**A multi locus test for qualitative and quantitative metabarcoding of diverse arthropod communities using Illumina amplicon sequencing**

**Reduction of PCR priming bias for optimized quantitative and qualitative metabarcoding from Illumina sequencing of diverse arthropod communities**

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

**Abstract**

1. Next generation sequencing based metabarcoding studies promise the rapid and cost efficient analysis of species compositions in whole ecosystems. It is still disputed, if species abundances can be derived from metabarcoding data. Amplicon based approaches using the barcode marker Cytochrome Oxidase subunit 1 (COI) are believed to suffer from pronounced PCR priming bias. PCR free approaches, have been suggested to mitigate these problems, but they come with a considerable increase of workload and cost.

2. Here, we test 8 different nuclear and mitochondrial markers for their potential to recover species richness and abundance in mock communities of arthropod taxa from 14 orders. We test numerous factors for their potential to alleviate priming bias, including sequence conservation, primer degeneracy, PCR template concentration and a reduction of PCR cycles. We also develop correction methods to reconstruct abundance of taxa.

3. Priming bias clearly affects barcoding results, but can be alleviated with proper experimental conditions. Using appropriate correction methods, amplicon sequencing can predict species richness and abundance. Copy number variation emerged as another potential factor driving read abundance differences, which would affect amplicon based and PCR free approaches alike. For most arthropod taxa, degenerate COI primers provide excellent recovery, reappraising COI as useful barcoding marker.

4.A good understanding of the taxonomic composition of a community provided, amplicon sequencing is a cost efficient and reliable approach for large-scale analyses of species richness and possibly abundance. The great simplicity and cost efficiency of Illumina amplicon sequencing, makes it the method of choice for large scale community analyses.

**Introduction**

Next generation sequencing technology has ushered in a revolution in evolutionary biology and ecology. This revolution has not passed by taxonomy and spurred various studies in the field of molecular barcoding. Next generation sequencing based barcoding comes with little work load, is cost efficient and requires comparably little taxonomic expertise. The resulting leap in throughput allows large-scale barcoding studies in whole ecosystems (Taberlet et al. 2012; Leray & Knowlton 2015; Gibson et al. 2014; Ji et al. 2013). The recovery of species richness, food web structure, cryptic species, hidden diversity like internal parasitoids, and the identification of juveniles from bulk samples promise unprecedented new insights into ecosystem function and assembly (Krehenwinkel et al. 2016; Shokralla et al. 2015; Shokralla et al. 2012; Kress et al. 2015; Kartzinel et al. 2015; Morinière et al. 2016). A critical, but not yet sufficiently understood application of such metabarcoding approaches is the potential estimation of species abundances (Elbrecht & Leese, 2015; Shokralla et al. 2016).

The commonly used PCR based approaches suffer from pronounced priming bias. Sequence divergence in priming sites, copy number variation of the target gene and several other factors can lead to flawed abundance estimates (Clarke et al. 2014; Deagle et al. 2014; Polz & Cavanaugh 1998). A short stretch of the mitochondrial Cytochrome Oxidase Subunit I (COI) gene is commonly used as a barcoding marker in animals (Hebert et al. 2003 & 2004; Folmer et al. 1994). 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 substitute for COI (Drummond et al. 2015; Clarke et al. 2014; Deagle et al. 2014; Elbrecht et al. 2016). Such novel markers however, are usually less variable (Tang et al. 2012) and no well-developed reference databases are available for them (Ratnasingham & Hebert, 2007). Another solution is found by using degenerate COI primers, which mitigate PCR bias and allow an amplification of a broader taxonomic range (Yu et al, 2012) or by designing taxon specific primers. PCR free approaches have also been suggested. The direct sequencing of genomic DNA or sequence capture of barcodes avoids PCR amplification and is thus assumed to provide more accurate predictions of abundance in communities (Crampton-Platt et al. 2016; Gomez-Rodriguez et al. 2015; Zhou et al. 2013, Shokralla et al. 2016). However, PCR free methods come with a considerable increase in workload and processing cost, e.g. for enrichment and library preparation. And while they mitigate priming bias, they are also sensitive to copy number variation of the target locus.

As PCR exponentially amplifies DNA templates, priming bias should significantly increase with the number of PCR cycles. A low number of PCR cycles should thus mitigate bias and allow for a more accurate correlation of input DNA and recovered reads (Suzuki & Giovannoni 1996; Polz & Cavanaugh 1998). Copy number variation instead, should be unaffected by cycle number and pose a constant problem to the study of abundance. Aside from reduced PCR cycle numbers, increased concentrations of DNA template were suggested to alleviate priming bias (Suzuki & Giovannoni 1996; Polz & Cavanaugh 1998).

Even though PCR bias is a problem in quantitative community analyses, PCR has proven as highly predictable and accurate, as evident by applications like quantitative PCR (Heid et al. 1996). Assuming that the PCR for a taxon in a community sample is not affected by other species in the extraction, the proportion of input DNA should be tightly correlated to the proportion of recovered reads. Priming bias or copy number variation should only affect the slope of this correlation. If this slope is known for taxa in a community, it might be possible to predict their relative abundance by using correction factors (Thomas et al. 2015; Angly et al. 2014; Saitoh et al. 2016; Evans et al. 2016). As PCR bias is induced by sequence divergence, it should also be similar in closely related taxonomic groups, as has been shown in bacteria (Angly et al. 2014; Kembel et al. 2012). Hence, similar correction methods could possibly be derived for closely related taxa allowing for community level abundance estimates, without calibrating the correction model for each taxon.

Considering this background, we hypothesize that priming bias can be alleviated by **1.** choosing barcode markers with high sequence conservation or high levels of primer degeneracy. **2.** by reducing the PCR cycle number during library preparation. **3.** by increasing the template concentration and **4.** by identifying the taxon specific PCR bias and correcting for it.

To test these hypotheses, we sequenced mock communities of a taxonomically diverse set of Hawaiian and Californian arthropods on the Illumina MiSeq. We aim to optimize quantitative assessments of arthropod biodiversity, e.g. taxon recovery from mixed communities. We also explicitly test for the possibility of qualitative analyses, e.g. abundance estimates from read counts. Using eight different barcode markers, we test for the effect of various factors on qualitative and quantitative community characterization. In particular, we explore the effect of sequence conservation, primer degeneracy and PCR cycle number.

**Methods**

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

Arthropod samples were collected in native rainforests on the Hawaiian Islands Maui and Hawaii and oak forest near the University of California Berkeley in spring 2015 and 2016 by beating vegetation. Specimens were stored in 99-% ethanol, morphologically identified to order and then assigned to morphotypes, which likely correspond to separate species. We extracted DNA from 44 of these morphotypes, representing 19 orders (in the classes Arachnida, Crustacea, Insecta & 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 Qbit Fluorometer with the high sensitivity assay (Thermo Scientific, Waltham, USA) and each sample subsequently diluted to a final concentration of 15 ng/µl. We prepared 23 mock communities by pooling randomized volumes of each of the 44 samples. Each pool contained all samples in randomized volumes from 0.7 to 5 µl per sample and increments of 0.1 µl.

We chose 8 different 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 each of 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 variable mitochondrial markers (See Table 2 for degeneracy and genetic distances). All primer pairs amplified sequences shorter than 500 bp to achieve an overlap of between paired 2 x 300 bp Illumina MiSeq reads.

PCRs were run in 10 µl volumes using the Qiagen Multiplex PCR kit according to the manufacturer’s protocols using 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 gradient PCRs. A first round PCR was run with 32 cycles using tailed primers. On these tails, a second indexing PCR was performed with 6 cycles, to introduce Illumina TruSeq adapters and dual indexes. The basic PCR layout followed that described in Lange et al (2014). We amplified 16 of the mock communities for each of the 8 markers. After each round of PCR, the product was cleaned up from remaining primer sequences by 1X AMpure XP Beads according to the manufacturer’s protocol (Beckman Coulter, Indianapolis, USA). The final libraries were quantified with a Qbit Fluorometer as described above and then all samples pooled in equimolar amounts.

*Tissue mock communities*

To test the applicability of our method under real conditions, we generated mock communities from tissue pools of different Hawaiian taxa. Specimens were identified to morphotypes as described above and defined amounts of tissue of approximately 20 taxa combined into 30 mock communities. Due to 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. To prepare the mock communities, specimens were dried for 1 hour on kimwipes at room temperature. Depending on the size of the specimens, they were either added completely 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 450 µl of lysis buffer and a 5-mm stainless steel bead and disrupted by shaking them 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. PCR and library preparation were performed as described above.

*PCR cycle reduction experiment*

Additionally, we ran a series of PCRs with varying cycle numbers. All 23 DNA mock communities were used for this experiment. Two PCRs were run as described above with the primer combination ArF1/Fol-degen-rev. 4 µl of template DNA (corresponding to 60 ng) were used in a 10 µl PCR 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 of indexing PCRs of 26, 22, 14 and 6 cycles, to make the total number of cycles 38 across experiments. Assuming that PCR priming bias leads to inaccurate predictions of species abundance in community samples, a low number of first round PCRs should greatly reduce this bias. As the indexing PCR is based on the same priming sites (5’-tails introduced in the first round PCR) for all samples, PCR bias should be of minor concern here (See Supplementary Figure 1 for a visualization of the concept). PCR and library preparation were performed as described above.

*Sequencing and sequence analysis*

The final pools were sequenced on a flow cell of an Illumina MiSeq, using V3 chemistry and 2 x 300 bp reads according to the manufacturer’s protocol (Illumina, San Diego, USA). The resulting paired reads were assembled using PEAR (Zhang et al. 2014) with a minimum overlap of 50 and a minimum quality of 30. The assemblies were quality filtered using the FastX Toolkit (Gordon & Hannon 2010) with a minimum of 90-% of bases ≥ Q30 and transformed into Fasta format. Samples for the separate primer pairs were demultiplexed by marker using the forward and reverse primer sequences as indexes and primer sequences trimmed from the resulting Fasta files using a custom UNIX script. Each of the previously generated alignments of reference specimens per marker was used to create BLAST databases. Using BLASTn (Altschul et al. 1990) 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 used a minimum overlap of 95-% and a sequence similarity of 98-% to assign a sequence to the reference. Only community samples with more than 1,000 reads were used for the subsequent analyses. 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 USEARCH (Edgar et al. 2011) with a minimum similarity of 97-%. The taxonomy of the resulting OTU centroid sequences was assigned as described above using BLAST. Taxon recovery and read abundance to input tissue proportion were analyzed like described above for the DNA pools.

*Qualitative and quantitative community analyses*

We analyzed three dependent variables from our experiments. **1.** We quantified the proportion (*α*) of samples from each taxon, that could be recovered from sequencing each mock community. This measure allowed a general prediction of the suitability of the markers for qualitative metabarcoding purposes.

By linear regression of the proportion of reads per specimen against the proportion of its DNA in each mock community, we identified **2.** the coefficient of determination (*R2*) and **3.** the slope (*m*) of the according regression line for each specimen and marker. Thereby *R2* served as a measure of correlation/predictability of the amount of input DNA per taxon vs. the proportion of reads to recover. 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 for that taxon.

In our three experiments, we tested the effect of different explanatory variables on *α, R2* & *m,* (Table 2) using generalized linear models in R (R Core Team 2016).

**1.** Marker type, conservation and primer degeneracy: Based on our experiment with eight different PCR markers, we tested for an effect of marker type (nuclear or mitochondrial), primer degeneracy and marker sequence conservation. Primer degeneracy and marker conservation were strongly associated, e.g. highly conserved priming sites and saturation of degenerate sites in the primer had similar effects. We thus combined both variables. We used the difference of pairwise genetic distance between all taxa in our mock community and the proportion of degenerate priming sites of the respective marker. This difference was assigned into one of three categories for each of the 8 targeted markers. These categories comprised low (≤ 10), medium (> 10 ≤ 20) and high (> 20), corresponding to increasing sequence variability and/or decreasing proportion of degenerate priming sites.

**2.** DNA vs. tissue pools: DNA pools constitute idealized mock communities. Tissue pools, on the other hand, can be affected by additional biases, e.g. taxon or tissue specific DNA content. In our second experiment, we estimated the effect of DNA pools and tissue pools on *α, R2* & *m*.

**3.** PCR cycle reduction and DNA concentration: Our third experiment allowed us to test for an effect of a reduction of first round PCR cycles and increasing concentration of DNA template on *α, R2* & *m*.

*Correcting abundance estimates using linear and Bayesian approaches*

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 and mock community by the taxon specific slope.

We also used a Dirichlet-multinomial model to determine how reliably the number of reads recovered for a given taxon could be predicted by that taxon’s abundance (either amount of DNA or tissue) in the mock community. These models were implemented in a Bayesian framework owing their hierarchical nature. The hierarchy results from modeling read capture as a two step process: first different amounts of input DNA interact with primer affinity and copy number to produce a latent probability that a read will be captured for a given taxon, modeled by a Dirichlet distribution; next these probabilities enter into a multinomial process whereby the total number of reads are stochastically assigned to taxa based on their latent read probability. This model was written and fit using the NIMB. Based on the estimated parameters of the model relating input DNA and recovered reads, we derived a correction method to estimate the predicted abundance of taxa. Given a vector of reads assigned to each taxon (*R*) and the vector of estimated parameters from the Dirichlet-multinomial model (*α*), the vector of predicted abundances (*x*) is the solution to the system

where *K* is the total number of taxa, *N* is the total number of reads and *Xtotal* is the total amount of input (either total amount of DNA as in this study, or total number of individuals). Writing this system in matrix form reveals that the solution is the lead eigen vector (normalized to sum to *Xtotal*) of the matrix:

where *IK* is the identity matrix of dimension *K*.

We tested this correction method using the posterior estimates from the 4 cycle PCR experiment to predict the input of the 32 cycle experiment, both using the nimble package (de Valpine et al. 2016) in the R language.

**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.

*Qualitative and quantitative community analyses*

**1.** Marker type, conservation and degeneracy: With the exception of 12SrDNA, all of the eight analyzed markers allowed a high recovery of taxa from mock communities (Table 3). For nearly all analyzed arthropod taxa, we also found a positive linear association and high predictive power of read counts and input DNA (Table 3 & Figure 1, Supplementary Figure 2). This association was independent of the amount of the target taxon or other taxa in the mock community. The slope of the association was variable between different taxa and different markers (Figure 1).

The marker type did not have any significant effect on the three tested dependent variables (*α, R2* & *m*). Mitochondrial and nuclear markers allowed for an equally high recovery of taxa, and did not yield different coefficients of determination or slopes.

In contrast, the difference of pairwise sequence divergence and primer degeneracy had a strong effect on the dependent variables (Figure 2). A lower pairwise sequence divergence and/or an increased degree of primer degeneracy resulted in significantly increased recovery of taxa (Figure 2A) and a higher *R2* (Figure 2B) of the association of input DNA and recovered reads. Decreasing sequence variation and/or increasing primer degeneracy also led to a slight increase of the mean slope and a pronounced decrease of the variation of the slope of the association (Figure 2C).

**2.** DNA vs. tissue pools: In comparison to DNA pools, our tissue based mock communities did not have any significant effect on the recovery of taxa or on the slope of the association of input DNA and recovered reads. However, our tissue pools showed a lower coefficient of determination per taxon, than DNA pools. Nevertheless, the amount of tissue per taxon was still well correlated to the recovered read count (Table 3 & Figure 3 & 4). Replicates of the same taxon (Collembola, Isopoda & Myriapoda) from DNA and tissue pools, recovered very similar slopes (Supplementary Figure 3).

**3.** PCR cycle reduction and DNA concentration: A fourfold increase of DNA amount during the PCR did not have a significant effect on any of the studied variables. The proportion of recovered taxa was not affected by first round PCR cycle numbers. Contrary to our expectation, PCR cycle numbers also did not change the mean, range or variation of the slope of the association of input DNA and recovered reads (Figure 5A). A reduction of PCR cycles did not positively affect the coefficient of determination. Instead, the two samples with higher cycle numbers show a significantly higher average *R2* (Figure 5B & Supplementary Figure 4).

*Correcting abundance estimates using linear and Bayesian approaches*

Each taxon shows a predictable fold change between the proportion of input DNA and recovered reads. But due to the 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 (Figure 6). By using 5 mock communities to derive taxon specific correction factors, a significant correlation is found (Figure 6A). This correlation is considerably increased, when 10 mock communities are used to derive corrections factors (Figure 6B). Using our Bayesian correction method, we also acquired a fairly accurate prediction of the amount of input DNA for all 44 arthropod taxa (Figure 7). The amount of input DNA can thus be fairly accurately predicted from mock communities for most studied taxa here. Both using linear regression and Bayesian correction methods.

**Discussion**

*Towards qualitative and quantitative metabarcoding of arthropod communities*

Metabarcoding studies can reliably predict the species richness of arthropod communities (Elbrecht et al. 2016), a finding which is well supported by our results. With the proper markers used, nearly all taxa in the mock communities could be recovered with high fidelity. The amount of recovered reads for our target taxa was correlated in a very predictable way with the amount of input DNA or tissue. Similar results have been found for microorganisms (Sohn et al. 2014; Giner et al. 2016). Irrespective of the amount of DNA of the target taxon or that of other taxa in the mock community, we could predict the taxon abundance with high reproducibility. These results suggest that a PCR based metabarcoding approaches should allow quantitative estimates of biomass in arthropod communities. Weighing or measuring specimens in community samples before sequencing could even allow to predict, e.g. biomass of taxa. However, it turned out to be quite difficult, to accurately measure the dry weight of arthropods. Especially larger specimens were often not fully dried, adding disproportionally more weight. This could also explain the lower coefficients of determination found for tissue, compared to DNA pools. Also, we could not always use members of the same species for different tissue pools, introducing a taxonomic bias. Moreover, we used different body parts for many taxa. It would be advisable to either use whole specimens or focus on only one body part for arthropod community analysis. Otherwise, taxon abundances and even species richness estimates could be highly skewed towards taxa with DNA rich tissues in the community sample.

The fold change between the proportion of input DNA and recovered reads is fairly variable, complicating quantitative analysis. Generally, this bias was not very pronounced and many taxa showed a near perfect correlation between input DNA and recovered reads. For an accurate quantitative analysis by metabarcoding, the expected taxa in the studied system and the taxon specific slopes need to be known. Ideally, mock communities of all major representative taxa in the community can be run to derive correction factors and predict the relative abundance of species (Thomas et al. 2015, Saitoh et al. 2016). Such correction factors from pooled mock communities might be easier to acquire than using qPCR on every single taxon. In our case, 5-10 mock communities are sufficient to quantify the relative abundance of more than 40 arthropod species with high accuracy. However, the identification of correction factors involves considerable effort. This method will thus not be feasible in unknown ecosystems or for simple exploratory work. But for large scale and long term studies in one ecosystem the effort might well pay off. Similar correction factors could possibly be derived for closely related taxa, e.g. reducing the necessary effort. Moreover, our approach seems suitable for comparative studies on abundance changes of a single target taxon, e.g. an invasive species across different sites. Considering that the PCR amplification of a single taxon will not change across samples, correction factors are not even needed.

*Abundance differences between taxa and markers - Priming bias or copy number variation?*

The accuracy of quantitative species recovery was tightly coupled to the use of conserved barcode markers and/or highly degenerate primers. While priming bias is affecting metabarcoding studies, it is alleviated by primer degeneracy. The most pronounced effect on priming efficiency is known from mutations in the last few 3’-prime bases of primers, while mutations further upstream have less effect (Stadhouders et al. 2010). But despite conserved priming sites for ribosomal RNA and degenerate primers for COI, we still found considerable variation in the slopes of the association of input DNA and recovered reads between different taxa. The observed differences in fold-change of recovery of input DNA between different taxa might thus also be affected by copy number variation (Rogers & Bendich 1987; Piotrowski et al. 2008). This assumption is supported by our experiments using PCR cycle reduction and increasing DNA template amounts. Both did not improve the accuracy of quantitative taxon recovery. Priming bias should be reduced by lower PCR cycles numbers and possibly template concentration, while bias based on copy number variation would be unaffected. However, other factors aside from simple sequence divergence are known to affect priming bias, even using degenerate primers (Suzuki & Giovannoni 1996; Polz & Cavanaugh 1998). Even though these factors should also be affected by PCR cycle number and DNA concentration, we currently can not rule out priming bias. Digital PCR or qPCR assays might help to identify copy number variation in the future.

PCR free analyses have been suggested as possible means for quantitative community analysis, as they exclude PCR priming bias (Crampton-Platt et al. 2016). PCR free methods were also shown to result in a better recovery of taxa from diverse communities (Shokralla et al. 2016). However, an amplicon sequencing based approach is much more cost efficient and involves a simpler workflow than PCR free methods. With current amplicon sequencing protocols, nearly 1,000 community samples can be analyzed in a single MiSeq run, reducing the sequencing cost to less than 2 $ per sample. Moreover, PCR free methods will be similarly sensitive to copy number variation of the target genes.

*Metabarcoding and mitochondrial COI – a perfect match?*

Most of the tested markers recover a high fraction of taxa from our mock communities. But interestingly, COI outperforms the other mitochondrial markers in its recovery of species and the prediction of species abundance. Even nuclear ribosomal markers with their highly conserved priming sites do not yield significantly better qualitative or quantitative results than COI. In contrast to rDNA however, COI is more variable and allows to distinguish even recently divergent species and intraspecific variation. Recent studies have suggested alternative primers to COI (Clarke et al. 2014; Deagle et al. 2014; Elbrecht et al. 2016). Indeed, different markers than COI might be advisable for certain taxa. E.g., we could not amplify some acari and hymenoptera with COI. As COI and other mitochondrial markers bring along problems like NUMTS (Benasasson et al. 2001) and their genealogy can be strongly affected by bacterial infections (Hurst and Jiggins 2005) or paternal gene flow (Chen et al. 2008), a suitable nuclear marker would be highly recommendable for future metabarcoding studies. While 18SrDNA and 28SrDNA performed very well in our study, they may be too conserved for many barcoding applications (Tang et al. 2012; Krehenwinkel et al. in prep.). A promising target are the internal transcribed spacers of the ribosomal cluster, which are already successfully applied in fungal taxonomy (Nilsson et al. 2009). With increasing genomic data available, a multitude of novel markers will be discovered in the coming years. Ideally, such markers should occur as a single copy and possess highly conserved priming sites. Ultra-conserved elements could provide a good solution in the future (Faircloth et al. 2012).

**Conclusion**

Our study shows that the abundance of taxa in community samples is tightly correlated to the recovered read count. Taxon specific priming bias can be mitigated by using highly degenerate primers for PCR amplification and correction methods based on mock communities. Apart from priming bias, copy number variation of the target locus could contribute to read abundance differences between taxa. Mitochondrial COI emerged as useful barcoding marker, uniting a high amplification efficiency across taxa with high sequence variation. A suitable nuclear backup marker is desirable.

**Acknowledgements**

We would like to thank Edward Greg Huang, Noriyuki Suzuki and Luis Cayetano for help during fieldwork. 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 thankfully provided access to the Genogrinder. HK was funded by a postdoctoral fellowship of the German Science Foundation (DFG). This work is funded by a grant of the National Science Foundation (NSF) to RG. 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

3. Analysis tables containing body length, weight and DNA content of each analyzed specimen

**Author contributions**

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

**References**

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403-410.

Angly, F. E., Dennis, P. G., Skarshewski, A., Vanwonterghem, I., Hugenholtz, P., & Tyson, G. W. (2014). CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome*, *2*(1), 1.

Barraclough, T. G., Hogan, J. E., & Vogler, A. P. (1999). 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*(1423), 1061-1067.

Bensasson, D., Zhang, D. X., Hartl, D. L., & Hewitt, G. M. (2001). Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in ecology & evolution*, *16*(6), 314-321.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic acids research*, *41*(D1), D36-D42.

Chen, S. F., Jones, G., & Rossiter, S. J. (2008). Sex‐biased gene flow and colonization in the Formosan lesser horseshoe bat: inference from nuclear and mitochondrial markers. *Journal of Zoology*, *274*(3), 207-215.

Clarke, L. J., Soubrier, J., Weyrich, L. S., & Cooper, A. (2014). Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. *Molecular ecology resources*, *14*(6), 1160-1170.

Colgan, D. J., McLauchlan, A., Wilson, G. D. F., Livingston, S. P., Edgecombe, G. D., Macaranas, J., ... & Gray, M. R. (1998). Histone H3 and U2 snRNA DNA sequences and arthropod molecular evolution. *Australian Journal of Zoology*, *46*(5), 419-437.

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

Deagle, B. E., Jarman, S. N., 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.

Drummond, A. J., Newcomb, R. D., Buckley, T. R., Xie, D., Dopheide, A., Potter, B. C., ... & Park, D. (2015). Evaluating a multigene environmental DNA approach for biodiversity assessment. *GigaScience*, *4*(1), 1.

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

Elbrecht, V., & Leese, F. (2015). Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass—sequence relationships with an innovative metabarcoding protocol. *PloS one*, *10*(7), e0130324.

Elbrecht, V., Taberlet, P., Dejean, T., Valentini, A., Usseglio-Polatera, P., Beisel, J. N., ... & Leese, F. (2016). Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects. *PeerJ*, *4*, e1966.

Evans, N. T., Olds, B. P., Renshaw, M. A., Turner, C. R., Li, Y., Jerde, C. L., ... & Lodge, D. M. (2016). Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Molecular ecology resources*, *16*(1), 29-41.

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic biology*, sys004.

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

Folmer, O. (1994). DNA primers for ampliation of mitochondrial cytochrome oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotechnol.*, *3*, 294-299.

Gibson, J., Shokralla, S., Porter, T. M., King, I., van Konynenburg, S., Janzen, D. H., ... & Hajibabaei, M. (2014). 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*(22), 8007-8012.

Giner, Caterina R., Irene Forn, Sarah Romac, Ramiro Logares, Colomban de Vargas, und Ramon Massana (2016). „Environmental Sequencing Provides Reasonable Estimates of the Relative Abundance of Specific Picoeukaryotes“. *Applied and Environmental Microbiology*, AEM.00560-16.

Gómez‐Rodríguez, C., Crampton‐Platt, A., Timmermans, M. J., Baselga, A., & Vogler, A. P. (2015). Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. *Methods in Ecology and Evolution*, *6*(8), 883-894.

Gordon, A. & Hannon, G.J. (2010) Fastx-toolkit. *Computer Program Distributed by the Author, Website http://hannonlab. cshl. edu/fastx\_toolkit/index. html [accessed 2014–2015]*.

Hebert, P. D., Ratnasingham, S., & de Waard, J. R. (2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London B: Biological Sciences*, *270*(Suppl 1), S96-S99.

Hebert, P. D., Penton, E. H., Burns, J. M., Janzen, D. H., & Hallwachs, W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(41), 14812-14817.

Heid, C. A., Stevens, J., Livak, K. J., & Williams, P. M. (1996). Real time quantitative PCR. *Genome research*, *6*(10), 986-994.

Hurst, G. D., & Jiggins, F. M. (2005). 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*(1572), 1525-1534.

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., ... & Larsen, T. H. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology letters*, *16*(10), 1245-1257.

Kartzinel, T. R., Chen, P. A., Coverdale, T. C., Erickson, D. L., Kress, W. J., Kuzmina, M. L., ... & Pringle, R. M. (2015). DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. *Proceedings of the National Academy of Sciences*, *112*(26), 8019-8024.

Kembel, S. W., Wu, M., Eisen, J. A., & Green, J. L. (2012). Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput Biol*, *8*(10), e1002743.

Kocher, T. D., Thomas, W. K., Meyer, A., Edwards, S. V., Pääbo, S., Villablanca, F. X., & Wilson, A. C. (1989). Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. *Proceedings of the National Academy of Sciences*, *86*(16), 6196-6200.

Krehenwinkel, H., Kennedy, S., Pekár, S., & Gillespie, R. G. (2016). 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*.

Kress, W. J., García-Robledo, C., Uriarte, M., & Erickson, D. L. (2015). DNA barcodes for ecology, evolution, and conservation. *Trends in ecology & evolution*, *30*(1), 25-35.

Lange, V., Böhme, I., Hofmann, J., Lang, K., Sauter, J., Schöne, B., Paul, P., Albrecht, V., Andreas, J.M., Baier, D.M., Nething, J., Ehninger, U., Schwarzelt, C., Pingel, J., Ehninger, G. & Schmidt, A.H. (2014) Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. *BMC Genomics*, **15**, 63.

Leray, M., Yang, J.Y., Meyer, C.P., Mills, S.C., Agudelo, N., Ranwez, V., Boehm, J.T. & Machida, R.J. (2013) A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, **10**, 1-14.

Leray, M., & Knowlton, N. (2015). DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences*, *112*(7), 2076-2081.

Machida, R. J., & Knowlton, N. (2012). PCR Primers for metazoan nuclear 18S and 28S ribosomal DNA sequences. *PLoS one*, *7*(9), e46180.

Morinière, J., de Araujo, B. C., Lam, A. W., Hausmann, A., Balke, M., Schmidt, S., ... & Haszprunar, G. (2016). Species Identification in Malaise Trap Samples by DNA Barcoding Based on NGS Technologies and a Scoring Matrix. *PloS one*, *11*(5), e0155497.

Nilsson, R. H., Ryberg, M., Abarenkov, K., Sjökvist, E., & Kristiansson, E. (2009). The ITS region as a target for characterization of fungal communities using emerging sequencing technologies. *FEMS Microbiology Letters*, *296*(1), 97-101.

Palumbi, S. R., Martin, A., Romano, S., McMillan, W. O., Stice, L. & Grabowski, G. (1991). The Simple Fool's Guide to PCR, Version 2.0. Privately published, Univ. Hawaii

Piotrowski, A., Bruder, C. E., Andersson, R., de Ståhl, T. D., Menzel, U., Sandgren, J., ... & Bartoszewski, R. (2008). Somatic mosaicism for copy number variation in differentiated human tissues. *Human mutation*, *29*(9), 1118-1124.

Polz, M. F., & Cavanaugh, C. M. (1998). Bias in template-to-product ratios in multitemplate PCR. *Applied and environmental Microbiology*, *64*(10), 3724-3730.

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

Ratnasingham, S., & Hebert, P. D. (2007). BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). *Molecular ecology notes*, *7*(3), 355-364.

Rogers, S. O., & Bendich, A. J. (1987). Ribosomal RNA genes in plants: variability in copy number and in the intergenic spacer. *Plant Molecular Biology*, *9*(5), 509-520.

Saitoh, S., Aoyama, H., Fujii, S., Sunagawa, H., Nagahama, H., Akutsu, M., ... & Nakamori, T. (2016). A quantitative protocol for DNA metabarcoding of springtails (Collembola) 1. *Genome*, *59*(9), 705-723.

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

Shokralla, S., Porter, T. M., Gibson, J. F., Dobosz, R., Janzen, D. H., Hallwachs, W., ... & Hajibabaei, M. (2015). Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific reports*, *5*.

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

Sohn, Michael B., Lingling An, Naruekamol Pookhao, und Qike Li (2014). Accurate genome relative abundance estimation for closely related species in a metagenomic sample. *BMC Bioinformatics* 15: 242.

Suzuki, M. T., & Giovannoni, S. J. (1996). Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and environmental microbiology*, *62*(2), 625-630.

Stadhouders, R., Pas, S. D., Anber, J., Voermans, J., Mes, T. H., & 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.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next‐generation biodiversity assessment using DNA metabarcoding. *Molecular ecology*, *21*(8), 2045-2050.

Tamura, K., Dudley, J., Nei, M., & Kumar, S. (2007). MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular biology and evolution*, *24*(8), 1596-1599.

Tang, C. Q., Leasi, F., Obertegger, U., Kieneke, A., Barraclough, T. G., & Fontaneto, D. (2012). 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*(40), 16208-16212.

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

de Valpine P., Paciorek C., Turek D., Anderson-Bergman C. & Temple Lang D. (2016). nimble: Flexible BUGS-Compatible System for Hierarchical Statistical. Modeling and Algorithm Development. R package version 0.6-2. https://CRAN.R-project.org/package=nimble.

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

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

Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X., Zhou, L., ... & Huang, Q. (2013). Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *Gigascience*, *2*(1), 1.

**Tables and Figures**

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

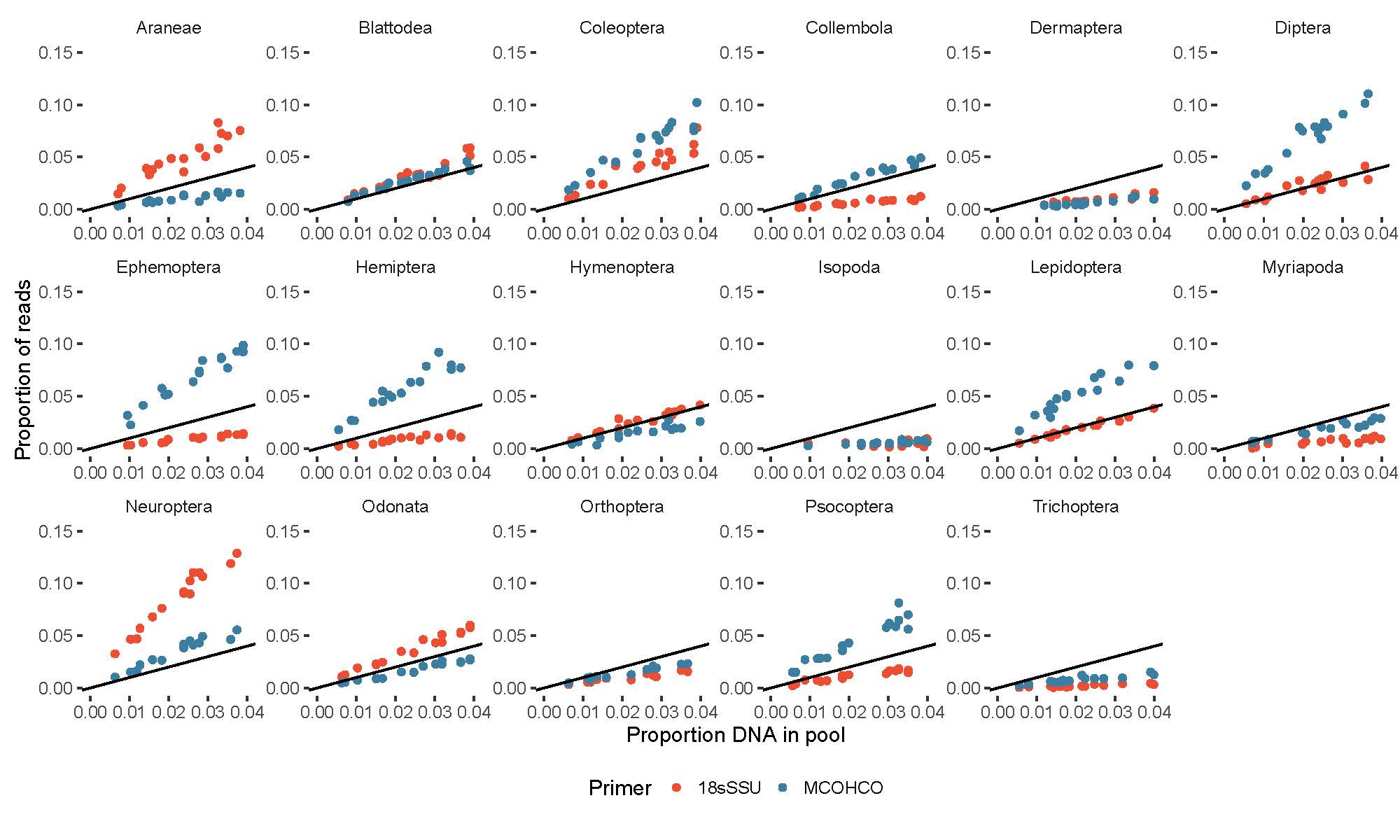
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Gene** | **Forward** | **Sequence 5'-3'** | **Reverse** | **Sequence 5'-3'** | |
| COI | ArF11 | GCNCCWGAYATRGCNTTYCCNCG | Fol-degen-rev2 | TANACYTCNGGRTGNCCRAARAAYCA | |
| COI | mlCOIintF3 | GGWACWGGWTGAACWGTWTAYCCYCC | Fol-degen-rev2 | TANACYTCNGGRTGNCCRAARAAYCA | |
| CytB | CB34 | GAGGAGCAACTGTAATTACTAA | CB44 | AAAAGAAARTATCATTCAGGTTGAAT | |
| 12SrDNA | 12sai5 | AAACTAGGATTAGATACCCTATTAT | 12sbi5 | AAGAGCGACGGGCGATGTGT | |
| 18SrDNA | SSU\_FO46 | GCTTGTCTCAAAGATTAAGCC | SSU\_R226 | GCCTGCTGCCTTCCTTGGA | |
| 18SrDNA | 18s\_2F7 | AACTTAAAGRAATTGACGGA | 18s\_4R7 | CKRAGGGCATYACWGACCTGTTAT | |
| 28SrDNA | 28s\_3F7 | TTTTGGTAAGCAGAACTGGYG | 28s\_4R7 | ABTYGCTACTRCCACYRAGATC | |
| Histone H3 | H3aF8 | ATGGCTCGTACCAAGCAGACVGC | H3aR8 | ATATCCTTRGGCATRATRGTGAC | |
| 1 Gibson et al. 2014; 2 Yu et al, 2012 3 Leray et al, 2013; 4 Barraclough et al. 1999; 5 Kocher et al. 1989; 6 Fonseca et al. 2010;  7 Machida & Knowlton 2012; 8 Colgan et al. 1998 | | | | |

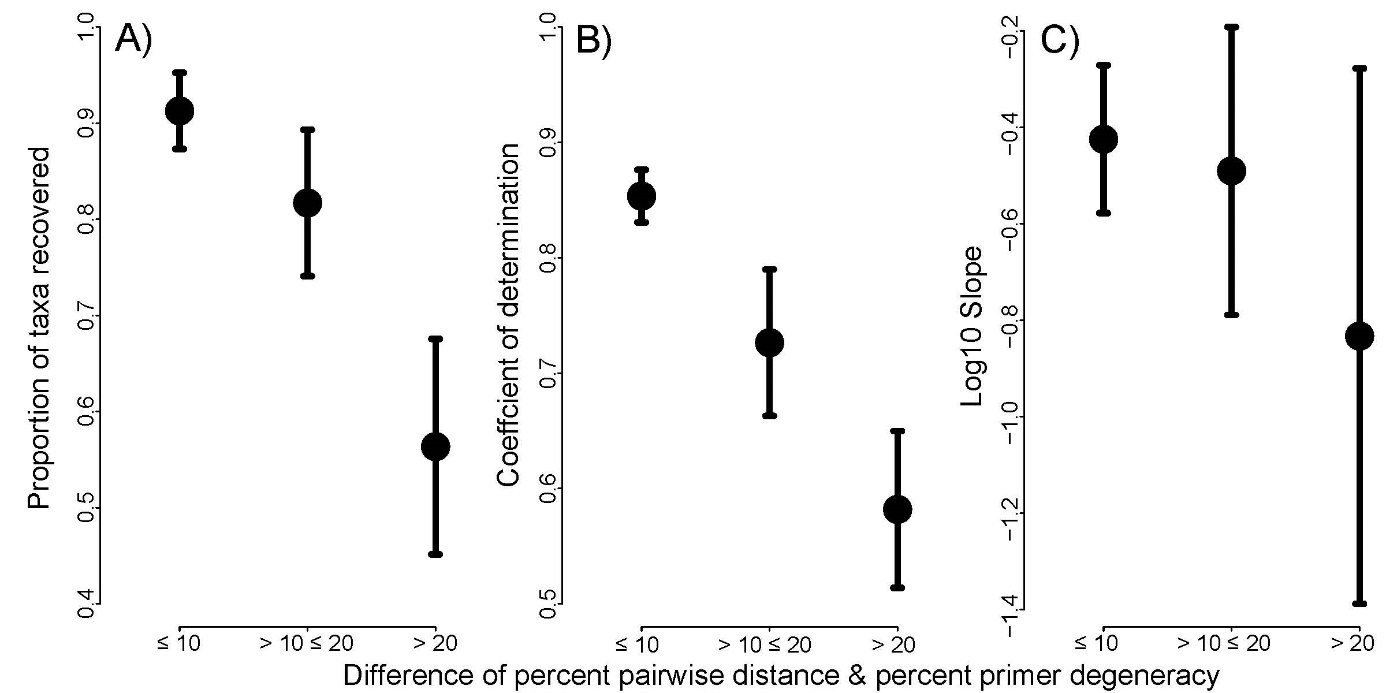
**Table 2** Summary of amplified genes, experimental conditions and experiment number. The tested variables include: marker type, amount of DNA per PCR, PCR cycle number, primer degeneracy, and average pairwise distance of marker in our mock communities. Experiment numbers correspond to: 1. Marker type, conservation and primer degeneracy, 2. DNA vs. tissue pools, 3. PCR cycle reduction and DNA concentration.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Gene** | **Primer** | **Experiment** | **Marker type** | **ng DNA/10 ul** | **PCR Cycle #** | **Tissue/DNA** | **Degeneracy** | **Pairw. dist.** |
| COI | ArF1/Fol-degen-rev | 3 | mito. | 60 | 4 | DNA | 21.4 | 26 |
| COI | ArF1/Fol-degen-rev | 3 | mito. | 60 | 8 | DNA | 21.4 | 26 |
| COI | ArF1/Fol-degen-rev | 3 | mito. | 60 | 16 | DNA | 21.4 | 26 |
| COI | ArF1/Fol-degen-rev | 3 | mito. | 60 | 32 | DNA | 21.4 | 26 |
| COI | ArF1/Fol-degen-rev | 1 & 3 | mito. | 15 | 32 | DNA | 21.4 | 26 |
| COI | mlCOIintF/Fol-degen-rev | 1 &2 | mito. | 15 | 32 | DNA | 17.3 | 25.4 |
| COI | mlCOIintF/Fol-degen-rev | 2 | mito. | 15 | 32 | Tissue | 17.3 | 25.4 |
| CytB | CB3/CB4 | 1 | mito. | 15 | 32 | DNA | 1 | 31.4 |
| 12S | 12sai/12sbi | 1 | mito. | 15 | 32 | DNA | 0 | 28.4 |
| 18S | SSU\_FO4/SSU\_R22 | 1 | nucl. | 15 | 32 | DNA | 0 | 12 |
| 18S | 18s\_2F/18s\_4R | 1 | nucl. | 15 | 32 | DNA | 5.7 | 9 |
| 28S | 28s\_3F/28s\_4R | 1 | nucl. | 15 | 32 | DNA | 7.6 | 16.3 |
| H3 | H3aF/H3aR | 1 | nucl. | 15 | 32 | DNA | 4.9 | 21.3 |

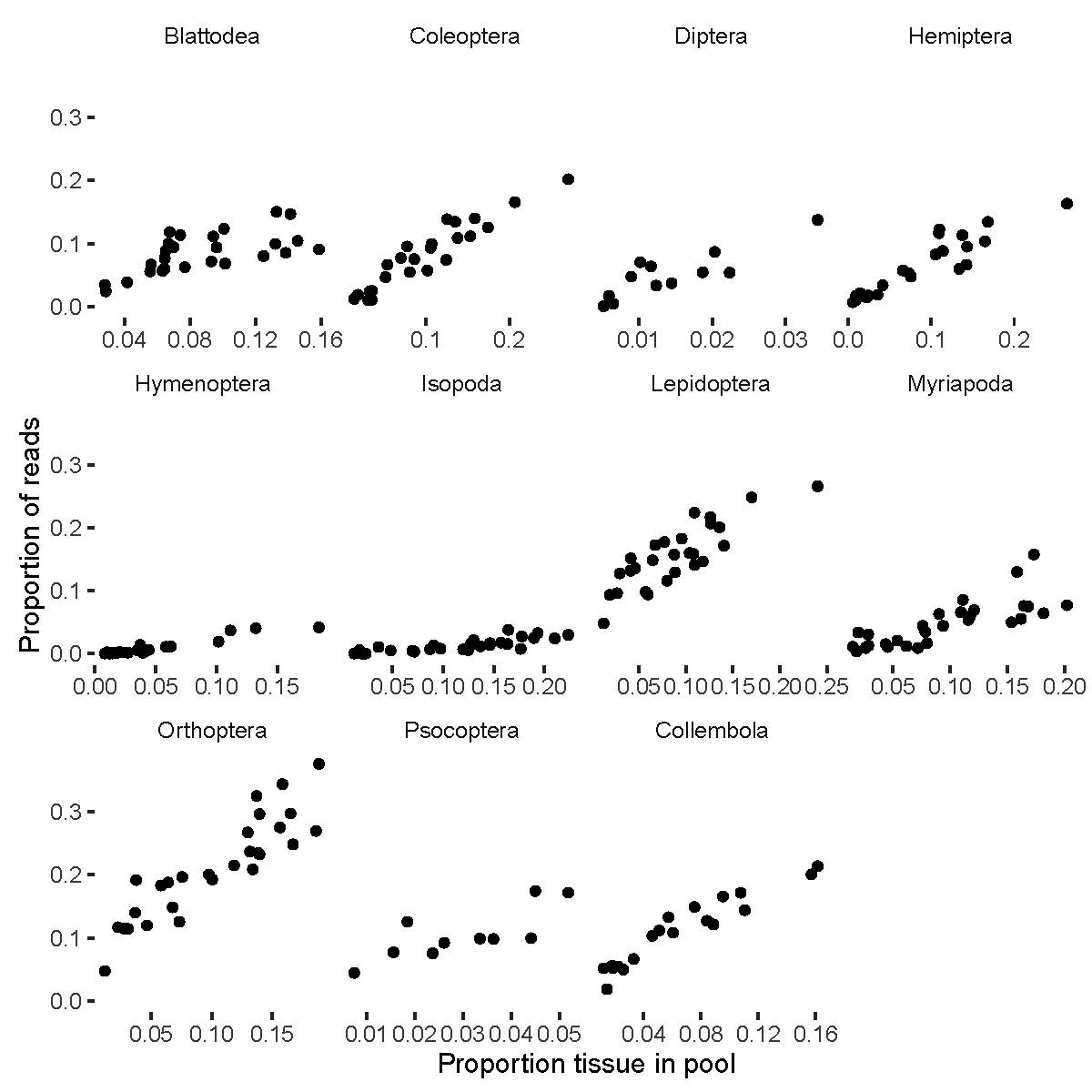
**Table 3** Summary of the result for the three experiments. The last three columns present the average proportion of recovered taxa, as well as the average coefficient of determination and slope of the association of input taxon and recovered read abundances including the respective standard deviation. Experiment numbers correspond to: 1. Marker type, conservation and primer degeneracy, 2. DNA vs. tissue pools, 3. PCR cycle reduction and DNA concentration.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Gene** | **Primer** | **Experiment** | ***α*** | ***R2*** | ***m*** |
| COI | ArF1/Fol-degen-rev | 3 | 0.92 ± 0.22 | 0.74 ± 0.17 | 1.04 ± 1.18 |
| COI | ArF1/Fol-degen-rev | 3 | 0.94 ± 0.21 | 0.67 ± 0.11 | 0.86 ± 1.05 |
| COI | ArF1/Fol-degen-rev | 3 | 0.94 ± 0.21 | 0.87 ± 0.08 | 1.03 ± 1.07 |
| COI | ArF1/Fol-degen-rev | 3 | 0.93 ± 0.22 | 0.86 ± 0.08 | 0.94 ± 0.8 |
| COI | ArF1/Fol-degen-rev | 1 & 3 | 0.94 ± 0.21 | 0.88 ± 0.08 | 0.94 ± 0.76 |
| COI | mlCOIintF/Fol-degen-rev | 1 &2 | 0.92 ± 0.26 | 0.88 ± 0.11 | 0.99 ± 0.84 |
| COI | mlCOIintF/Fol-degen-rev | 2 | 0.97 ± 0.1 | 0.72 ± 0.21 | 0.86 ± 0.68 |
| CytB | CB3/CB4 | 1 | 0.71 ± 0.39 | 0.55 ± 0.23 | 1.02 ± 1.72 |
| 12S | 12sai/12sbi | 1 | 0.41 ± 0.47 | 0.64 ± 0.19 | 2.47 ± 2.82 |
| 18S | SSU\_FO4/SSU\_R22 | 1 | 0.94 ± 0.22 | 0.86 ± 0.16 | 0.99 ± 1.1 |
| 18S | 18s\_2F/18s\_4R | 1 | 0.94 ± 0.22 | 0.84 ± 0.17 | 0.97 ± 0.97 |
| 28S | 28s\_3F/28s\_4R | 1 | 0.85 ± 0.34 | 0.8 ± 0.18 | 1.06 ± 1.12 |
| H3 | H3aF/H3aR | 1 | 0.69 ± 0.41 | 0.54 ± 0.27 | 1.21 ± 1.02 |

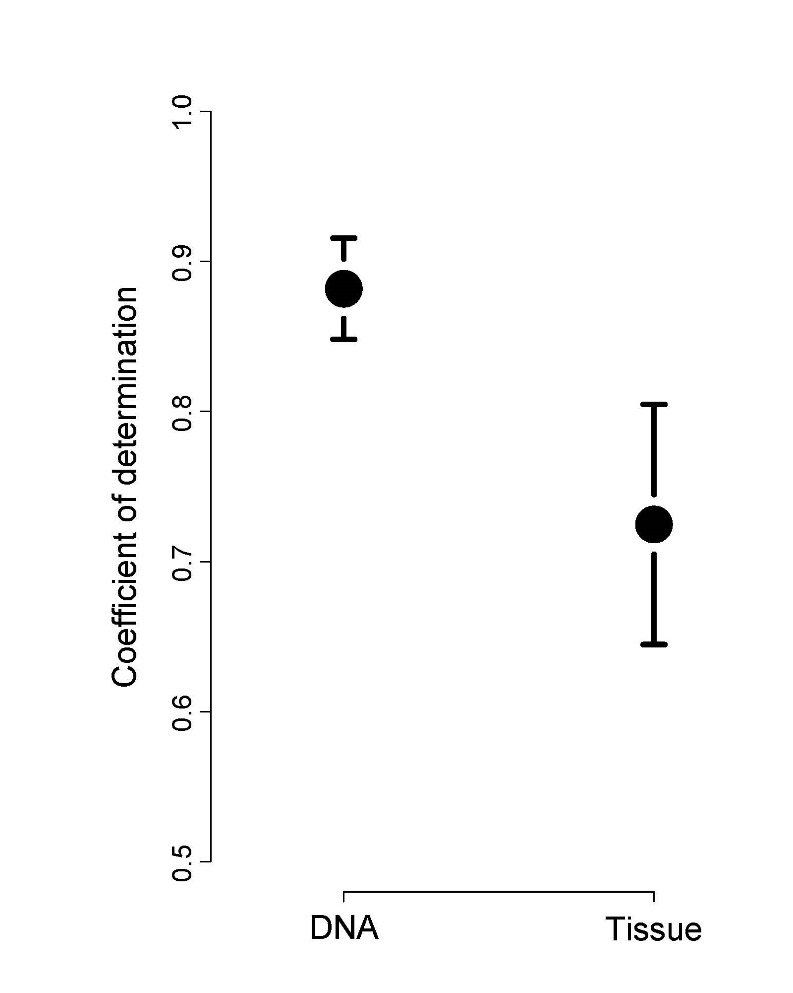
***Figure 1*** *Exemplary associations of proportion of DNA and proportion of recovered reads per taxon for various arthropod orders from DNA mock communities using mitochondrial COI and nuclear 18SrDNA as marker. The Y-axis is scaled the same way in each plot to visualize differences in the slope of the association. The black lines represent 1:1 lines.*



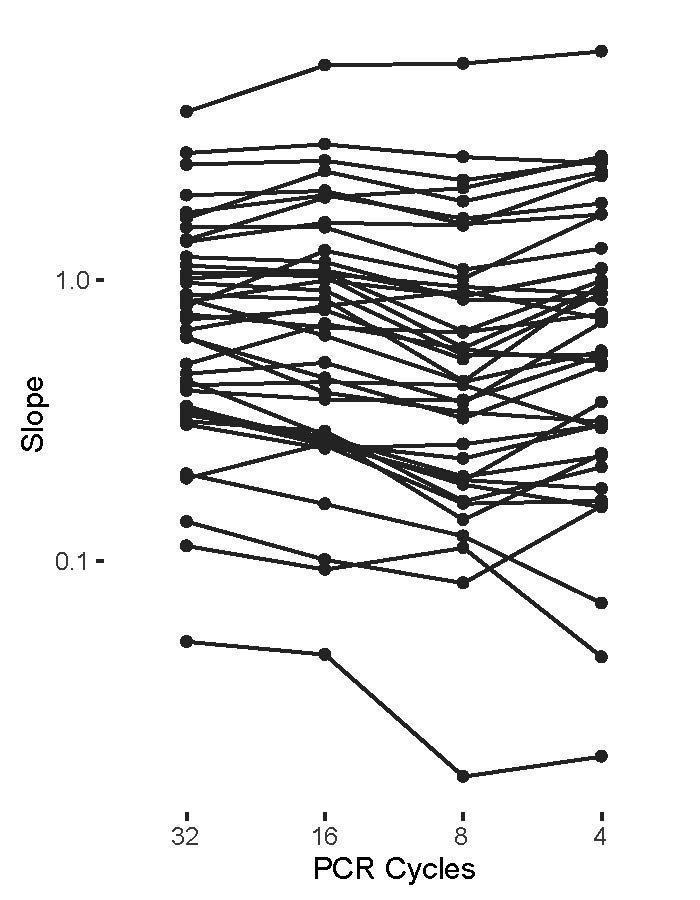
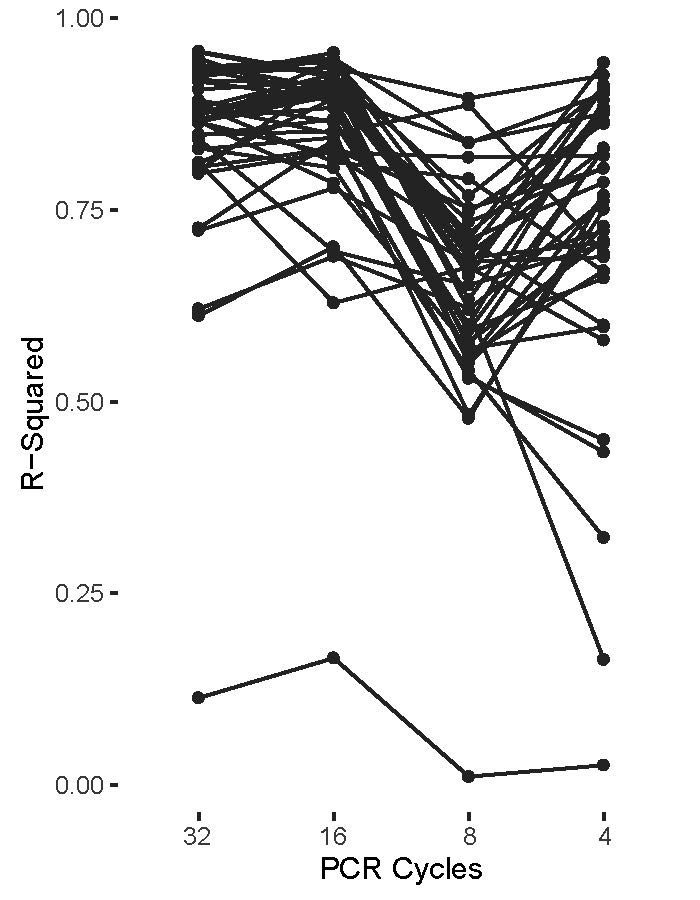
**Figure 2** **A)** Average Proportion of recovered taxa from mock communities of 44 different arthropod taxa for three categories of primer variation (calculated as difference between average pairwise percent distance of marker in community and percent of degenerate sites of primer) from 0-10, 10-20 and > 20. These three categories correspond to increasing variability of the targeted gene sequence and/or decreasing degeneracy of the primer sequences. **B)** Coefficients of determination and **C)** slopes of the associations between read count and taxon abundance for the same mock communities and categories. Bars depict the 95 % confidence intervals.



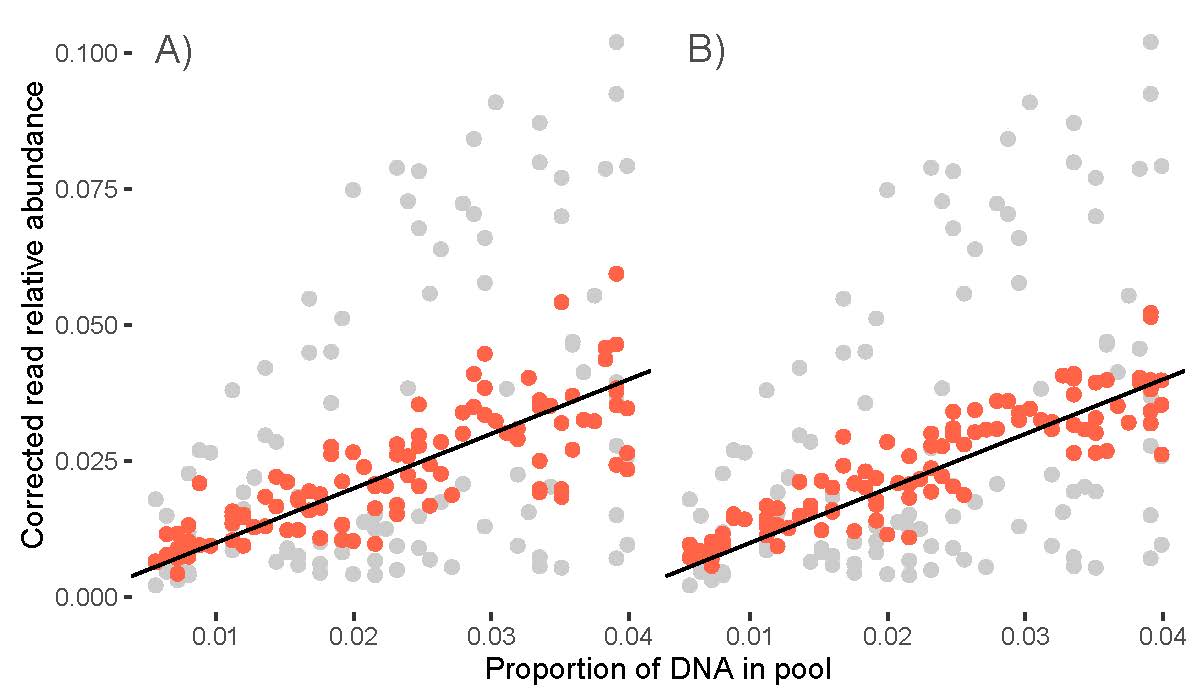
**Figure 3** Exemplary associations of proportion of tissue and proportion of recovered reads per taxon for tissue mock communities using mitochondrial COI for various arthropod orders. The Y-axis is scaled the same way to visualize differences in the slope of the association.



***Figure 4*** *Average coefficient of determination (R2) of the association of the abundance of taxa and the recovered read counts from mock communities based on DNA pools and tissue pools. Bars depict the 95 % confidence interval.*



***Figure 5******A)*** *Slope (m) and* ***B)*** *Coefficient of determination (R2) of the association of input DNA and recovered read count for 44 arthropod taxa amplified for mitochondrial COI at four different first round PCR cycle numbers. Replicates of the same taxon at different PCR cycle numbers are connected by black lines.*



**Figure 6** Uncorrected association of actual abundance and recovered read proportion for 44 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.

Macintosh HD:Users:ajr:Dropbox:hawaiiDimensions:metabarcoding:poc:ms:fig_correctedDNA.pdf

**Figure 7** Association of the amount of actual input DNA with the predicted estimates of input DNA for 44 arthropod taxa. Gray open circles represent a naïve estimate using the read count per taxon. Closed black circles are calculated using our Bayesian correction method. The red line corresponds to the 1:1 line.