**A systematic assessment of biases in qualitative and quantitative arthropod metabarcoding studies**

Henrik Krehenwinkel1 2 \*, Andrew J. Rominger1, Jun Jing Lim1, Madeline Wolf1, Marisa Fong1, Jingfei Zhang1, Brian Simison2, Rosemary 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

**Abstract**

1. Next generation sequencing based metabarcoding promises the rapid and cost efficient analysis of arthropod biodiversity. However, current amplicon based metabarcoding approaches are believed to suffer from pronounced PCR biases, leading to flawed diversity estimates. Novel barcode markers or PCR free approaches have been suggested to mitigate this problem. Aside from PCR bias, many other confounding factors remain largely untested. Especially, taxon specific DNA degradation and DNA content pose a potential issue in community analyses. Here, we systematically test the effect of these biases on qualitative and quantitative recovery of arthropod communities.

2. We test eight different nuclear and mitochondrial PCR markers of varying sequence conservation for their potential to recover species richness and abundance in diverse arthropod communities. We also test the possibility of mitigating priming bias by reducing PCR cycle number and increasing the ratio of template DNA to primer. We develop a Bayesian correction method to reconstruct taxon specific abundance. Last, we estimate the effect of taxon specific DNA content and DNA degradation biases on community analyses.

3. Amplicon sequencing approaches can accurately predict species richness and abundance. PCR bias can be avoided by employing our Bayesian correction method and increasing the ratio of template DNA to primer. A reduction of PCR cycles does not improve the accuracy of abundance estimates, suggesting copy number variation rather than priming bias drives read abundance differences. For most arthropod taxa, degenerate COI primers provide the best taxonomic resolution and species recovery, reappraising COI as the most useful barcoding marker. We find pronounced differences of DNA content between body parts and increasing taxonomic DNA degradation bias with less optimal sample storage.

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 abundance. Additional biasing factors have to be considered in metabarcoding studies, including taxon specific DNA degradation and tissue specific DNA content.

Especially important for work using passive trapping methods exposing specimens to unfavorable conditions over long term

**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 food web structure, cryptic species, identification of juveniles and hidden diversity, e.g. internal parasitoids and species richness 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 actual species abundances from mixed samples (Elbrecht & Leese, 2015, Collembola paper).

The commonly used PCR based approaches suffer from pronounced bias. Sequence divergence in priming sites or copy number variation of the target genes can lead to flawed abundance estimates (Clarke et al. 2014; Deagle et al. 2014). 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, and might be more suitable for quantitative community analysis (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 to identify sequences (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). PCR free approaches have also been suggested. The direct sequencing of genomic DNA and analysis of the recovered barcode sequences is assumed to provide more accurate predictions of abundance in communities (Crampton-Platt et al. 2016; Gomez-Rodriguez et al. 2015; Zhou et al. 2013). 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 will also be 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 problems with bias and allow for a more accurate correlation of input DNA and recovered reads. Copy number variation instead should be unaffected by cycle number and pose a constant problem to the study of abundance.

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 taxa in the extraction, the proportion of input DNA should be tightly correlated to the proportion of recovered reads for that taxon. PCR priming bias or copy number variation should merely affect the slope of the correlation. If this slope for taxa in a community is known, it might be possible to accurately predict their relative abundance by using correction factors (Thomas et al. 2015; Angly et al. 2014; Collembola paper). 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.

Aside from PCR bias, other factors could affect the qualitative and quantitative efficiency of metabarcoding studies. A largely neglected problem in arthropod metabarcoding concerns taxonomic biases in DNA recovery. This can affect an analysis in several ways. First, DNA could degrade at different paces in different taxa leading to a taxonomic degradation bias. Arthropod community samples are often collected by passive trapping methods, exposing the samples to varying times of suboptimal storage, e.g. in a malaise or pitfall trap in a tropical forest. Such storage conditions could additionally amplify degradation bias. Different arthropod taxa are also distinguished by different bodyplans, which translate into different soft to hard tissue ratios and thus most likely different DNA contents. And even within an arthropod specimen, separate body parts are expected to contain varying amounts of cells and consequently DNA. The usage of different body parts or different taxa in a community analysis could thus already introduce a considerable bias. Depending on the strength of all these taxonomic biases, qualitative and quantitative metabarcoding results could be severely distorted, even in the absence of PCR priming bias.

Here, we test various biasing factors in metabarcoding analyses of arthropod communities. Using a simple dual indexing protocol on the Illumina MiSeq system, we sequence tissue and DNA mock communities of a taxonomically diverse set of Hawaiian and Californian arthropods. We aim to optimize qualitative assessments of arthropod biodiversity and explicitly test for the possibility of species abundance estimates. We try to optimize our analyses by: **1.** choosing appropriate barcode markers out of a selection of four nuclear and four mitochondrial fragments. These primers amplify sequences of varying degrees of conservation, and accordingly should be affected by small to moderate priming bias. **2.** By reducing the PCR cycle number during library preparation and increasing the ration of template DNA to primer and 3. by identifying the taxon specific PCR bias and correcting for it. We also quantify the effect of degradation bias at two different sample storage conditions by size selecting and sequencing degraded and non-degraded DNA fractions out of community samples. And last, we estimate taxon and tissue specific biases and association of body size, weight and associated DNA content for different taxa.

**Methods**

*Sample collection, mock community preparation and PCR amplification*

Arthropod samples were collected in native rainforests on the Hawaiian Islands Maui and Hawaii Island 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 (Arachnida, Crustacea, Insecta & Myriapoda, see supplementary Table 1 for details on taxonomy). 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 then 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 each of the 44 specimens in the mock communities for each of the 8 primer combinations. The primers amplified sequences of varying degrees of conservation, from the highly conserved nuclear 18srDNA to the variable mitochondrial COI (See supplementary material). This increasing sequence variation should also be reflected in increasing PCR priming bias. All primer pairs amplified sequences shorter than 500 bp to achieve an overlap of 2 x 300 bp Illumina MiSeq reads. The COI primers were selected from a list of xxx possible primer combinations. We tested these primer combinations extensively, to find those primers, which most successfully amplified the largest fraction of a diverse set of arthropods (See supplementary xxx).

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.

Additionally, we ran a series of PCRs with varying cycle numbers and increasing DNA template concentration. All 23 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 of were run, followed by second round indexing PCRs of 26, 22, 14 and 6 cycles, to make the total number of cycles 32 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 following 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. 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.

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 | | | | |

*Tissue mock communities and DNA degradation experiment*

To test the applicability of our method under real conditions and to estimate taxonomic bias in DNA degradation, we generated mock communities from tissue pools of different Hawaiian taxa. After collection, the according samples were stored at two different conditions for four weeks. One was constantly frozen, while the other was kept at room temperature for the same time. Specimens were identified to morphotypes as described above and defined amounts of tissue of approximately 20 taxa combined into 15 mock communities for each of the two temperature conditions. 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 different 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 microscale (XX). 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. Tissue pools were 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. Each sample was brought to a concentration of 35 ng/µl and a separation of high and low molecular weight DNA performed using Ampure Beads XP (Beckman Coulter, Brea, CSA, USA**)**, as described in Krehenwinkel et al. (2016). This step resulted in 60 final DNA samples, (30 freezer stored and 30 room temperature stored). Each community sample was thus split into a high and a low molecular weight fraction, with the latter containing degraded DNA. The mitochondrial COI and nuclear 18srDNA were amplified from each sample using the primer pairs mlCOIintF/Fol-degen-rev and SSU\_FO4/SSU\_R22. PCR, 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. Each of the previously generated alignments of 44 reference specimens per marker was used to create seven 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 and 18s sequences from the tissue pools was performed using USEARCH (Edgar et al. 2011) with a minimum similarity of 95-%. The taxonomy of the resulting OTU centroid sequences was assigned as described above. Taxon recovery and read abundance to input tissue proportion were analyzed like described above for the DNA pools. We refrained from analysis of abundance and read count for specific taxa for the 18srDNA dataset. As this marker did not provide a sufficient reference database and taxonomic resolution to distinguish between some of the different taxa in the tissue pools.

*Qualitative and quantitative community analyses*

We quantified the proportion 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 according markers for qualitative metabarcoding purposes. 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. We calculated a Bayesian R2 value to summarize the predictability of the amount of input DNA per taxon vs. the number of reads recover. 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 NIMBLE package (cite) in the R language (cite).

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 26 cycle PCR experiment to predict the input of the 6 cycle experiment, both using ARFHCO.

*DNA degradation and DNA content bias*

We correlated the read abundance estimates for each recovered OTU cluster for the high and the low molecular weight fraction of our tissue mock communities, and calculated Bray Curtis distances for each of the sample pairs to detect taxonomic bias in DNA degradation. The correlations of read abundances and Bray Curtis distances were compared to a previously generated dataset of 15 mock communities of Hawaiian arthropods (Krehenwinkel unpublished data), in which we estimate the effect of simple PCR replicates and changes in annealing temperature on read abundances. For this purpose, exact PCR replicates of the same 15 mock communities were run 1. Under exactly the same conditions and 2. Using an elevated annealing temperature of 51°C instead of 46°C. Assuming that taxon bias in DNA degradation is marginal, the association between exact PCR replicates of high and low molecular DNA weight of the same sample should be similar. Moreover, we identified the association of bodysize, bodyweight and DNA content for 9 exemplary arthropod taxa and estimated the bias of DNA content between different body parts from the same specimen from our data (see Supplementary Material for more details).

**Results**

Most recovered samples were of high quality and good sequence coverage. After quality filtering, we recovered 8889 reads per DNA mock community and 15077 reads per tissue mock community on average. 4 of the 60 tissue community samples and 6 of the 220 DNA pools had to be excluded due to too low coverage.

We found considerable differences in taxa recovery from the mock communities for different markers. In our DNA pools, nuclear ribosomal markers and mitochondrial COI show a significantly better recovery than H3, CytochromeB or 12s (for each the mean). Neither a reduction of PCR cycles, nor an increase of the amount of template DNA did show significant effect on taxon recovery. Also, the degree of conservation of the targeted sequence did not show an association with recovery. The highly conserved ribosomal DNA, as well as the variable COI showed comparable recovery. We also did not find a significant difference between taxon recovery for the tissue pools. Irrespective of the DNA integrity (high vs. low molecular weight) we consistently find a high taxon recovery ().

Our Dirichlet-multinomial model accurately predicted the number of recovered reads per taxon based on DNA input (high Bayesian R2 in Fig. 2).

Throughout arthropoda holds for all taxa

Very good association

Different species and different markers markers show different trends e.g. different slopes ofassocitions

Highly conserved 18s not less steep and varying slope than coi

Tissue pools show similar results, with association not that clear cut 🡪 more noise

But interstinlgy replicates taxa (collembola, isopoda, myriapoda)🡪 tissue and dna pools show very similar slopes

Nuclear ribosomal markers and mitochondrial COI result in the best association of DNA per taxon and read count, substantially better than 12s, CytB and H3. The COI primer combination MCOHCO yields a significantly better correlation than ARF1HCO. No association of sequence conservation and coefficient of determination was found. A reduction of PCR cycles in the first round PCR did not positively affect the posterior coefficient of determination. A significant effect was found for an increase of the amount of PCR template. A four-fold increase of starting template for ARf1HCo resulted in a xxx fold higher coefficient of determination. Our tissue pools generally showed a lower coefficient of determination per taxon, than DNA pools. Nevertheless, the amount of tissue per taxon is usually well correlated to the recovered read count. The R2 95% credible intervals overlapped between high or low molecular weight DNA samples.

While we usually found a high coefficient of determination between read count and tissue or DNA amount, the taxon-specific association between DNA input and read count was fairly variable (see Fig. 3 for examples). However, because these differences are predictable, the parameter estimates from one experiment can be used to estimate the input DNA from another. We find a tight association between the input DNA and predicted output DNA using this method, especially compared to the naïve estimate of input DNA based on the raw reads (Fig. 4). Moreover, the recovered slopes for the same taxon based on DNA pools or tissue pools is very similar (XXX). A comparison of three taxonomic replicates of DNA and tissue pools shows a very similar slope for the association of taxon and read abundance (see Supplementary Figure xxx).

**Taxon bias**

An exact PCR replicate of the same sample does result in a very low Bray Curtis distance and very narrow association of read abundances for different taxa. Simple PCR thus introduces a negligible bias into metabarcoding analyses. In contrast, we found a significantly higher Bray Curtis distance between high and low molecular weight fractions of the same mock community amplified for mitochondrial COI. The distance is higher for samples stored at room temperature than for those kept in the freezer. A similar trend is found for the 18srDNA, with room temperature samples showing a higher distance than freezer ones. However, the observed distances for 18srDNA are significantly lower than those for COI. An association of read counts for each taxon from the high and low molecular weight fraction confirms these analysis, with freezer samples showing a narrower distribution then those stored at room temperature. Even though we found significantly lower community distances for PCR replicates than for high and low molecular weight fractions of DNA extractions, a much higher Bray Curtis distance was found for PCR replicates with an annealing temperature increase of 5°C (). Considering a more than 10-fold decrease of community similarity between 46 and 51 C annealing temperature, the bias between high and low molecular weight samples at freezer and room temperature storage seems insignificant.

**DNA content and body parts**

All analyzed taxa show a trend of

Legs contribute realtivel signfacntly less dna than psorsoma from spiders

Pronounced bias for all experimental condntions

Also shows taxon bias

**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 for qualitative as well as quantitative estimates of arthropod communities.

While DNA pools constitute optimized communities, our tissue pools were affected by additional biases, explaining the lower coefficients of determination. 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. Also, we could not always use members of the same species for different pools, introducing a taxonomic bias. Moreover, we used different body parts for many taxa (See Supplementary Material). It would be highly 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. Probably due to bias in priming efficiency or variable copy numbers, some taxa were overabundant, while others were underrepresented in the mock community (Angly et al. 2014; Deagle et al. 2014). 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 need to be known. In ideal case, mock communities of all major representative taxa in the community can be run to calibrate the Bayesian correction method proposed here to predict the abundances of species of community samples. Such correction methods would allow for an accurate identification of relative abundances. In our case, 5-10 mock communities are sufficient to quantify the relative abundance of ~40 Californian arthropod species with high accuracy. The calibration of such correction models involves considerable effort. Our 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. It might be possible that calibrated correction models could be derived for closely related taxa, e.g. reducing the necessary effort. Moreover, our approach seems suitable for comparative studies on abundance changes of target taxa, e.g. an invasive species across different sites.

An overabundance of template DNA in relation to primers during the PCR significantly increases the correlation of input DNA and recovered read proportions. Such an overabundance might simply increase the chance for even rare templates to be primed and amplified. In combination with the use of correction factors, this improves accurate estimates of community composition.

Contrary to our expectations, a reduction of the first round PCR cycle number did not improve the accuracy of quantitative species recovery. These results suggest that divergence in priming sites might not necessarily lead to strong bias. The most pronounced effect on priming efficiency is known from mutation in the last few 3’-prime bases of primers, while mutations further upstream have less effect (Stadhouders et al. 2010). Saturated with degenerate bases, many of the primers used here, could be well protected from such bias. The observed differences in fold-change of recovery of input DNA between different taxa might thus rather be affected by copy number variation, than by priming bias (Rogers & Bendich 1987; Piotrowski et al. 2008).

PCR free analyses have been suggested as possible means for quantitative community analysis, as they exclude PCR priming bias (Crampton-Platt et al. 2016). However, an amplicon sequencing based approach is much more cost efficient and involves a simpler workflow than current 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 genome sequencing methods (Gomez-Rodriguez et al. 2015) or enrichment protocols (Zhou et al. 2013) 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). But the good performance of the COI primers used here (Leray et al. 2013; Gibson et al. 2014), is probably associated with their saturation with degenerate sites, which allow to reliably amplify most arthropods. A large number of markers did not increase the predictive power of our metabarcoding study. Instead degenerate COI primers alone will allow for a detailed estimation of species richness and abundance for most taxa. Different markers might be advisable for certain taxa. E.g., we could not amplify some 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 as a backup for COI in future metabarcoding studies. While 18s and 28s rDNA performed very well in our study, they may be too conserved for many barcoding applications (Tang et al. 2012). A promising target might be the internal transcribed spacers of the ribosomal cluster, which are already successfully applied in fungal taxonomy (Nilsson et al. 2009). With increasing genome data available, a multitude of novel markers might be discovered in the coming years.

*The effect of taxonomic bias in DNA degradation*

Arthropod metabarcoding is often based on specimens collected with passive trapping methods, e.g. pitfall or malaise traps. Such traps expose the samples to suboptimal storage conditions for extended time periods. Moreover, field conditions usually don’t allow to store samples at low temperatures. This can lead significant DNA degradation. A taxonomic bias in DNA integrity would skew metabarcoding results. We show that there is in fact taxonomic bias in DNA degradation and that it is significantly reduced by an immediate cold storage of samples. However, the observed bias is low, even after suboptimal storage and not much different from simple PCR replicates. Only few taxa show a more pronounced degradation bias. For Illumina sequencing, very short sequence stretches of multi copy loci are amplified, e.g. mitochondrial or ribosomal DNA. The degradation bias will be additionally mitigated by this. Passively collected samples should thus still be well suitable for metabarcoding applications, even after certain times of suboptimal storage. Interestingly, nuclear ribosomal markers showed a less pronounced bias than mitochondrial ones. Possibly, DNA in the nucleus is better protected from degradation than cytoplasmic mitochondria. The histone packing of nuclear chromosomes, might additionally contribute to their stability compared to the mitochondrial DNA ().

**Conclusion**

Our study shows that the abundance of taxa in community samples is tightly correlated to the recovered read count. Taxon specific biases in read abundance can be mitigated by using a Bayesian correction method from mock communities. Using highly degenerate primers for PCR amplification, might mitigate priming bias. Our results suggest that copy number variation of the targeted marker might be a prime problem for abundance estimates. Mitochondrial COI emerged as most useful barcoding marker, even though a suitable nuclear backup would be desirable. Passive sampling and field conditions introduce an additional taxonomic degradation bias into metabarcode analysis. Same holds true for the usage of different body parts for the analysis.

**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 giving us access to a microscale. Anna Sellas helped during 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, read counts and abundances, for each mock community

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

**Author contributions**

Devised the study Analyzed the data Wrote the manuscript Collected the data

**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.

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.

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.

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*.

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.

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*.

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.

Figures

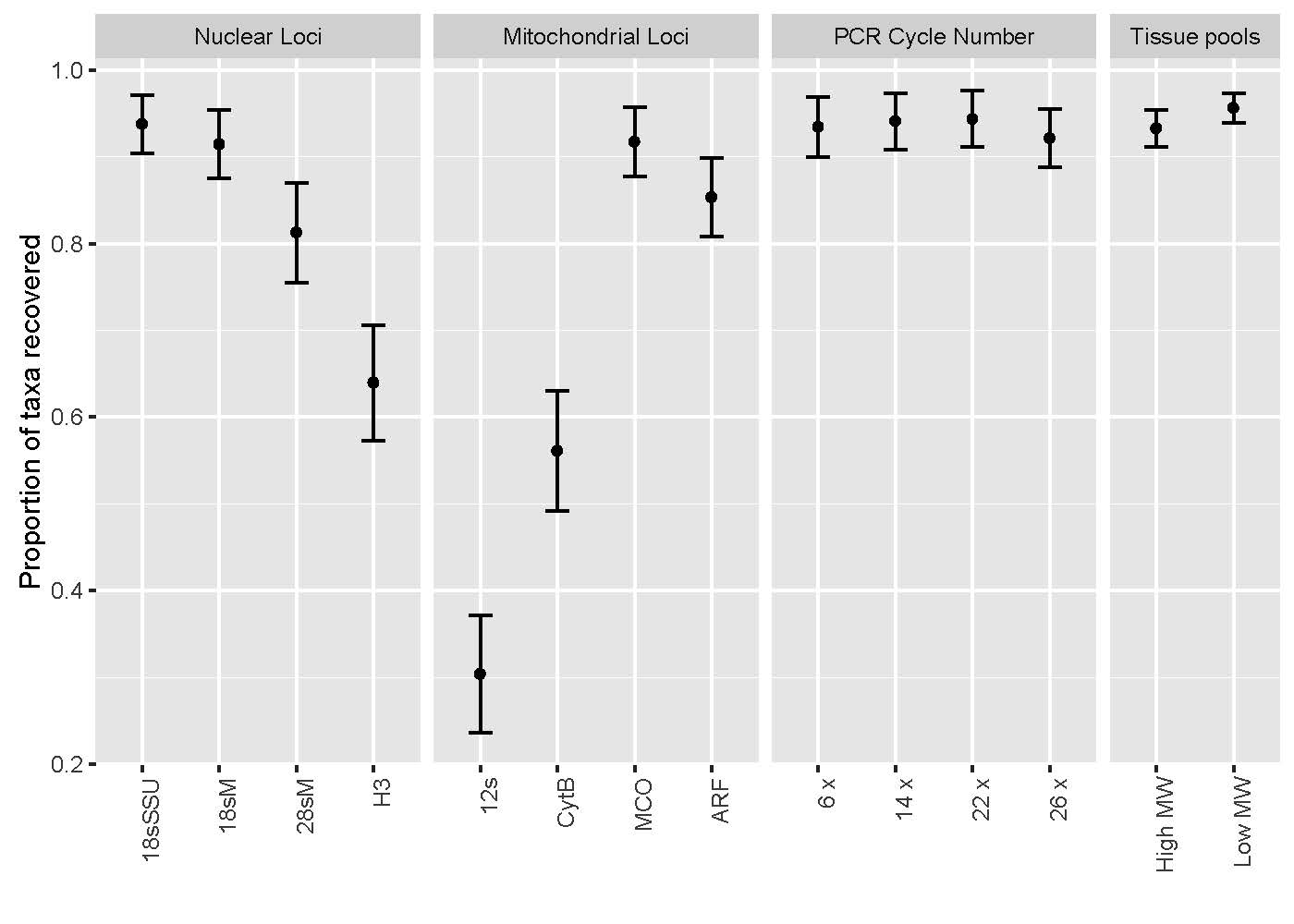


Figure 1. (A) Average proportion of each taxon recovered from mock communities of (A) nuclear markers (B) for mitochondrial markers (C) for marker ARF with different PCR cycles (D) for tissue pools with marker MCO

Macintosh HD:Users:ajr:Dropbox:hawaiiDimensions:metabarcoding:poc:ms:fig_r2.pdf

Figure 2. Bayesian R2 values across markers (A: nuclear; B: mitochondrial), PCR cycles (C) and tissue experiments (D). Error bars (often obscured by data point) represent 95% credible intervals.

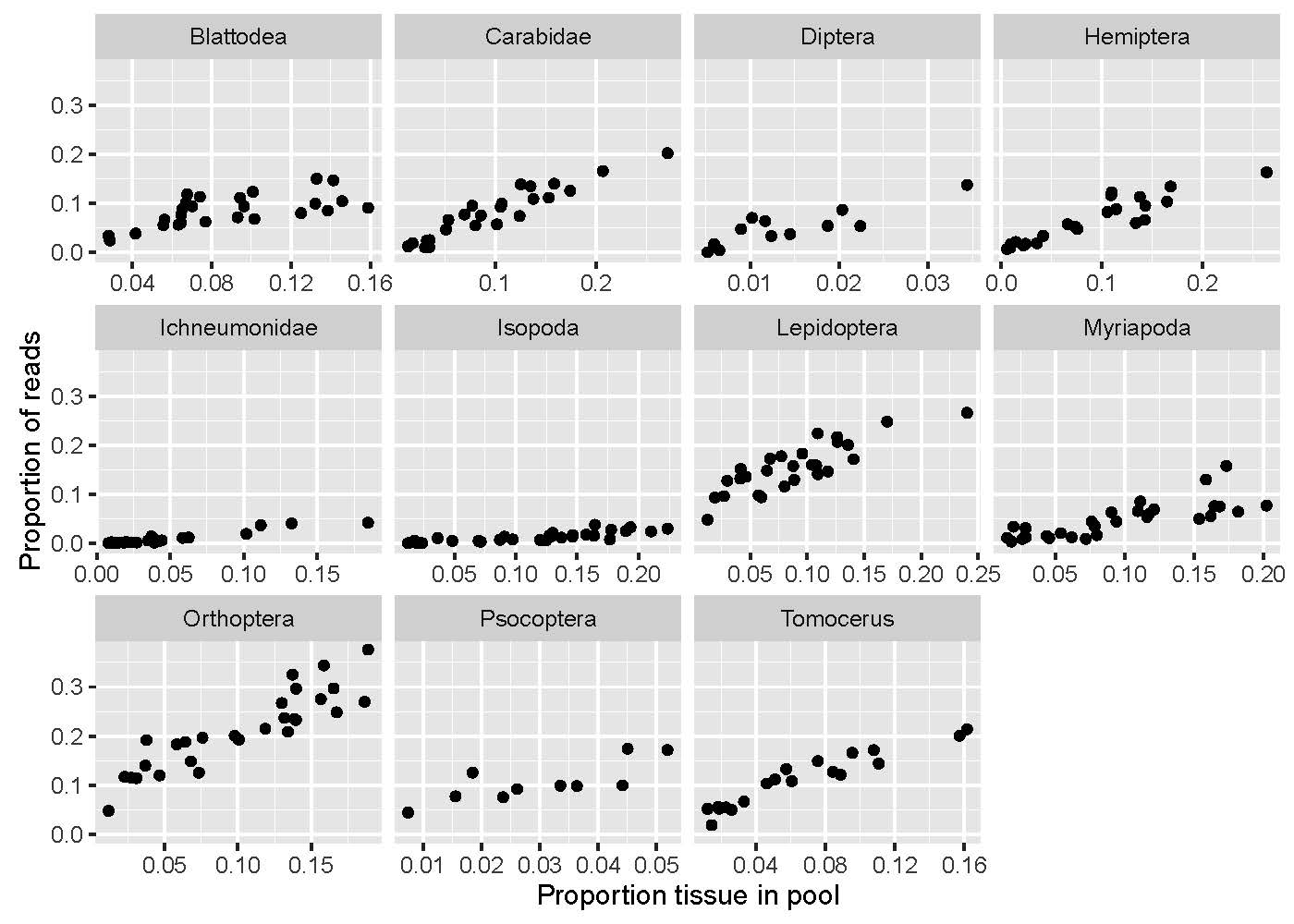
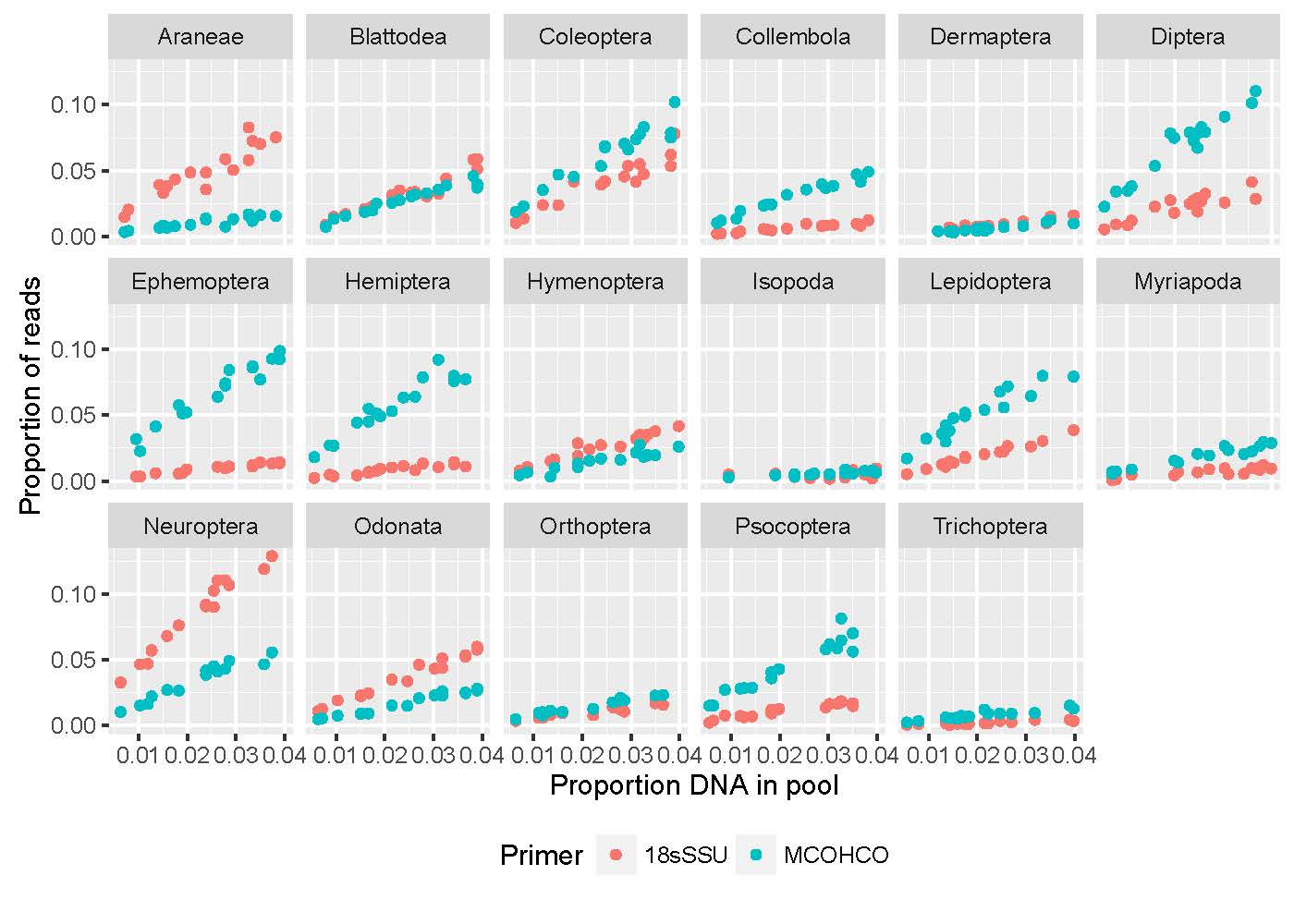


Figure 3. (A) Exemplary associations of proportion of DNA and proportion of recovered reads per taxon for DNA mock communities using mitochondrial COI and nuclear 18srDNA as marker. (B) Exemplary associations of proportion of tissue and proportion of recovered reads per taxon for tissue mock communities using mitochondrial COI.

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

Figure 4. Association of actual input DNA with two predicted estimates using the 6 cycle PCR experiment. Gray open circles represent a naïve estimate using the raw read count per taxon (input DNA is estimated as (proportion read count) \* (total DNA input). Closed black circles are calculated using our Bayesian correction method. This method was calibrated using the 26 cycle experiment and tested here on the largely independent 6 cycle experiment. The red line corresponds to the 1:1 line.

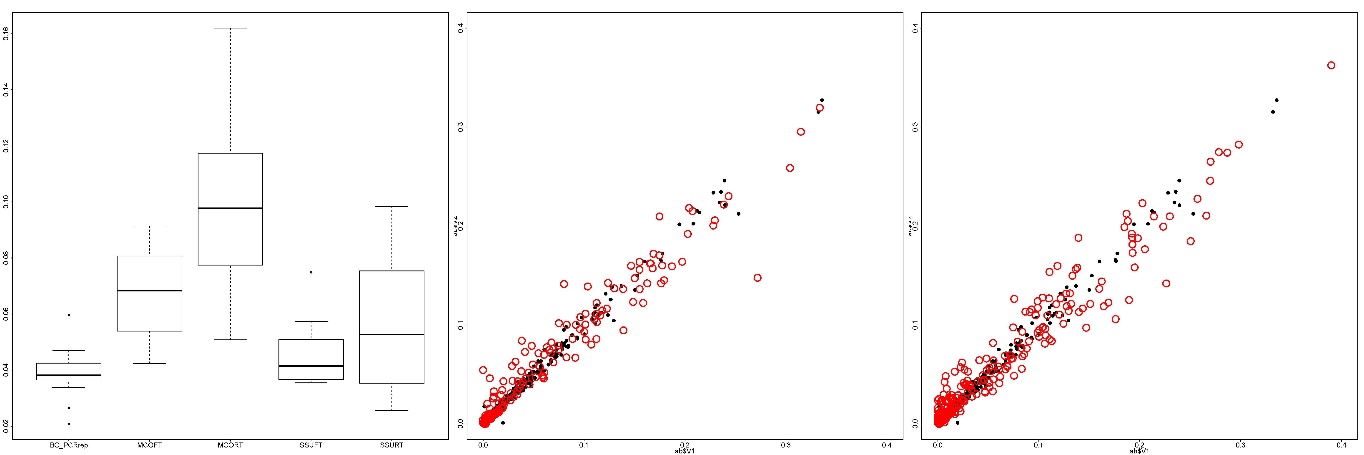


Figure 5. (A) Bray Curtis distance between exact PCR replicates of the same mock community samples, and high vs low molecular weight fractions for samples stored under freezer and room temperature conditions. (B) Association of mitochondrial COI read proportions from high and low molecular weight samples for all taxa of mock communities stored in the freezer and (C) for all samples stored at room temperature.