Environmental DNA Barcode Sequence Capture: Targeted, PCR-free Sequence Capture for Biodiversity Analysis from Bulk Environmental Samples

Shadi Shokralla[1], Joel F. Gibson [1, 2], Ian King[1], Donald J. Baird[3], Daniel H. Janzen[4], Winnie Hallwachs[4], and Mehrdad Hajibabaei[1*]

[1]Centre for Biodiversity Genomics and Department of Integrative Biology, University of Guelph, Guelph, Ontario, N1G 2W1, Canada

[2]Royal BC Museum, Victoria, British Columbia, V8W 9W2, Canada

[3]Environment Canada, Canadian Rivers Institute, University of New Brunswick, Fredericton, New Brunswick, E3B 5A3, Canada

[4]Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania, 19104-6303, USA

* To whom correspondence should be addressed. Mehrdad Hajibabaei, Tel: 1-519-824-4120 x52487; Fax: 1-519-824-5703; Email: mhajibab@uoguelph.ca.

Keywords: DNA barcoding, DNA metabarcoding, Environmental DNA, Sequence capture, Biodiversity

ABSTRACT

Environmental DNA analysis using PCR amplified marker genes has been a key application of high-throughput sequencing (HTS). However, PCR bias is a major drawback to gain accurate qualitative and quantitative biodiversity data. We developed a PCR-free approach using enrichment baits for species-specific mitochondrial cytochrome c oxidase 1(COI) DNA barcodes. The sequence capture was tested on species-rich bulk terrestrial and aquatic benthic samples. Hybridization capture recovered an average of 6 and 4.7 more arthropod orders than amplicon sequencing for terrestrial and benthic samples, respectively. For the terrestrial sample, the four most abundant arthropod orders comprised 94.0% of the sample biomass. These same four orders comprised 95.5% and 97.5% of the COI sequences recovered by amplification and capture, respectively. Hybridization capture recovered three arthropod orders that were detected by biomass analysis, but not by amplicon sequencing and two other insect orders that were not detected by either biomass or amplicon methods. These results indicate the advantage of using sequence capture for a more accurate analysis of biodiversity in bulk environmental samples. The protocol can be easily customized to other DNA barcode markers or gene regions of interest for a wide range of taxa or for a specific target group.

INTRODUCTION

The application of DNA sequence information for the identification of living organisms (e.g., DNA barcoding) has revolutionized biodiversity science (1,2). Customized, public databases of DNA barcodes have been assembled through concerted efforts such as the International Barcode of Life (iBOL) project. These databases employ standardized gene regions for different groups of organisms. For example, cytochrome *c* oxidase subunit I (*COI*) has been established as the standard DNA barcode for animals (1). An increasing number of animal species have been DNA barcoded and their sequences are available publicly through the GenBank and BOLD online data portals (3,4). Similarly, other DNA

barcode markers have been established and public libraries built for other kingdoms of life: 16S ribosomal DNA (*16S*) for bacteria and Archaea (5); 18S ribosomal DNA (*18S*) for protists (6,7); the nuclear internal transcribed spacer (*ITS*) for fungi (8); and *rbcL* and *matK* gene regions for plants (9).

Initial DNA barcoding efforts were based on dideoxy chain-termination ("Sanger") sequencing (10). While scalable, this approach is limited to producing single DNA barcode sequences for one individual at a time. High-throughput sequencing (HTS) (e.g., Illumina MiSeq) has been established as a viable means of assessing DNA barcode-based biodiversity from individuals (11-13) and mixed tissue samples (14-16). The HTS approach can produce a large number of sequences at a greatly reduced overall cost (11-13,15,17). All of these previous examples of massively parallel recovery of DNA barcode gene(s) have relied upon PCR amplification prior to HTS.

Protein-coding genes, such as *COI*, display more sequence variation between individuals and species as compared to ribosomal markers (18). Hence, they are generally considered more discriminating for the purposes of species identification and biodiversity analysis. However, any oligonucleotide primers designed for amplification of protein-coding gene regions may require the addition of degenerate or inert sites to improve taxonomic coverage. The use of degenerate PCR primer sets can still lead to amplification bias, especially when utilized to amplify environmental mixtures with high sequence diversity. Meanwhile, this bias can be exaggerated by the exponential nature of PCR amplification throughout the PCR cycles. This phenomenon can lead to over-amplification of some DNA templates at the expense of non- or under-amplification of other templates. Another possible effect of PCR amplification is the erroneous recovery of nuclear copies of mitochondrial DNA (i.e., NUMTs or pseudogenes). This amplification, sequencing, and reporting of paralogous gene regions can reach substantial rates in some instances (11,19).

It is important to note that primer amplification bias is not due solely to degeneracy and primer mismatch. Primer amplification bias exists even when highly conserved primers are used on mixed tissues. This bias is likely due to differences in GC content of amplicons, amplicon length, annealing temperature, overall genetic diversity of the initial mixture, and the binding energies of the primers

68 themselves (20-22). In order to account for PCR bias, most HTS studies design primers to amplify
69 taxonomic groups of interest. For example, Zeale et al. (23) designed *COI* primers to amplify 12 orders of
70 insects known to make up the diet of bats prior to amplifying and sequencing bat gut contents. In addition
71 to taxon-specific primer design, the employment of multiple PCR reactions with different primer sets has
72 been shown to lessen the impact of primer bias in HTS research. The multiple primer set approach has
73 been especially effective when attempting to recover sequence data from a broad taxonomic range. In
74 mixtures of arthropod tissue from both freshwater benthos (24) and terrestrial Malaise samples (15), the
75 use of multiple primer sets recovers a greater proportion of taxa present than any single primer set alone.
76 In addition to the possibility of PCR primer bias producing false negative results due to selective
77 exclusion of some species present in a mixture, the issue of organismal abundance has been raised. With
78 the possibility of some DNA templates being selectively amplified over others during the PCR
79 amplification process, relative abundance of DNA sequences cannot be used as a proxy for organismal
80 abundance. Some have attempted to reconcile this discrepancy through the estimation of DNA barcode
81 copy numbers, but this must be calculated one species at a time under controlled lab conditions (25).
82 The necessity of shorter target gene regions suitable for HTS has resulted in a move towards the
83 use of shorter regions of *COI* contained wholly within the standard *COI* barcode region (14,15, 26,27). In
84 each case, these shorter gene regions are tested *in silico* to ensure that the shorter region can still provide
85 adequate species-level discrimination. In contrast, similar tests of species-level discrimination for other
86 potential marker sequence regions (especially non-protein-coding loci) have given mixed results. For
87 example, short segments of the *ITS* region have proven to be valuable for identification in fungi (28). On
88 the other hand, the nuclear 16S/18S SSU rDNA region has been shown to lump species together
89 producing artificially low reports of biodiversity (29,30).
90 Hybridization capture followed by HTS has been used in previous instances, usually focusing on
91 individual test species. For instance, the capture method has been used to investigate vertebrate/virus
92 genomic integration (31-33), the phylogeography of small sets of vertebrate species (34,35), and ancient
93 plant DNA (36). When used to recover DNA regions associated with human genetic disease, the capture

method was able to supply accurate copy number information for multiple gene regions (37). Also using humans as test organisms, capture protocols were used to recover and reassemble complete individual mitochondrial genomes from mixed starting material (38).

We propose the use of hybridization capture methods, short DNA barcode regions, and HTS to eliminate the effects of PCR bias on biodiversity assessments of mixed environmental samples. We hypothesize that this approach can improve recovery of DNA barcode data from different types of bulk environmental samples as compared to PCR-based analysis. Hence, a capture-based approach should report fewer taxonomic false negatives as compared to a PCR-based method. Also, relative sequence abundance data should be more detailed and accurate than PCR-based relative sequence abundance data. This hypothesis is tested using capture baits designed from a DNA library representing 26 orders of Arthropoda and the tested tissue samples collected via two, distinct arthropod-targeting sampling methods. Malaise traps typically capture terrestrial insects, whereas benthic sampling using a pond net recovers a broad range of aquatic arthropods.

MATERIAL AND METHODS

Field sampling

A Malaise sample was collected at Bosque Humedo, Area de Conservación Guanacaste, northwestern Costa Rica (latitude 10.85145˚N; longitude -85.60801˚W; altitude 290 m; date January 24–31, 2011). The sample was collected directly into 95% ethanol, and frozen at -20°C until thawed and processed. In addition, three benthic samples (A, B, and C) were collected from Wood Buffalo National Park in Alberta, Canada (58.60273˚N; 111.52612˚W) in June, 2011. The samples were collected 50 meters apart from each other and were taken from the edge of the emergent vegetation zone into the submerged vegetation zone at each site. Samples were collected using a standard pond net with a sterile 400µm mesh net and attached collecting cup attached to a pole. Effort was standardized at 2 minutes per sample. Sampling was conducted by moving the net up and down through the vegetation in a sinusoidal pattern while maintaining constant forward motion. Samples were preserved in 95% ethanol in the field, and kept cold until processed. All DNA-friendly best practices—including wearing clean gloves to collect and

handle samples in the field and laboratory and decontaminating nets between samples—were followed to minimize the risk of DNA contamination between sites.

Biomass calculation

For the Malaise sample, all 1,066 morphologically identifiable individuals were isolated, identified to order morphologically, photographed, measured for total body length (excluding antennae) and maximum thoracic width, and tissue subsampled for individual DNA extraction. Dry biomass of each individual was estimated using order-specific length-width formulae (39).

DNA extraction

The Malaise sample and each of the benthic samples were individually blended in 95% ethanol and the resultant slurry was transferred to multiple sterile 50 mL Falcon tubes. After ethanol evaporation of the slurry at 56°C, the dried mixture was divided into three lysing matrix tubes "A" (about 100 mg each) and homogenized using the MP FastPrep-24 Instrument (MP Biomedicals Inc.) at speed 6 for 40 sec. Total DNA of this homogenized slurry was extracted using the Nucleospin tissue kit (Macherey-Nagel Inc.) following the manufacturer's instructions and eluted in 50 µL of molecular biology grade water.

Amplification-based approach

For all samples, two fragments within the standard *COI* DNA barcode region were amplified with two primer sets (F230R and BR5 for the Malaise sample; AD and BE for the three benthic samples), in a two-step PCR amplification regime (14). The primer sequences are as follows, F: 5`.GGTCAACAAATCATAAAGATATTGG .3` (40), 230_R: 5`. CTTATRTTRTTTATICGIGGRAAIGC.3` (this paper), A_F: 5`. GGIGGITTTGGIAATTGAYTIGTICC.3`, B_F:5`.CCIGAYATRGCI-TTYCCICG.3`, D_R: 5`.CCTARIATIGAIGARAYICCIGC.3` (24); and R5: 5`.GTRATIGCICCIG-CIARIAC.3` (15). The first PCR used *COI*-specific primers and the second PCR involved Illumina-tailed primers. The PCR reactions were assembled in 25 µL volumes. Each reaction contained 2 µL DNA template, 17.5 µL molecular biology grade water, 2.5 µL 10× reaction buffer (200 mM Tris–HCl, 500 mM KCl, pH 8.4), 1 µL $MgCl_2$

(50 mM), 0.5 µL dNTPs mix (10 mM), 0.5 µL forward primer (10 mM), 0.5 µL reverse primer (10 mM), and 0.5 µL Invitrogen's Platinum *Taq* polymerase (5 U/µL). The PCR conditions were initiated with heated lid at 95°C for 5 min., followed by a total of 30 cycles of 94°C for 40 s., 46°C (for both primer sets) for 1 min., and 72°C for 30 s., and a final extension at 72°C for 5 min., and hold at 4°C. Amplicons from each sample were purified using Qiagen's MiniElute PCR purification columns and eluted in 30 µL molecular biology grade water. The purified amplicons from the first PCR were used as templates in the second PCR with the same amplification condition used in the first PCR with the exception of using Illumina-tailed primers in a 15-cycle amplification regime. All PCRs were done using Eppendorf Mastercycler ep gradient S thermal cyclers and negative control reactions (no DNA template) were included in all experiments. PCR products were visualized on a 1.5% agarose gel to check the amplification success.

Selecting capture targets and probes design

The most important factor in designing a successful Sequence Capture Developer experiment is the quality of the input sequence used to select the target and capture probes. Conservative, hypervariable regions and copy number variation can introduce problems at the design stage (31,35). These problems increase significantly when creating designs for multiple species where some of the taxa are less represented in the database. In this study, a total of 79,215 *COI* DNA barcodes (47,631,471 bp) representing 26 arthropod orders, 559 families, and 4,035 genera were downloaded from both GenBank and BOLD in October, 2012. The 98% similarity clustering of the downloaded sequences resulted in 46,762 unique clusters (28,365,304 bp) (Supplementary Table S1). Sequence Capture Developer probes were designed with the help of the NimbleGen probes design team. A total of 2.1 million 50-105 mer probes (baits) were designed to cover all target clusters. Uniformity of probe abundance was considered, especially for highly conservative regions. The baits design covers 26,918,673 target bases (94.9%) from 46,762 (100%) of the clusters. Four Sequence Capture Developer probes reactions were ordered from NimbleGen, catalog number 06740278001.

Library preparation for Illumina MiSeq sequencing

For each of the Malaise and benthic samples, a total of 3 µg total DNA of each sample was sheared using a Covaris S220 Focused ultra-sonicator, resulting in a range of 300-800 bp fragments of each sample. The fragmentation efficiency was tested by running the fragmented DNA on an Agilent Bioanalyzer and DNA 1000 chip. Duplicate libraries with unique indexes were prepared with KAPA LTP library preparation kit (Catalog no. KK8230) according to the manufacturer's protocol.

Pre-capture pooling and hybridization

For the three benthic samples, the indexed libraries were quantified and equimolar concentrations of each were pooled before hybridization. According to the NimbleGen SeqCap EZ SR user guide v4.2, the pooled and un-pooled indexed libraries were blocked with a blend of COT DNA (Sigma Aldrich, catalog number: 05480647001), MB-grade fish sperm DNA (Roche Diagnostics, catalog number: 11467140001), and plant capture enhancer as a part of Sequence Capture Developer Reagent (Roche Diagnostics, catalog number: 06684335001) in addition to the universal sequencing adaptors and the used indexes. The blocked indexed libraries were hybridized to aliquots of SeqCap EZ library at 47°C for 72 hours with the thermocycler's head lid maintained at 57°C. The captured libraries were recovered using Streptavidin Dynabeads and washed twice with 1x stringent wash buffer preheated to 47°C followed by subsequent washes with wash buffer I, II, and III and finally recovered with 50 µL of molecular biology grade water. The captured libraries were amplified, purified, and quantified to be ready for sequencing.

Illumina MiSeq sequencing

The generated amplicons and captured libraries were sequenced in two MiSeq flow cells (one for amplicons and one for captured libraries) using a V3 MiSeq sequencing kit (300 × 2)(FC-131-1002 and MS-102-3001).

Bioinformatic processing

For the three benthic samples, a total of 24.4 million sequences were generated from PCR amplification and capture libraries, while a total of 11.0 million sequences were generated from PCR amplification and capture libraries of the Malaise sample. For each sample, the forward and reverse raw sequences were kept un-merged as the size of the capture libraries could be variable. All sequences were filtered for

quality using PRINSEQ software (41) with a minimum Phred score of 20, window of 10, step of 5, and a minimum length of 150 bp. A total of 17.6 million sequences for benthic samples (mean: 2.0 million reads/sample) and a total of 9.3 million reads for the Malaise sample (mean: 3.1 million reads/sample) were retained for further processing. USEARCH v6.0.307 (42), with the UCLUST algorithm, was used to de-replicate and cluster the remaining sequences using a 99% sequence similarity cutoff. This was done to de-noise any potential sequencing errors prior to further processing. Chimera filtering was performed using USEARCH with the 'de novo UCHIME' algorithm (43). At each step, cluster sizes were retained, singletons were retained, and only putatively non-chimeric reads were retained for further processing. All good quality, non-chimera clusters were identified using the MEGABLAST algorithm (44) against a reference library. This reference library contained all verified *COI* sequences downloaded from the GenBank database January 15, 2015 (N = 883,612 sequences). All MEGABLAST searches were conducted with a minimum alignment length of 100 bp and a minimum similarity of 90% for *COI* taxonomic identification recovery based on unambiguous top matches, and 98% for order-, family-, genus-, and species-level identification recovery. All sequencing data generated has been submitted to Dryad and can be accessed at (link will be added once accepted)

RESULTS

PCR-based sequencing

PCR amplification followed by sequencing produced between 285,680 and 2,559,402 total sequences for each sample (Table 1). Following quality filtering between 144,356 (50.5%) and 2,342,372 (91.5%) high-quality sequences remained for analysis. Of these high-quality sequences, between 135,560 (93.9%) and 2,075,555 (88.6%) sequences were identified as *COI* sequences (i.e., had at a minimum of 90% similarity to a reference *COI* sequence in GenBank).

Hybridization capture sequencing

Hybridization capture followed by sequencing produced between 3,010,292 and 4,444,412 total sequences for each sample (Table 1). Following quality filtering, between 1,700,512 and 3,713,290 (up to

85.73%) high-quality sequences per sample remained for analysis. Of these sequences, between 1,001,123 and 2,278,390 (up to 68.2%) sequences per sample were identified as *COI* sequences as mentioned before.

Taxonomic data recovery

For the benthic samples, hybridization capture followed by HTS recovered an average of 18 more orders than PCR amplification followed by HTS. For the Malaise sample, the average increase in order richness was 7. For the benthic samples, hybridization capture followed by HTS recovered an average of 31.8 more families than PCR amplification followed by HTS. For the Malaise sample, the average increase in family richness was 23. For the benthic samples, hybridization capture followed by HTS recovered an average of 37.2 more genera than PCR amplification followed by HTS. For the Malaise sample, the average increase in genus richness was 19.5.

Differences in taxonomic recovery considering only arthropod orders—for which the sampling and capture baits were designed—were also calculated. For the benthic samples, hybridization capture followed by HTS recovered an average of 4.7 more arthropod orders than PCR amplification followed by HTS (Figure 1). For the Malaise sample, the average increase in order richness was 6. For the benthic samples, hybridization capture followed by HTS recovered an average of 15 more arthropod families than PCR amplification followed by HTS. For the Malaise sample, the average increase in arthropod family richness was 24. For the benthic samples, hybridization capture followed by HTS recovered an average of 18.2 more arthropod genera than PCR amplification followed by HTS. For the Malaise sample, the average increase in arthropod genus richness was 21.5.

Relative sequence abundance

For the Malaise sample, the four most abundant orders (Diptera, Lepidoptera, Coleoptera, and Hymenoptera) were all arthropods and comprised 94.0% of the biomass of the sample (Supplementary Table S2). These same four orders comprised 95.5%, 97.5%, and 97.4% of the *COI* sequences recovered by PCR amplification, Capture 1, and Capture 2, respectively. Three arthropod orders (Blattodea,

Trombidiformes, and Psocoptera) were detected by morphological identification and biomass analysis, but not by any molecular method. Also, Coleoptera was present as a much higher proportion of biomass (35.6%), than as sequence proportion by any of the three methods (amplification 0.08%; capture 0.23% and 0.29%). Hybridization capture recovered three arthropod orders (Collembola, Neuroptera, Trichoptera) that were detected by biomass analysis, but not by PCR amplification. Furthermore, hybridization capture detected two insect orders (Mantodea, Plecoptera), in very low sequence numbers, which were not detected by either biomass analysis or PCR amplification (Supplementary Table S2).

For the Benthic A sample, at the order level, the three most abundant arthropod orders (Diptera, Hemiptera, Podocopida) combined represented 56.0%, 56.4%, and 59.3% of the *COI* sequences recovered for PCR amplification, Capture 1, and Capture 2, respectively. For the Benthic B sample, at the order level, the mentioned three most abundant arthropod orders combined represented 60.8%, 56.2%, and 55.0% of the *COI* sequences recovered for PCR amplification, Capture 1, and Capture 2, respectively. For the Benthic C sample, at the order level, the three most abundant arthropod orders combined composed 93.3%, 92.7%, and 92.9% of the *COI* sequences recovered for PCR amplification, Capture 1, and Capture 2, respectively (Figure 2).

A comparison between methods of the proportion of sequences assigned to each arthropod order, family, and genus was used to generate scatter plots and correlation values (Figure 3). Regardless of sample or taxonomic level, the two rounds of hybridization capture produced highly correlated (0.99 to 1.00) proportions assigned to each taxon. The degree of correlation between PCR amplification and hybridization capture varied (0.43 to 0.93), but was consistently below that of hybridization capture round one versus round two. Similarly, Bray-Curtis dissimilarity values calculated from the proportion by taxon matrices at the arthropod order, family, and genus level were calculated (Figure 4). Likewise, dissimilarity values were much lower for hybridization capture round one versus round two comparisons (0.011 to 0.077), than those for PCR amplification versus hybridization capture comparisons (0.166 to 0.615).

DISCUSSION

This study is the first targeted attempt at PCR-free DNA barcode recovery from mixed environmental samples. Previous work employed mitochondrial enrichment followed by sequencing of total enriched mtDNA (45). This method required the exclusion of 99.5% of sequence data in order to retain only the 0.5% that comprises informative *COI* DNA barcode data. Liu et al. (46) were able to increase the efficiency of whole mitochondrial genome sequencing to about 42.5% by using a capture microarray followed by HTS in a mock community sample containing equal aliquots of genomic DNA extracts of 49 species. The present method retains 52.4 to 68.2% of the passed filter sequences produced through Illumina MiSeq sequencing for final analysis (Table 1). Whether pooled as three lower diversity samples or retained as one single, high diversity, high sequencing coverage sample, this produces one to two million *COI* sequences per sample for further analysis. This rate of analyzable sequences does not differ between benthic- and Malaise-derived tissue samples.

The recovery of non-target DNA regions has been noted in previous use of capture enrichment in resequencing studies where multiple regions of the genome need to be enriched and analyzed (33,34,37). This occurrence is likely due to the capture of adjacent ("flanking") DNA regions during the hybridization process. The proportion of recovered sequences not matching targets ranges between 18 and 42% in these previous studies. In the present study, 31.8 to 47.6% of high-quality sequences produced through hybridization capture recovered were not positively identified as *COI*. This rate in target recovery, while similar to previous resequencing studies, differs fundamentally from past mitochondrial enrichment experiments. Previously, hybridization baits were designed and implemented to target a number of different gene regions for a narrow range of organisms (e.g., the entire mitochondrial genome of one species). The present use of hybridization capture targets many versions of one small gene region in order to retain sequence data from a taxonomically wide range of organisms. As such, the rate of non-target sequence generation can be interpreted as either capture of non-*COI* sequences with some affinity for the oligonucleotide baits used, or else sequencing error within the Illumina MiSeq resulting in uninformative sequences. Any truly "flanking" sequences adjacent to the *COI* region targeted could still be retained and identified as *COI* sequences.

Hybridization capture followed by HTS recovered approximately twice as many order, family, and genus names as compared to PCR amplification for the three benthic samples (Table 1, Figure 1). Likewise, hybridization capture followed by HTS recovered an increase in order, family, and genus richness for the Malaise sample. These increases in taxonomic recovery are more pronounced when only the Arthropoda are considered. These increases in biodiversity detection reflect a reduced false negative rate due to the absence of primer amplification bias. The difference in increased taxonomic recovery at each level reveals an important difference between the benthic and Malaise samples. For both Malaise and benthic samples, the majority of the orders recovered only through hybridization capture were non-insects, including other arthropods, vertebrates, diatoms, fungi, and plants. Benthic samples, being aquatic in nature, will include a number of orders not found in a terrestrial sample, especially zooplankton and phytoplankton. To rule out false positives, we considered only the sequencing clusters containing more than ten sequences per cluster. The PCR amplification primers employed were designed for insects and are unlikely to have amplified non-insect orders. The hybridization capture, despite also being designed from insect DNA sequences, employs multiple longer oligos complementary to multiple fragments of the *COI* gene region and was better able to recover non-insect taxa.

It has been stated that PCR primer bias is likely to make conventional barcode copy abundance calculations from HTS environmental DNA studies impossible (e.g., 47). The removal of this bias facilitates a shift towards more meaningful interpretations of sequence numbers recovered from environmental samples and decreases chances of false negatives for taxa with a small biomass or rare taxa (15,17). For all benthic and Malaise samples, there was consistency in relative sequence frequencies between hybridization capture attempts at the arthropod genus, family, and order levels (Figure 2, Table 2). In all samples, PCR amplification recovered a different community profile from hybridization capture (Figures 2, 3; Table 1).

In the Malaise sample, when compared to biomass calculations, the difference in proportion of sequences assigned to each taxon is noticed as an overabundance of Lepidoptera sequences at the expense of Coleoptera. In addition, three orders (Blattodea, Trombidiformes, and Psocoptera) are not detected by

HTS methods at all. This result is likely a product of the taxonomic composition of GenBank *COI* libraries. An abundance of tropical Lepidoptera are represented as GenBank records whereas few Costa Rican Blattodea, Psocoptera, and Trombidiformes are included. This lack of likely matches, coupled with our strict 98% similarity cut-off, resulted in a failure to recover these orders. As a further analysis, when similarity cut-offs were loosened to 90% similarity, with the same GenBank library, Blattodea, Trombidiformes, and Psocoptera were recovered with HTS methods, and Coleoptera was recovered as a much greater proportion of sequences (results not shown). However, a relaxed cut-off for sequence analyses such as BLAST could also result in reduced confidence in taxonomic identification of sequences or may result in false positives. As reference DNA barcode libraries are populated with diverse taxa and include better representation of local populations (haplotypes), sequence capture studies can benefit from these additional sequences both for the design of capture probes and in subsequent analyses of sequences. In fact, the use of standard DNA markers (e.g., barcodes) and provisioning reference databases with voucher specimens could facilitate large-scale analyses of environmental samples.

The use of HTS to allow parallel sequencing of multiple genetic markers in both library building and environmental mixture research has been previously proposed and demonstrated (11,12,15). While the present study focused on *COI* region as the most widely used DNA barcodes, the protocol could be easily adapted to include other DNA barcode regions, phylogenetic markers, or functional genes. The presented approach offers the flexibility to be customized to specific target groups and multiple barcoding markers. It also overcomes the challenges in PCR-based methods including primer design in hypervariable markers and the associated bias. The ability to multiplex many samples and recover individual sample data at sufficient sequencing depth is shown here to have no negative effect on biodiversity data recovery. The push to include as many samples and genetic markers as desired in future HTS-based biodiversity studies would be readily accommodated by the current protocol.

ACKNOWLEDGEMENT

We are grateful to Area de Conservación Guanacaste for protecting the forest habitat that we sampled.

FUNDING

This project was funded by the Government of Canada through Genome Canada, Ontario Genomics Institute, and Environment and Climate Change Canada through the Biomonitoring 2.0 project to M.H. and the National Science Foundation [grant number DEB 0515699 to D.H.J]. J.F.G. was also funded by a Natural Sciences and Engineering Research Council of Canada Postdoctoral Fellowship.

REFERENCES

1. Hebert, P.D.N., Cywinska, A., Ball, S.L. and deWaard, J.R. (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society B, 270, 313-321.
2. Hajibabaei,M., Singer,G.A., Clare,E.L. and Hebert,P.D.N. (2007) Design and applicability of DNA arrays and DNA barcodes in biodiversity monitoring. BMC Biology, 5, 24.
3. Ratnasingham,S. and Hebert,P.D.N. (2007) BOLD: the barcode of life data system (www.barcodinglife.org). Molecular Ecology Notes, 7, 355-364.
4. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2010) GenBank. Nucleic Acids Research, 39, D32-D37.
5. Tringe,S.G. and Hugenholtz,P. (2008) A renaissance for the pioneering 16S rRNA gene. Current Opinions in Microbiology, 11, 442-446.
6. Creer,S., Fonseca,V.G., Porazinska,D.L., Giblin-Davis,R.M., Sung,W., Power,D.M., Packer,M., Carvalho,G.R., Blaxter,M.L., Lambshead,P.J. et al. (2010) Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. Molecular Ecology, 19 Suppl 1, 4-20.
7. Weber,A.A. and Pawlowski,J. (2013) Can abundance of protists be inferred from sequence data: a case study of foraminifera. PLoS One, 8(2), e56739.
8. Schoch,C.L., Seifert,K.A., Huhndorf,S., Robert,V., Spouge,J.L., Levesque,C.A., Chen,W. and the Fungal Barcoding Consortium (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proceedings of the National Academy of Science USA, 109, 6241-6246.
9. CBOL Plant Working Group (2009) A DNA barcode for land plants. Proceedings of the National Academy of Science USA, 106, 12794-12797.
10. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Science USA, 74, 5463-5467.
11. Shokralla,S., Gibson,J.F., Nikbakht,H., Janzen,D.H., Hallwachs,W. and Hajibabaei,M. (2014) Next-generation DNA barcoding: using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. Molecular Ecology Resources, 14, 892-901.

12. Shokralla,S., Porter,T.M., Gibson,J.F., Dobosz,R., Janzen,D.H., Hallwachs,W., Golding,G.B. and Hajibabaei,M. (2015) Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. Scientific Reports, 5, 9687.
13. Meier,R., Wong,W., Srivathsan,A. and Foo,M. (2015) $1 DNA barcodes for reconstructing complex phenomes and finding rare species in specimen-rich samples. Cladistics, 32, 100-110.
14. Hajibabaei,M., Shokralla,S., Zhou,X., Singer,G.A.C. and Baird,D.J. (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. PLoS One, 6, e17497.
15. Gibson,J., Shokralla,S., Porter,T.M., King,I., vanKonynenburg,S., Janzen,D.H., Hallwachs,W. and 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 Science USA, 111, 8007-8012.
16. Leray,M. and Knowlton,N. (2015) DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. Proceedings of the National Academy of Science USA, 112, 2076-2081.
17. Shokralla,S., Spall,J.L., Gibson,J.F. and Hajibabaei,M. (2012) Next-generation sequencing technologies for environmental DNA research. Molecular Ecology, 21, 1794-1805.
18. Gibson,J.F., Skevington,J.H. and Kelso,S. (2010) Placement of Conopidae (Diptera) within Schizophora based on mtDNA and nrDNA gene regions. Molecular Phylogenetics and Evolution, 56, 91-103.
19. Haran,J., Koutroumpa,F., Magnoux,E., Roques,A. and Roux,G. (2015) Ghost mtDNA haplotypes generated by fortuitous NUMTs can deeply disturb infra-specific genetic diversity and phylogeographic pattern. Journal of Zoological Systematics and Evolutionary Research, 53, 109-115.
20. Suzuki,M.T. and Giovannoni,S.J. (1996) Bias caused by template annealing in the amplification of mixtures of 16S genes by PCR Applied and Environmental Microbiology, 62, 625-630.
21. Polz,M.F. and Cavanaugh,C.M. (1998) Bias in template-to-product ratios in multitemplate PCR. Applied and Environmental Microbiology, 64, 3724-3730.
22. Gonzalez,J.M., Portillo,M.C., Belda-Ferre,P. and Mira,A. (2012) Amplification by PCR artificially reduces the proportion of the rare biosphere in microbial communities. PLoS One, 7, e29973.

413 23. Zeale,M.R., Butlin,R.K., Barker,G.L., Lees,D.C. and Jones,G. (2011) Taxon-specific PCR for
414 DNA barcoding arthropod prey in bat faeces. Molecular Ecology Resources, 11, 236-244.
415 24. Hajibabaei,M., Spall,J.L., Shokralla,S. and vanKonynenburg,S. (2012) Assessing biodiversity of
416 a freshwater benthic macroinvertebrate community through nondestructive environmental
417 barcoding of DNA from preservative ethanol. BMC Ecology, 12, 28.
418 25. Darby,B.J., Todd,T.C. and Herman,M.A. (2013) High-throughput amplicon sequencing of rRNA
419 genes requires a copy number correction to accurately reflect the effects of management practices
420 on soil nematode community structure. Molecular Ecology, 22, 5456-5471.
421 26. Meusnier,I., Singer,G.A.C., Landry,J.-F., Hickey,D.A., Hebert,P.D.N. and Hajibabaei,M. (2008)
422 A universal DNA mini-barcode for biodiversity analysis. BMC Genomics, 9, 214.
423 27. Shokralla,S., Hellberg,R.S., Handy,S.M., King,I. and Hajibabaei,M. (2015) A DNA mini-
424 barcoding system for authentication of processes fish products. Scientific Reports, 5, 15894.
425 28. Porter,T.M. and Golding,G.B. (2011) Are similarity- or phylogeny-based methods more
426 appropriate for classifying internal transcribed spacer (ITS) metagenomic amplicons? New
427 Phytologist, 192, 775-782.
428 29. Tang,C.Q., Leasi,F., Obertegger,U., Kieneke,A., Barraclough,T.G. and Fontaneto,D. (2012) The
429 widely used small subunit 18S rDNA molecule greatly underestimates true diversity in
430 biodiversity surveys of the meiofauna. Proceedings of the National Academy of Science USA,
431 109, 16208-16212.
432 30. Zhan,A., Bailey,S.A., Heath,D.D. and MacIsaac,H.J. (2014) Performance comparison of genetic
433 markers for high-throughput sequencing-based biodiversity assessment in complex communities.
434 Molecular Ecology Resources, 14, 1049-1059.
435 31. Duncavage,E.J., Magrini,V., Becker,N., Armstrong,J.R., Demeter,R.T., Wylie,T., Abel,H.J. and
436 Pfeifer,J.D. (2011) Hybrid capture and next-generation sequencing identify viral integration sites
437 from formalin-fixed, paraffin-embedded tissue. The Journal of Molecular Diagnostics, 13, 325-
438 333.

439 32. Tsangaras,K., Siracusa,M.C., Nikolaidis,N., Ishida,Y., Cui,P., Vielgrader,H., Helgen,K.M.,
440 Roca,A.L. and Greenwood,A.D. (2014) Hybridization capture reveals evolution and conservation
441 across the entire Koala retrovirus genome. PLoS One, 9, e95633.
442 33. Tsangaras,K., Wales,N., Sicheritz-Pontén,T., Rasmussen,S., Michaux,J., Ishida,Y., Morand,S.,
443 Kampmann,M.-L., Gilbert,M.T.P. and Greenwood,A.D. (2014) Hybridization capture using short
444 PCR products enriches small genomes by capturing flanking sequences (CapFlank). PLoS One, 9,
445 e109101.
446 34. Mason,V.C., Li,G., Helgen,K.M. and Murphy,W.J. (2011) Efficient cross-species capture
447 hybridization and next-generation sequencing of mitochondrial genomes from noninvasively
448 sampled museum specimens. Genome Research, 21, 1695-1704.
449 35. Peñalba,J.V., Smith,L.L., Tonione,M.A., Sass,C., Hykin,S.M., Skipwith,P.L., McGuire,J.A.,
450 Bowie,R.C.K. and Moritz,C. (2014) Sequence capture using PCR-generated probes: a cost-
451 effective method of targeted high-throughput sequencing for nonmodel organisms. Molecular
452 Ecology Resources, 14, 1000-1010.
453 36. Ávila-Arcos,M.C., Cappellini,E., Romero-Navarro,J.A., Wales,N., Moreno-Mayar,J.V.,
454 Rasmussen,M., Fordyce,S.L., Montiel,R., Vielle-Calzada,J.-P., Willerslev,E. et al. (2011)
455 Application and comparison of large-scale solution-based DNA capture-enrichment methods on
456 ancient DNA. Scientific Reports, 1, 74.
457 37. Herman,D.S., Hovingh,G.K., Iartchouk,O., Rehm,H.L., Kucherlapati,R., Seidman,J.G. and
458 Siedman,C.E. (2009) Filter-based hybridization capture of subgenomes enables resequencing and
459 copy-number detection. Nature Methods, 6, 507-510.
460 38. Maricic,T., Whitten,M. and Pääbo,S. (2010) Multiplexed DNA sequence capture of
461 mitochondrial genomes using PCR products. PLoS One, 5, e14004.
462 39. Gruner,D.S. (2003) Regressions of length and width to predict arthropod biomass in the Hawaiian
463 Islands. Pacific Science, 57, 325-336.

40. Folmer,O., Black,M., Hoeh,W., Lutz,R. and Vrijenhoek,R. (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Molecular Marine Biology and Biotechnology, 3, 294-299.
41. Schmieder,R. and Edwards,R. (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics, 27, 863-864.
42. Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26, 2460-2461.
43. Edgar,R.C., Haas,B.J., Clemente,J.C., Quince,C. and Knight,R. (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinformatics, 27, 2194-2200.
44. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. Journal of Computational Biology, 7, 203-214.
45. Zhou,X., Li,Y., Liu,S., Yang,Q., Su,X., Zhou,L., Tang,M., Fu,R., Li,J. & Huang,Q. (2013) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. GigaScience, 2, 4.
46. Liu S., Wang X., Xie L., Tan M., Li Z., Zhang H., Misof B., Kjer KM., Tang M., Niehuis O., Jiang H., and Zhou X. (2016) Mitochondrial capture enriches mito-DNA 100 fold, enabling PCR-free mitogenomics biodiversity analysis. Molecular Ecology Resources, 16, 470–479.
47. Piñol, J., Mir, G., Gomez-Polo, P. & Agustí, N. (2014) Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. Molecular Ecology Resources, 15, 819-830.

485 **Figure legneds**

486 **Figure 1.** Taxonomic richness recovery at three levels for four different subsamples using PCR
487 amplification and hybridization capture.

488 **Figure 2.** Proportion of DNA sequences identified to order for four different subsamples using PCR
489 amplification and hybridization capture. Amp – sequences recovered via PCR amplification. Cap1, Cap2 -
490 sequences recovered via hybridization capture in the first and second attempts.

491 **Figure 3.** Scatter plots and correlation values of proportion of sequences assigned to each arthropod
492 order, family, and genus for four different subsamples using PCR amplification and hybridization capture.

493 **Figure 4.** Bray-Curtis dissimilarity values based on proportion of sequences assigned to each arthropod
494 order, family, and genus for four different subsamples using PCR amplification and hybridization capture.

495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512

513 **Tables**

514

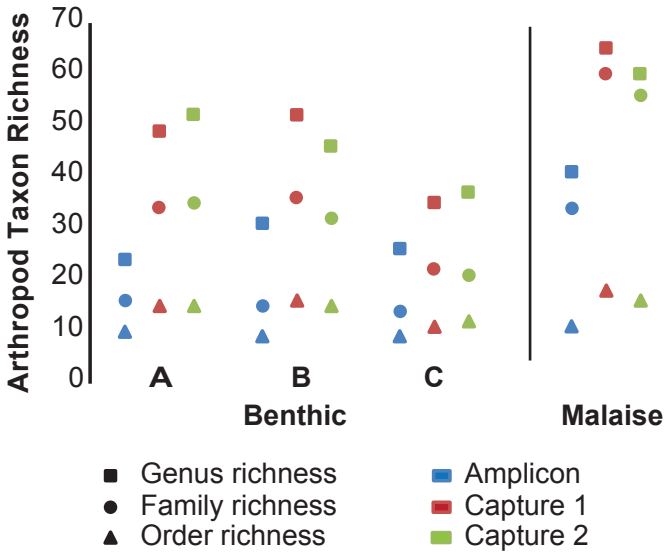
515 Table 1. Summary of number of DNA sequences produced, filtered, and used via four different samples

516 and two different methods.

Sample	Process	# raw DNA sequences	# sequences passing quality filter	# *COI* sequences	% high quality sequences used in *COI* analysis
	Amplicon	285,680	144,356	135,560	93.9
Benthic A	**Capture 1**	4,331,460	3,713,290	1,954,225	52.4
	Capture 2	4,023,194	2,993,002	1,939,291	64.8
	Amplicon	582,590	341,528	304,548	89.2
Benthic B	**Capture 1**	4,042,778	2,703,271	1,695,684	62.7
	Capture 2	3,010,292	1,700,512	1,001,123	58.9
	Amplicon	510,396	411,362	373,668	90.8
Benthic C	**Capture 1**	3,438,490	2,558,027	1,744,542	68.2
	Capture 2	4,177,500	3,027,892	2,046,278	67.6
	Amplicon	2,559,402	2,342,372	2,075,555	88.6
Malaise	**Capture 1**	4,444,412	3,499,795	2,278,390	65.1
	Capture 2	4,042,698	3,436,537	2,041,300	59.4

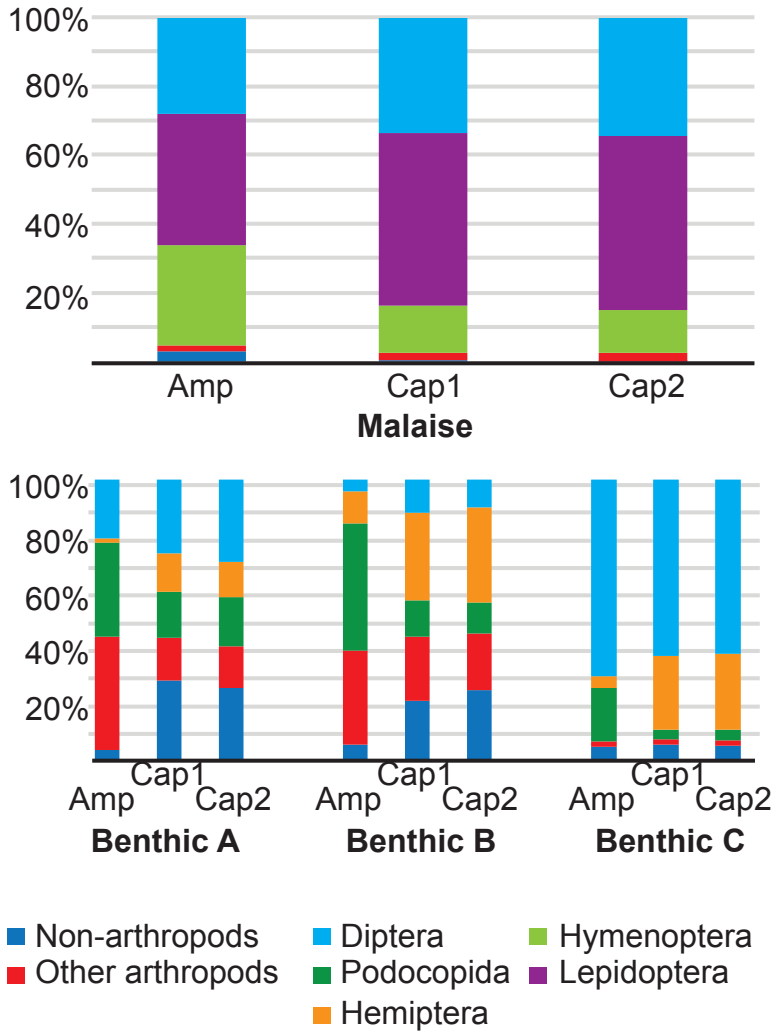
517

518



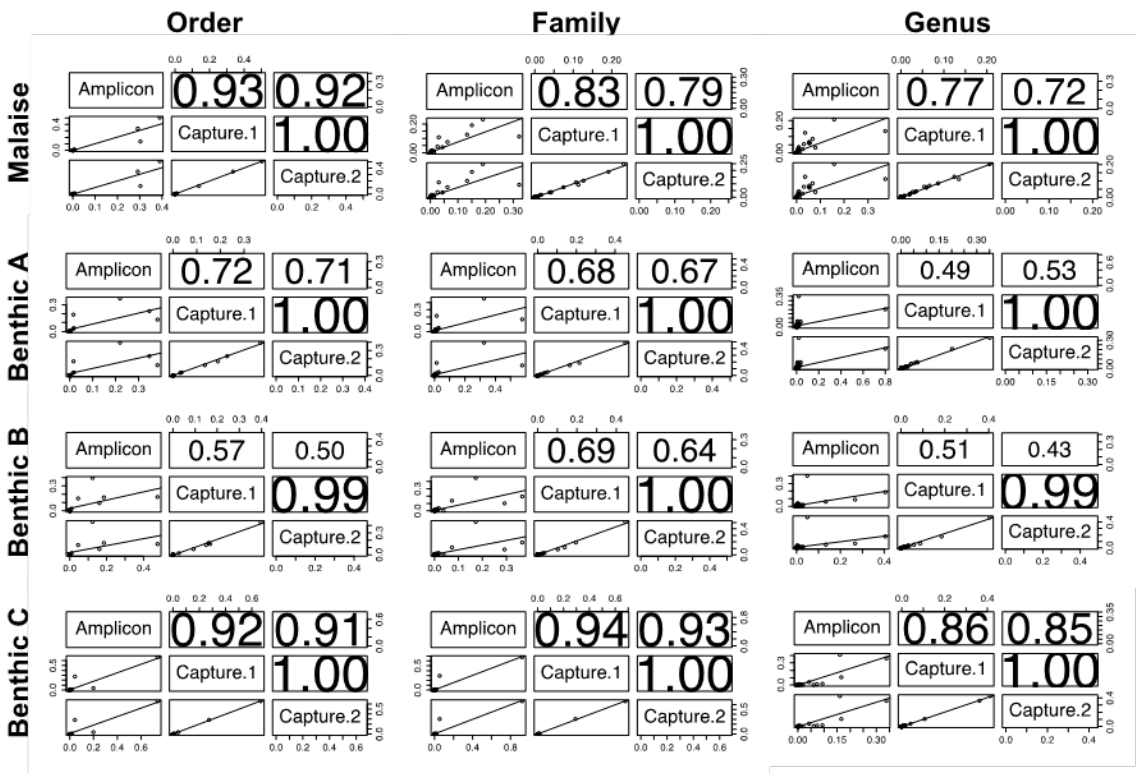
519

520



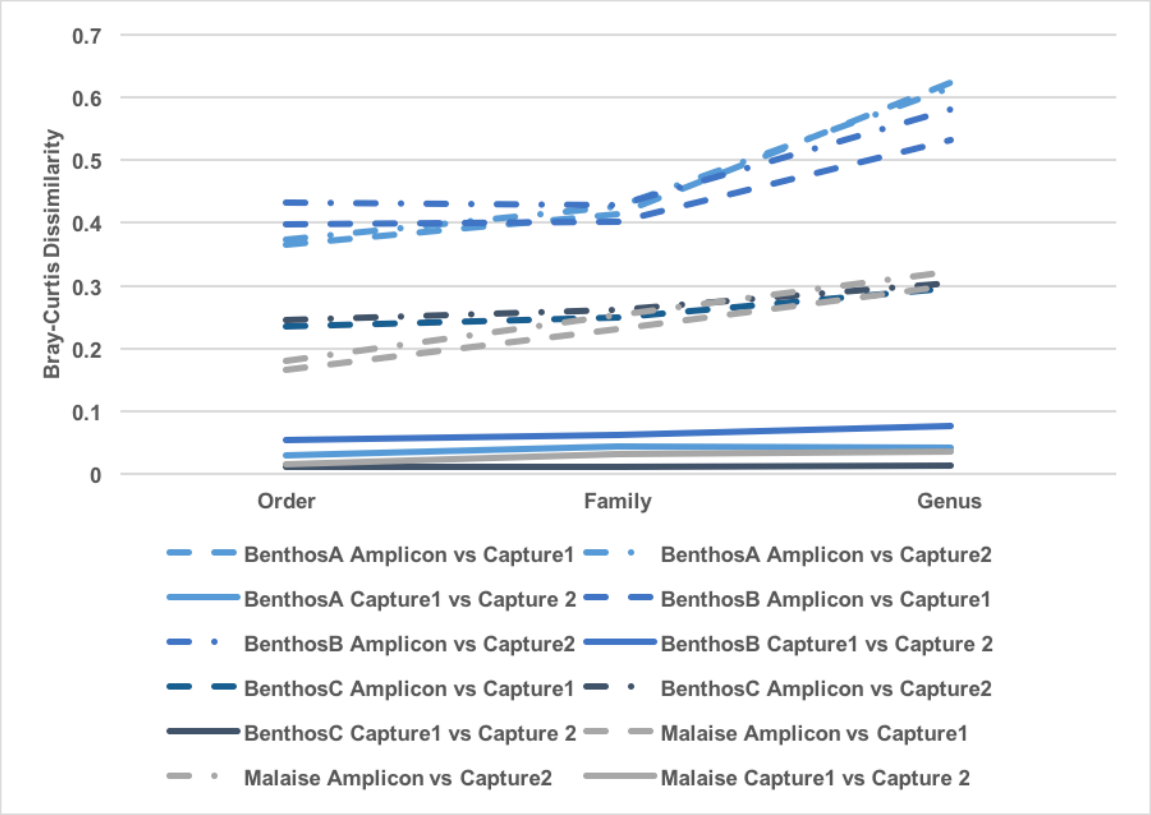
521

522



523

524



525

526

527 **Supplementary material**

528 **Table S1.** Summary of sequences included in oligo development.

Order	# Families	# Genera	#of sequences	# of bases
COLEOPTERA	80	764	6,531	3,756,110
DIPTERA	62	640	11,056	6,885,416
HEMIPTERA	60	407	4,010	2,390,925
HYMENOPTERA	36	257	14,358	8,745,477
ARACHNIDA	68	498	8,587	5,389,853
ORTHOPTERA	14	189	1,586	875,654
PSOCOPTERA **STREPSIPTERA** **THYSANOPTERA**	6	25	497	276,365
DERMAPTERA	1	1	34	21,751
ISOPODA	28	92	1193	702,654
EPHEMEROPTERA	15	47	2895	1,786,921
Blattaria **Dermaptera** **Isoptera** **Mecoptera** **Megaloptera** **Neuroptera** **Odonata**	49	172	1,964	1,196,335
PLECOPTERA	11	42	1,624	1,005,905
TRICHOPTERA	43	196	6,986	4,421,067

MANTODEA	12	90	135	79,097
TROMBIDIFORMES	17	22	11,453	6,095,705
COLLEMBOLA	16	38	3,265	2,056,296
LEPIDOPTERA	39	552	2,997	1,918,309
PHASMATODEA	2	3	44	27631

Table S2. Proportion of calculated total biomass and proportion of *COI* DNA sequences identified to insect, arthropod, and non-arthropod orders via two different methods for a single Malaise sample.

	Order	Biomass	Amplicon	Capture 1	Capture 2
Insects	**Blattodea**	0.710%			
	Coleoptera	35.600%	0.080%	0.235%	0.291%
	Diptera	27.500%	28.288%	33.595%	34.687%
	Ephemeroptera		0.018%	0.031%	0.024%
	Hemiptera	1.390%	0.580%	0.953%	0.992%
	Hymenoptera	26.310%	29.461%	13.775%	12.243%
	Lepidoptera	4.620%	37.835%	50.324%	50.635%
	Mantodea			0.005%	0.002%
	Neuroptera	0.030%		0.008%	0.005%
	Orthoptera	2.470%	0.817%	0.886%	0.933%
	Plecoptera			0.002%	0.002%
	Psocoptera	1.050%			
	Thysanoptera	0.080%	0.013%	0.016%	0.025%
	Trichoptera	0.020%		0.008%	0.023%
Non-Insect	**Amphipoda**			0.002%	0.002%
Arthropods	**Araneae**	0.090%	0.031%	0.002%	
	Collembola	0.020%		0.013%	0.008%