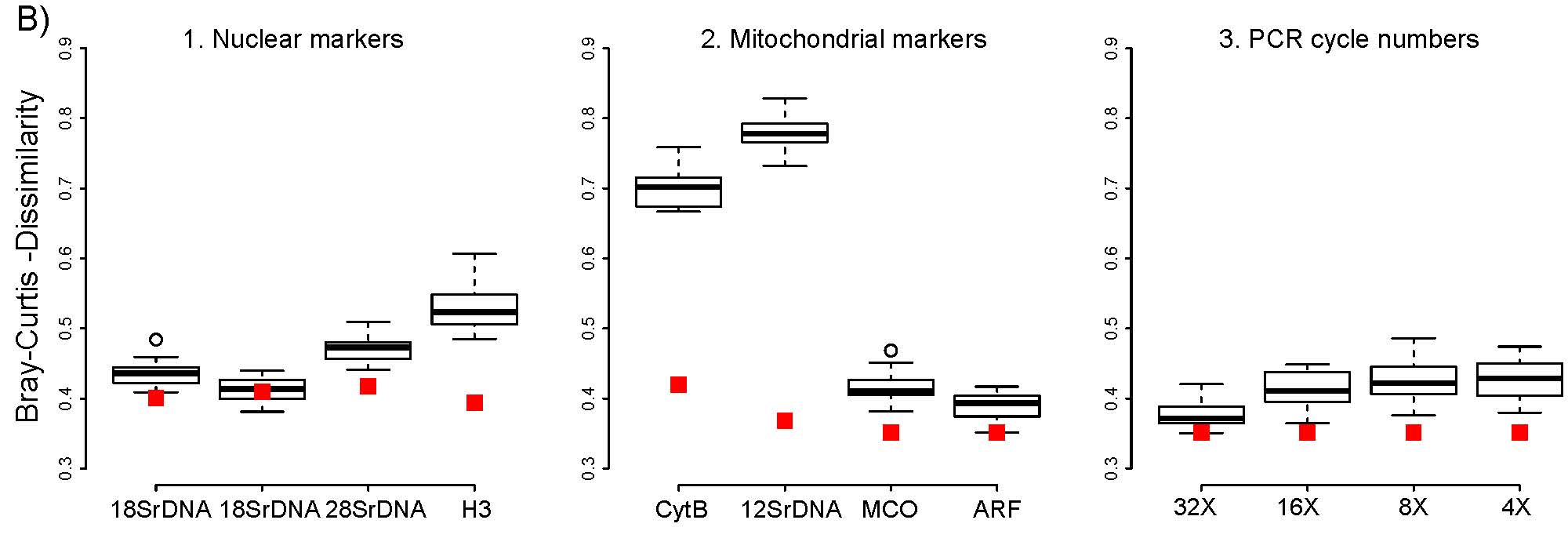
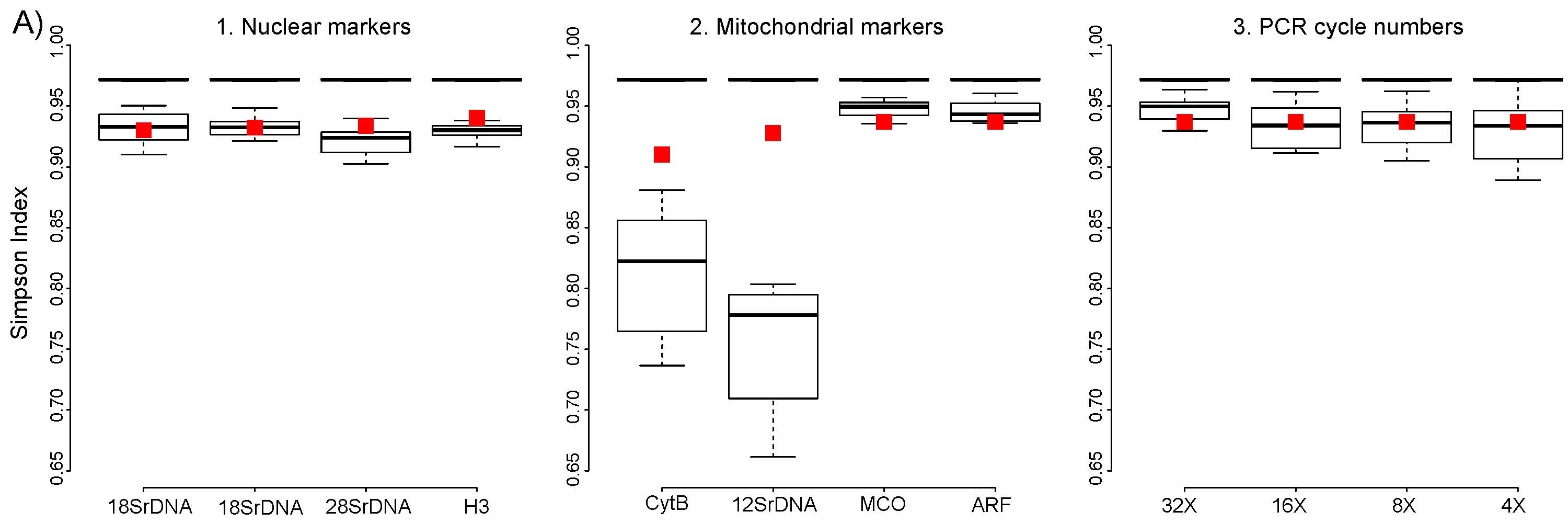
The coefficient of determination between input DNA and recovered reads was relatively high for most targeted primer pairs (Table 2, Figure 2A). The fold change between input DNA and recovered reads was mostly narrowly distributed around an actual 1:1 association (Table 2, Figure 2B). We found a strong association of primer degeneracy and amplicon sequence conservation with the coefficient of determination as well as fold change. A high conservation of the targeted amplicon or high degeneracy of the used primer pair led to significantly better correlation between input DNA and recovered reads (Figure 2 & 4A) (Pairwise Wilcoxon test, FDR corrected *P* < 0.05). At the same time, the variation of fold change was significantly reduced by sequence conservation and primer degeneracy (Levene’s test, *P* < 0.05) (Table 2, Figure 2 & 4B). The lowest R2 and highest variation for fold change was consistently found for 12SrDNA, CytochromeB and H3, which all showed a fairly high amount of sequence variation coupled with little primer degeneracy. While the two targeted COI amplicons also had a very high sequence variation, the primers used here were highly degenerate. The nuclear ribosomal markers in contrast, were highly conserved. We did not find any effect of DNA template concentration on either R2 or fold change. Fold change was also unaffected by first round PCR cycle numbers, while R2 showed a significant drop at lower PCR cycles (Pairwise Wilcoxon test, FDR corrected *P* < 0.05) (Figure 2A & B, Table 2).

The fold change between input DNA and recovered reads was very similar between amplicon libraries and our PCR free gDNA library. However, the variation of fold change was lower for the gDNA libraries (Figure 2B, Suppl. Table 1). A major difference was found for markers, which showed a significant increase of variation in the amplicon libraries (e.g. 12SrDNA, CytochromeB & H3) (Pairwise Wilcoxon test, FDR corrected *P* < 0.05). In the gDNA libraries, the variation of fold change for these loci was considerably reduced and well comparable to the other loci (Figure 2B).

Similar to R2 and fold change, alpha diversity was also strongly associated with primer degeneracy and amplicon sequence conservation (Table 2, Figure 3A & 4C, Suppl. Figure 2A). Significantly increased Simpson indexes and species richness were found for loci with high sequence conservation or highly degenerate primers (Pairwise Wilcoxon test, FDR corrected *P* < 0.05). A similar association was found for beta diversity. Jaccard distance and Bray Curtis dissimilarity between specimen based and sequence based communities decreased significantly with amplicon conservation and primer degeneracy (Pairwise Wilcoxon test, FDR corrected *P* < 0.05) (Figure 3B & 4D, Suppl. Figure 2B). A slight, but significant decrease of alpha diversity and increase of beta diversity was also observed for a decrease of PCR cycle numbers (Pairwise Wilcoxon test, FDR corrected *P* < 0.05). In contrast, DNA template concentration did not have a significant effect (Table 2, Figure 3A & B).

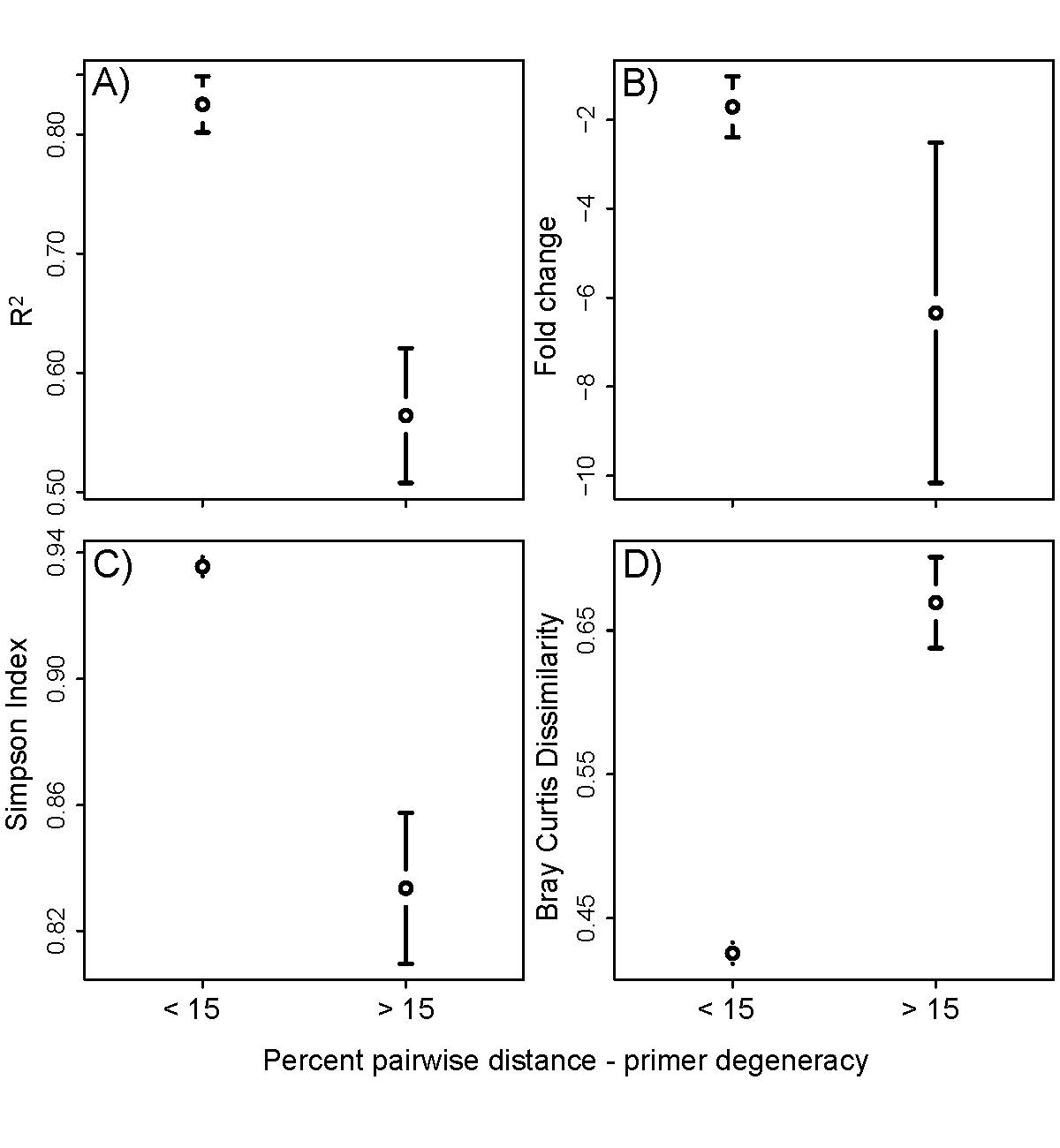
The average recovered species richness and Simpson indexes for our amplicon sequencing samples was significantly lower than the actual richness of the mock community (Pairwise Wilcoxon test, FDR corrected *P* < 0.05). However, the difference was small for most loci (Table 2, Figure 3A, Suppl. Figure 2A). We found a pronounced difference between qualitative and quantitative estimates of beta diversity. The Jaccard distances between specimen based communities and sequence based communities for most amplicons were very low (Suppl. Figure 2B). Bray Curtis dissimilarity, which incorporates taxon abundances, was significantly higher for all loci (Table 2, Figure 3B) (Pairwise Wilcoxon test, FDR corrected *P* < 0.05).

The gDNA based library generally showed slightly lower Bray Curtis dissimilarities and higher Simpson indexes than the amplicon libraries. However, a pronounced effect was only found for amplicons with high sequence variation and low primer degeneracy (e.g. 12SrDNA, CytochromeB & H3) (Figure 3A & B, Suppl. Table 1).



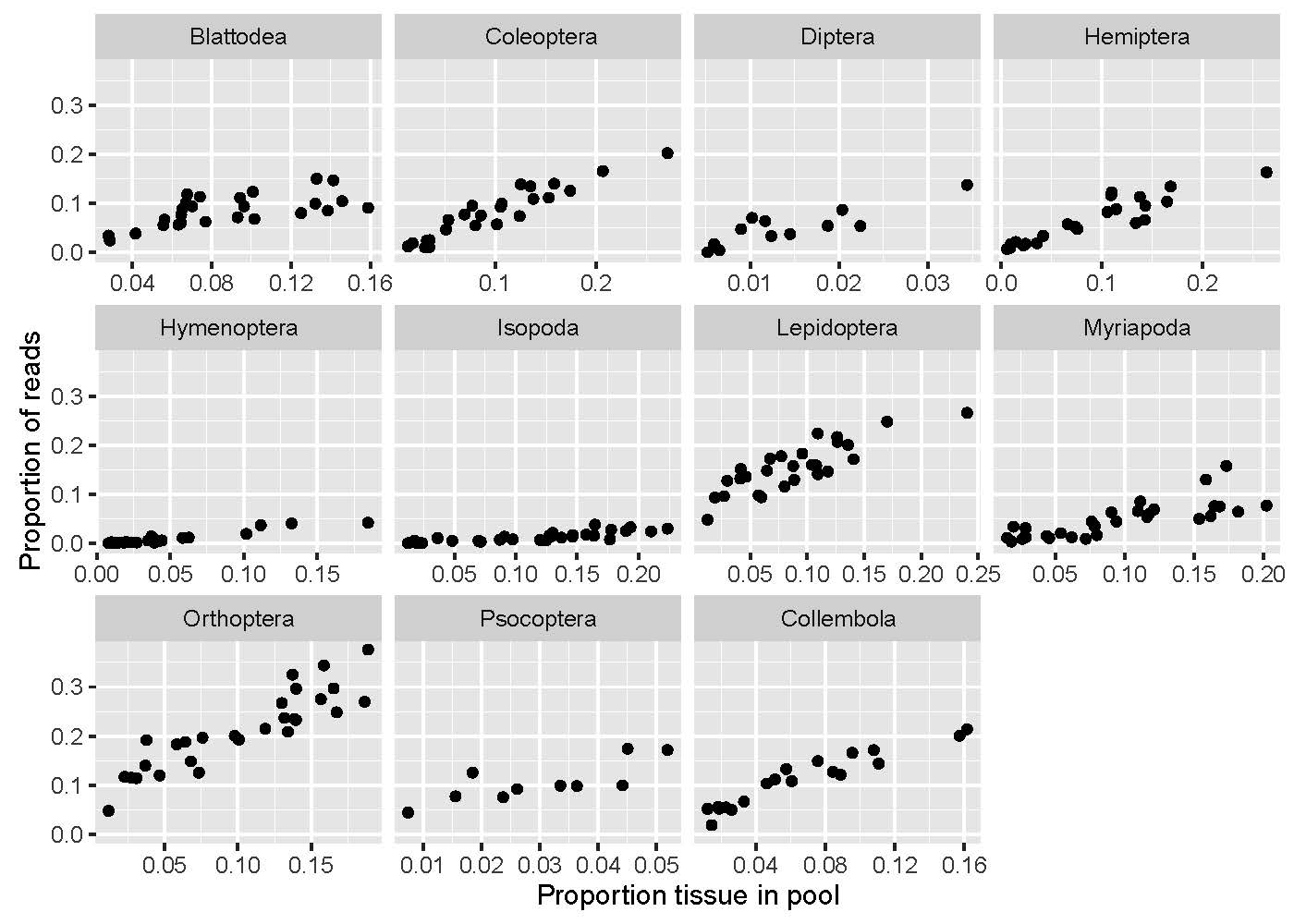
***Figure 3******A)*** *Alpha diversity (Simpson Index) of arthropod mock communities. The upper black bar shows the median alpha diversity of the actual communities based on morphospecies assignments. The boxplots show alpha diversity for the same communities based on DNA sequencing for* ***1.*** *nuclear and* ***2.*** *mitochondrial markers, and* ***3.*** *for mitochondrial COI at varying PCR cycle numbers and increased DNA template amount during PCR. Red squares indicate alpha diversity for the same loci based on a genomic DNA sample prepared without locus specific amplification.* ***B)*** *Beta diversity (Bray Curtis dissimilarity) between actual morphospecies based mock communities and sequence based analyses. The boxplots and red present the same experimental conditions as described above. Red squares indicate beta diversity for the same loci and based on a genomic DNA sample prepared without locus specific amplification.*

In summary, targeting highly conserved loci, or using highly degenerate primers, led to a considerable improvement of the association of input DNA and recovered read count and more reliable qualitative and quantitative recovery of species diversity from communities (Figure 4).



***Figure 4 A)*** *Coefficient of determination (R2) of the linear association between input DNA and recovered read abundance for two marker groups and 43 arthropod taxa. The groups are based on the difference between the average pairwise genetic distance of taxa for the according marker and the degeneracy of the primer pair used to amplify the locus. Group one comprises amplicons with a high sequence conservation and/or a high primer degeneracy. Group two comprises variable amplicons, with little primer degeneracy.* ***B)*** *Fold change between input DNA and recovered reads for the same taxa and markers.* ***C)*** *Alpha diversity (Simpson index) for the same marker groups.* ***D)*** *Beta diversity (Bray Curtis dissimilarity) between specimen based and sequence based communities for the same marker groups. The plots show the mean and the 95 % confidence interval.*

*Tissue pools*

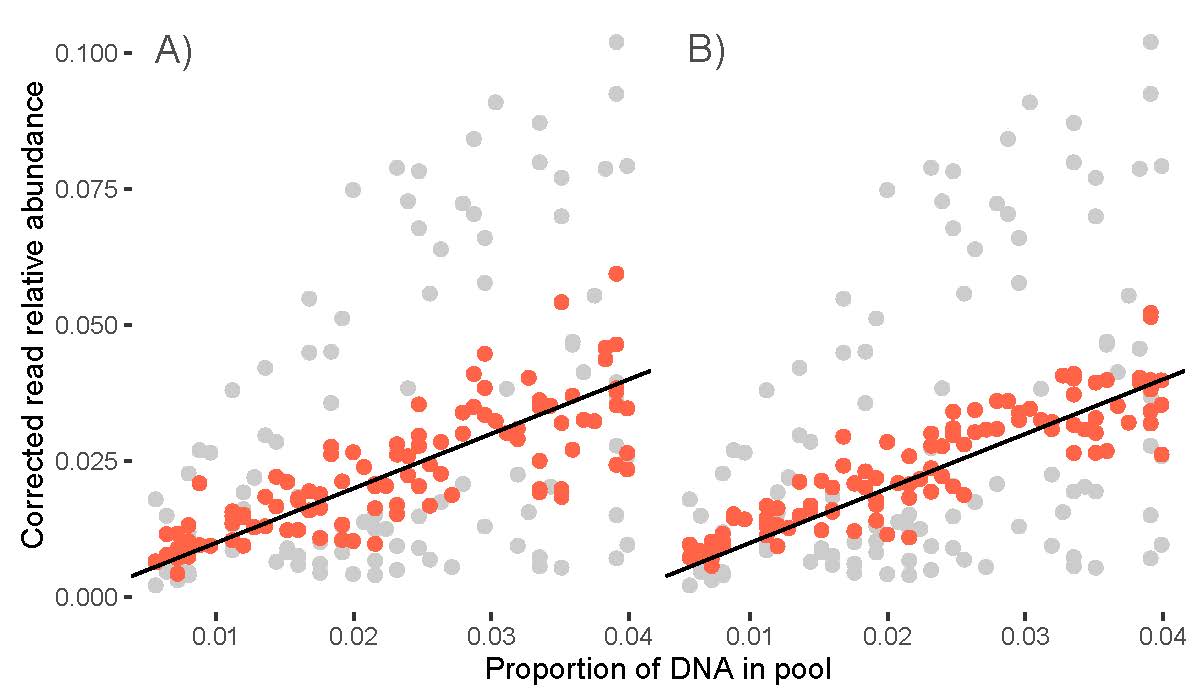


**Figure 5 A)** Exemplary associations of proportion of tissue and proportion of recovered reads per taxon. The plots are based on tissue mock communities using mitochondrial COI for various arthropod orders.

In comparison to DNA pools, tissue-based mock communities did not have any significant effect on the fold change between input DNA and recovered reads (Suppl. Figure 3B). However, they showed a lower coefficient of determination per taxon, than DNA pools (Suppl. Figure 3A) (Mann Whitney test, *P* < 0.001), e.g. the association between input tissue and recovered reads was not as predictable. Nevertheless, the amount of tissue per taxon was still well correlated with the read count (Figure 5). The fold change between input tissue amount and recovered read count was not different from that found for DNA pools (Suppl. Figure 3B). Replicates of the same taxon (Collembola, Isopoda & Myriapoda) from DNA and tissue pools, recovered very similar associations between input tissue/DNA and recovered reads (Fold change read count vs. input tissue|input DNA: FCCollembola = 1.152|1.158; FCIsopoda = 0.132|0.169; FCMyriapoda = 0.509|0.671; Suppl. Figure 4). The Simpson indexes of the actual tissue based community were significantly correlated to those derived from sequencing (R2 = 0.532) (Suppl. Figure 3C). Moreover, the recovered Bray Curtis dissimilarities between specimen and sequence based communities were not higher than those found for our DNA pools (Suppl. Figure 3D).

*Correcting abundance estimates*

Each taxon shows a predictable fold change between the proportion of input DNA and recovered reads. But, due to taxon-specific slopes, a simple association of the proportion of input DNA and recovered reads for all taxa in six mock communities suggests no correlation (R2 = 0.09; *P* > 0.05; Figure 6). By using 5 mock communities to derive taxon-specific correction factors, a significant correlation is found (R2 = 0.59; *P* < 0.05; Figure 6A). This correlation is increased when 10 mock communities are used to derive corrections factors (R2 = 0.82, *P* < 0.05; Figure 6B). The amount of input DNA can thus be fairly accurately predicted from mock communities for most taxa. Read abundance correction also led to significantly decreased Bray Curtis dissimilarities between specimen based and sequence based communities (Pairwise Wilcoxon test, FDR corrected *P* < 0.05) (Supplementary Figure 5).



**Figure 6** Effect of correcting read abundances on quantitative taxon recovery. Uncorrected association of actual abundance and recovered read proportion for 43 arthropod taxa (grey dots) and after applying the taxon specific slope of the association between input DNA and read count as correction factor for the read abundance (red dots) using **A)** 5 mock communities or **B)** 10 mock communities to derive the correction factors. The black lines represent the 1:1 lines.

**Discussion**

*Mitigating amplification bias in metabarcoding*

We show that metabarcoding accurately recovers the qualitative species composition of diverse arthropod communities, giving rise to very similar alpha diversities and low Jaccard distance between specimen based and sequence based community samples. However, we found pronounced quantitative bias in our sequence based community analyses. This bias can partly be attributed to differential amplification due to priming efficiency during PCR. We show that this amplification bias can be greatly alleviated by using degenerate primers and/or targeting amplicons with a high sequence conservation, but even under optimized PCR conditions or completely avoiding locus specific amplification in our metagenomic library, we found considerable differences in read abundances across taxa. These differences are possibly caused by the six cycles of indexing PCR, but considering the fact that indexing PCR primers for all amplicons are targeting exactly the same priming sequence, priming bias alone seems unlikely. Factors inherent to the target sequence could cause amplification bias during indexing PCR, e.g. length variation or GC bias. However, these factors should affect the first round PCR as well and a reduction of first round PCR cycle number did not yield any effect. Copy number variation of the target loci13 is thus the most likely reason for read abundance differences. All the amplified loci in our study are present in multiple copies in each cell. Mitochondrial copy number even varies considerably between different organs in a single organism39. And different arthropod taxa carry different ratios of tissue types with different mitochondrial content, e.g. muscles in flying and non-flying species.

PCR-free analyses have been suggested as possible means for quantitative community analysis. This approach circumvents amplification bias28 and has been shown to result in better recovery of taxa from diverse communities40. Indeed, even our best primer combinations did not recover all taxa from mock communities, as indicated by consistently lower species richness of sequence based over actual communities. This suggests PCR free approaches as the method of choice for exhaustive community analyses, where the recovery of all taxa is of critical importance. However, quantitative analyses using PCR free methods will be similarly sensitive to CNVs of the target genes. Also, an amplicon sequencing-based approach is much more cost efficient and involves a greatly simplified workflow, making it the method of choice for large scale community analyses.

*Abundance estimates by metabarcoding*

Due to biases in read abundance, metabarcoding does not allow estimating actual species abundances. However, despite the observed taxonomic bias of read abundances, the amount of recovered reads was correlated in a very predictable way with the amount of input DNA. Similar results have been found for microorganisms41,42. The correction of read abundances can thus yield an approximation of taxon abundances in a community17,32. For a quantitative analysis by metabarcoding, the expected taxa in the studied system and the taxon specific PCR amplification need to be known. The identification of correction factors involves considerable effort and is not feasible in unknown ecosystems or for simple exploratory work. But for large scale and long-term studies in one ecosystem, the effort could pay off. As every primer combination results in different fold change for different taxa, it is advisable to focus on only a few or even a single marker for such quantitative optimization. This approach seems particularly suitable for comparative studies on abundance changes of a subset of target taxa, e.g. invasive species across different sites.

*Metabarcoding and mitochondrial COI – a perfect match?*

Even nuclear ribosomal markers with highly conserved priming sites did not yield significantly better qualitative or quantitative results than degenerate COI primers. In contrast to nuclear rDNA, COI is more variable and can distinguish even recently diverged species. While 18SrDNA and 28SrDNA performed well in our analysis, they may be too conserved for many barcoding applications24. Our study was mostly based on quite divergent taxa. These are still picked up using conserved markers like nuclear rDNA. However, nuclear rDNA would likely fail to distinguish recently diverged species. Compared to other mitochondrial markers, COI is distinguished by an exceptionally well-developed reference database25, which often allows species identification. Recent studies suggested alternative primers to COI22,23. Indeed, different markers are advisable for certain taxa; for example, we were unable to amplify some Acari and Hymenoptera with COI. Mitochondrial markers bring along problems such as NUMTS12 and their genealogy can be strongly affected by bacterial infections43 or paternal gene flow44. Hence, a suitable nuclear marker would be recommendable for future studies. The internal transcribed spacers of the ribosomal cluster are promising targets, as they have already been successfully applied in fungal taxonomy45. With more genomic data available, a multitude of novel markers will be discovered in the coming years.

**Conclusion**

PCR amplification bias can be significantly mitigated by standardized PCR conditions. Apart from PCR bias, copy number variation of the target locus could contribute to read abundance differences between taxa, affecting PCR-free and amplicon-based approaches alike. A good understanding of the taxonomic composition of a community provided, taxon-specific correction factors can be applied to derive abundance estimates.

**Acknowledgements**

We thank Edward Greg Huang, Noriyuki Suzuki and Luis Cayetano for help during fieldwork. Jingfei Zhang helped prepare mock communities. Susan Kennedy thankfully provided help with the taxonomic identification of spiders.Summer Ahmed and Heather Dang are acknowledged for granting access to a microscale. Anna Sellas performed the Illumina sequencing. Forrest Freund and Ingrid Jordon-Thaden provided access to the Genogrinder. We acknowledge the Hawaii Department of Land and Natural Resources for granting us permission to sample arthropods in native rainforests. HK was funded by a postdoctoral fellowship of the German Science Foundation (DFG). This work is funded by a National Science Foundation grant (NSF DEB 1241253) to RGG. The authors declare no conflict of interest.

**Data accessibility**

The following data will be made available in the Dryad Digital Repository upon acceptance of the manuscript:

1. Read files for all analyzed sequences

2. Analysis tables containing DNA or tissue proportions and read counts for each taxon and each mock community

**Author contributions**

HK devised the study, HK & MW Collected the data, HK, JYL & AJR Analyzed the data, HK, JYL, RGG & WBS wrote the manuscript.

**References**

1. Meier, R., Wong, W., Srivathsan, A., & Foo, M. $1 DNA barcodes for reconstructing complex phenomes and finding rare species in specimen‐rich samples. *Cladistics*. **32**, 100-110 (2016).

2. Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. Towards next‐generation biodiversity assessment using DNA metabarcoding. *Molecular ecology*. **21**, 2045-2050 (2012).

3. Leray, M., & Knowlton, N. DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences*. **112**, 2076-2081 (2015).

4. Gibson, J. et al. Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *Proceedings of the National Academy of Sciences*. **111**, 8007-8012 (2014).

5. Ji, Y. et al. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology letters*.**16**, 1245-1257 (2013).

6. Krehenwinkel, H., Kennedy, S., Pekár, S., & Gillespie, R. G. A cost‐efficient and simple protocol to enrich prey DNA from extractions of predatory arthropods for large‐scale gut content analysis by Illumina sequencing. *Methods in Ecology and Evolution.* (2017).

7. Shokralla, S. et al. Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific reports*. ***5*** (2015).

8. Shokralla, S., Spall, J. L., Gibson, J. F., & Hajibabaei, M. Next‐generation sequencing technologies for environmental DNA research. *Molecular ecology*. 21, 1794-1805 (2012).

9. Kress, W. J., García-Robledo, C., Uriarte, M., & Erickson, D. L. DNA barcodes for ecology, evolution, and conservation. *Trends in Ecology & Evolution*. **30**, 25-35 (2015).

10. Elbrecht, V., & Leese, F. Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass—sequence relationships with an innovative metabarcoding protocol. *PloS one*. **10**, e0130324 (2015).

11. Stadhouders, R. et al. 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**, 109-117 (2010).

12. Bensasson, D., Zhang, D. X., Hartl, D. L., & Hewitt, G. M. Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in ecology & evolution*. **16**, 314-321 (2001).

13. Angly, F. E. et al. CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome*. **2**, 1 (2014).

14. Aird, D. et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome biology*. **12**, 1 (2011).

15. Polz, M. F., & Cavanaugh, C. M. Bias in template-to-product ratios in multitemplate PCR. *Applied and environmental Microbiology*. **64**, 3724-3730 (1998).

16. Gotelli NJ, Chao A. Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. In Levin S, ed, Encyclopedia of Biodiversity, Waltham, MA, USA. 195–211 (2013)

17. Saitoh, S. et al. A quantitative protocol for DNA metabarcoding of springtails (Collembola) 1. *Genome*. **59**, 705-723 (2016).

18. Tang, M. et al. High‐throughput monitoring of wild bee diversity and abundance via mitogenomics. *Methods in Ecology and Evolution*. **6**, 1034-1043 (2015).

19. Hebert, P. D., Ratnasingham, S., & de Waard, J. R. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London B: Biological Sciences*. **270**, 96-99 (2003).

20. Drummond, A. J. et al. Evaluating a multigene environmental DNA approach for biodiversity assessment. *GigaScience*. **4**, 1 (2015).

21. Clarke, L. J., Soubrier, J., Weyrich, L. S., & Cooper, A. Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. *Molecular ecology resources*. **14**, 1160-1170 (2014).

22. Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., & Taberlet, P. DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology letters*. **10**, 20140562 (2014).

23. Elbrecht, V. et al. Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects. *PeerJ*. **4**, e1966 (2016).

24. Tang, C. Q. et al. The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. *Proceedings of the National Academy of Sciences*. **109**, 16208-16212 (2012).

25. Ratnasingham, S., & Hebert, P. D. BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). *Molecular ecology notes*. **7**, 355-364 (2007).

26. Yu, D.W. et al. Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*. **3**, 613-623 (2012).

27. Suzuki, M. T., & Giovannoni, S. J. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and environmental microbiology*. **62**, 625-630 (1996).

28. Crampton-Platt, A., Douglas, W. Y., Zhou, X., & Vogler, A. P. Mitochondrial metagenomics: letting the genes out of the bottle. *GigaScience*. **5**, 1 (2016).

29. Gómez‐Rodríguez, C., Crampton‐Platt, A., Timmermans, M. J., Baselga, A., & Vogler, A. P. Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. *Methods in Ecology and Evolution*. **6**, 883-894 (2015).

30. Zhou, X. *et al.* Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *Gigascience*. **2**, 1 (2013).

31. Heid, C. A., Stevens, J., Livak, K. J., & Williams, P. M. Real time quantitative PCR. *Genome research*. **6**, 986-994 (1996).

32. Thomas, A. C., Deagle, B. E., Eveson, J. P., Harsch, C. H., & Trites, A. W. Quantitative DNA metabarcoding: improved estimates of species proportional biomass using correction factors derived from control material. *Molecular ecology resources* (2015).

33. Evans, N. T. et al. Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Molecular ecology resources*. **16**, 29-41 (2016).

34. Kembel, S. W., Wu, M., Eisen, J. A., & Green, J. L. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput Biol*. **8**, e1002743 (2012).

35. Lange, V. Et al. Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. *BMC Genomics*. **15**, 63 (2014).

36. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*. **30**, 614-620 (2014).

37. Gordon, A. & Hannon, G.J. Fastx-toolkit. *Computer Program Distributed by the Author, Website http://hannonlab. cshl. edu/fastx\_toolkit/index. html [accessed 2014–2016]* (2010).

38. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*. **27**, 2194-2200 (2011).

39. Wiesner, R. J., Rüegg, J. C., & Morano, I. Counting target molecules by exponential polymerase chain reaction: copy number of mitochondrial DNA in rat tissues. *Biochemical and biophysical research communications*. **183**, 553-559 (1992).

40. Shokralla, S., Gibson, J., King, I., Baird, D., Janzen, D., Hallwachs, W., & Hajibabaei, M. Environmental DNA Barcode Sequence Capture: Targeted, PCR-free Sequence Capture for Biodiversity Analysis from Bulk Environmental Samples. *bioRxiv*. 087437 (2016).

41. Sohn, M. B., Lingling A., Naruekamol P., und Qike L. Accurate genome relative abundance estimation for closely related species in a metagenomic sample. *BMC Bioinformatics.* **15**, 242 (2014).

42. Giner, C. R. et al. Environmental Sequencing Provides Reasonable Estimates of the Relative Abundance of Specific Picoeukaryotes. *Applied and Environmental Microbiology*. AEM.00560-16 (2016).

43. Hurst, G. D., & Jiggins, F. M. Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. *Proceedings of the Royal Society of London B: Biological Sciences*. **272**, 1525-1534 (2005).

44. Chen, S. F., Jones, G., & Rossiter, S. J. Sex‐biased gene flow and colonization in the Formosan lesser horseshoe bat: inference from nuclear and mitochondrial markers. *Journal of Zoology*. **274**, 207-215 (2008).

45. Nilsson, R. H., Ryberg, M., Abarenkov, K., Sjökvist, E., & Kristiansson, E. The ITS region as a target for characterization of fungal communities using emerging sequencing technologies. *FEMS Microbiology Letters*. **296**, 97-101 (2009).

46. Leray, M. et al. 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**, 1-14 (2013).

47. Barraclough, T. G., Hogan, J. E., & Vogler, A. P. Testing whether ecological factors promote cladogenesis in a group of tiger beetles (Coleoptera: Cicindelidae). *Proceedings of the Royal Society of London B: Biological Sciences*. **266**, 1061-1067 (1999).

48. Kocher, T. D. et al. Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. *Proceedings of the National Academy of Sciences*. **86**, 6196-6200 (1989).

49. Fonseca, V. G. et al. Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature communications*. **1**, 98 (2010).

50. Machida, R. J., & Knowlton, N. PCR Primers for metazoan nuclear 18S and 28S ribosomal DNA sequences. *PLoS one*. **7**, e46180 (2012).

51. Colgan, D. J. et al. (1998) Histone H3 and U2 snRNA DNA sequences and arthropod molecular evolution. *Australian Journal of Zoology*. **46**, 419-437.

52. Oksanen J, Kindt R, Legendre P, O’Hara B, Simpson GL, Solymos P, Stevens MHH, Wagner H (2007) The vegan package. *Community Ecology Package*. **10**, 631-637.

53. Goslee SC, Urban DL (2007) The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software*, **22**, 1-19.

54. R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

**Figures and Tables**