

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

3

4 Shadi Shokralla¹, Joel F. Gibson^{1,2}, Ian King¹, Donald J. Baird³, Daniel H. Janzen⁴, Winnie Hallwachs⁴,
5 and Mehrdad Hajibabaei^{1*}

6

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

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

10 ³Environment Canada, Canadian Rivers Institute, University of New Brunswick, Fredericton, New
11 Brunswick, E3B 5A3, Canada

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

13 * To whom correspondence should be addressed. Mehrdad Hajibabaei, Tel: 1-519-824-4120 x52487; Fax:
14 1-519-824-5703; Email: mhajibab@uoguelph.ca.

15

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

17

18 **ABSTRACT**

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

34 **INTRODUCTION**

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

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

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

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

64 It is important to note that primer amplification bias is not due solely to degeneracy and primer
65 mismatch. Primer amplification bias exists even when highly conserved primers are used on mixed
66 tissues. This bias is likely due to differences in GC content of amplicons, amplicon length, annealing
67 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

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

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

107 **MATERIAL AND METHODS**

108 **Field sampling**

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

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

122 **Biomass calculation**

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

127 **DNA extraction**

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

134 **Amplification-based approach**

135 For all samples, two fragments within the standard *COI* DNA barcode region were amplified with two
136 primer sets (F230R and BR5 for the Malaise sample; AD and BE for the three benthic samples), in a two-
137 step PCR amplification regime (14). The primer sequences are as follows, F:

138 5'.GGTCAACAAATCATAAAGATATTGG .3' (40), 230_R: 5'.

139 CTTATRTTRTTTATICGIGGRAAIGC.3' (this paper), A_F: 5'.

140 GGIGGITTTGGIAATTGAYTIGTICC.3', B_F:5'.CCIGAYATRGCI-TTYCCICG.3', D_R:

141 5'.CCTARIATIGAIGARAYICCIGC.3' (24); and R5: 5'.GTRATIGCICIG-CIARIAC.3' (15). The first

142 PCR used *COI*-specific primers and the second PCR involved Illumina-tailed primers. The PCR reactions

143 were assembled in 25 µL volumes. Each reaction contained 2 µL DNA template, 17.5 µL molecular

144 biology grade water, 2.5 µL 10× reaction buffer (200 mM Tris–HCl, 500 mM KCl, pH 8.4), 1 µL MgCl₂

145 (50 mM), 0.5 μ L dNTPs mix (10 mM), 0.5 μ L forward primer (10 mM), 0.5 μ L reverse primer (10 mM),
146 and 0.5 μ L Invitrogen's Platinum *Taq* polymerase (5 U/ μ L). The PCR conditions were initiated with
147 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
148 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
149 from each sample were purified using Qiagen's MiniElute PCR purification columns and eluted in 30 μ L
150 molecular biology grade water. The purified amplicons from the first PCR were used as templates in the
151 second PCR with the same amplification condition used in the first PCR with the exception of using
152 Illumina-tailed primers in a 15-cycle amplification regime. All PCRs were done using Eppendorf
153 Mastercycler ep gradient S thermal cyclers and negative control reactions (no DNA template) were
154 included in all experiments. PCR products were visualized on a 1.5% agarose gel to check the
155 amplification success.

156 **Selecting capture targets and probes design**

157 The most important factor in designing a successful Sequence Capture Developer experiment is the
158 quality of the input sequence used to select the target and capture probes. Conservative, hypervariable
159 regions and copy number variation can introduce problems at the design stage (31,35). These problems
160 increase significantly when creating designs for multiple species where some of the taxa are less
161 represented in the database. In this study, a total of 79,215 *COI* DNA barcodes (47,631,471 bp)
162 representing 26 arthropod orders, 559 families, and 4,035 genera were downloaded from both GenBank
163 and BOLD in October, 2012. The 98% similarity clustering of the downloaded sequences resulted in
164 46,762 unique clusters (28,365,304 bp) (Supplementary Table S1). Sequence Capture Developer probes
165 were designed with the help of the NimbleGen probes design team. A total of 2.1 million 50-105 mer
166 probes (baits) were designed to cover all target clusters. Uniformity of probe abundance was considered,
167 especially for highly conservative regions. The baits design covers 26,918,673 target bases (94.9%) from
168 46,762 (100%) of the clusters. Four Sequence Capture Developer probes reactions were ordered from
169 NimbleGen, catalog number 06740278001.

170 **Library preparation for Illumina MiSeq sequencing**

171 For each of the Malaise and benthic samples, a total of 3 µg total DNA of each sample was sheared using
172 a Covaris S220 Focused ultra-sonicator, resulting in a range of 300-800 bp fragments of each sample. The
173 fragmentation efficiency was tested by running the fragmented DNA on an Agilent Bioanalyzer and DNA
174 1000 chip. Duplicate libraries with unique indexes were prepared with KAPA LTP library preparation kit
175 (Catalog no. KK8230) according to the manufacturer's protocol.

176 **Pre-capture pooling and hybridization**

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

188 **Illumina MiSeq sequencing**

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

192 **Bioinformatic processing**

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

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

212 **RESULTS**

213 **PCR-based sequencing**

214 PCR amplification followed by sequencing produced between 285,680 and 2,559,402 total
215 sequences for each sample (Table 1). Following quality filtering between 144,356 (50.5%) and 2,342,372
216 (91.5%) high-quality sequences remained for analysis. Of these high-quality sequences, between 135,560
217 (93.9%) and 2,075,555 (88.6%) sequences were identified as *COI* sequences (i.e., had at a minimum of
218 90% similarity to a reference *COI* sequence in GenBank).

219 **Hybridization capture sequencing**

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

222 85.73%) high-quality sequences per sample remained for analysis. Of these sequences, between
223 1,001,123 and 2,278,390 (up to 68.2%) sequences per sample were identified as *COI* sequences as
224 mentioned before.

225 **Taxonomic data recovery**

226 For the benthic samples, hybridization capture followed by HTS recovered an average of 18 more
227 orders than PCR amplification followed by HTS. For the Malaise sample, the average increase in order
228 richness was 7. For the benthic samples, hybridization capture followed by HTS recovered an average of
229 31.8 more families than PCR amplification followed by HTS. For the Malaise sample, the average
230 increase in family richness was 23. For the benthic samples, hybridization capture followed by HTS
231 recovered an average of 37.2 more genera than PCR amplification followed by HTS. For the Malaise
232 sample, the average increase in genus richness was 19.5.

233 Differences in taxonomic recovery considering only arthropod orders—for which the sampling
234 and capture baits were designed—were also calculated. For the benthic samples, hybridization capture
235 followed by HTS recovered an average of 4.7 more arthropod orders than PCR amplification followed by
236 HTS (Figure 1). For the Malaise sample, the average increase in order richness was 6. For the benthic
237 samples, hybridization capture followed by HTS recovered an average of 15 more arthropod families than
238 PCR amplification followed by HTS. For the Malaise sample, the average increase in arthropod family
239 richness was 24. For the benthic samples, hybridization capture followed by HTS recovered an average of
240 18.2 more arthropod genera than PCR amplification followed by HTS. For the Malaise sample, the
241 average increase in arthropod genus richness was 21.5.

242 **Relative sequence abundance**

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

247 Trombidiformes, and Psocoptera) were detected by morphological identification and biomass analysis,
248 but not by any molecular method. Also, Coleoptera was present as a much higher proportion of biomass
249 (35.6%), than as sequence proportion by any of the three methods (amplification 0.08%; capture 0.23%
250 and 0.29%). Hybridization capture recovered three arthropod orders (Collembola, Neuroptera,
251 Trichoptera) that were detected by biomass analysis, but not by PCR amplification. Furthermore,
252 hybridization capture detected two insect orders (Mantodea, Plecoptera), in very low sequence numbers,
253 which were not detected by either biomass analysis or PCR amplification (Supplementary Table S2).

254 For the Benthic A sample, at the order level, the three most abundant arthropod orders (Diptera,
255 Hemiptera, Podocopida) combined represented 56.0%, 56.4%, and 59.3% of the *COI* sequences recovered
256 for PCR amplification, Capture 1, and Capture 2, respectively. For the Benthic B sample, at the order
257 level, the mentioned three most abundant arthropod orders combined represented 60.8%, 56.2%, and
258 55.0% of the *COI* sequences recovered for PCR amplification, Capture 1, and Capture 2, respectively. For
259 the Benthic C sample, at the order level, the three most abundant arthropod orders combined composed
260 93.3%, 92.7%, and 92.9% of the *COI* sequences recovered for PCR amplification, Capture 1, and Capture
261 2, respectively (Figure 2).

262 A comparison between methods of the proportion of sequences assigned to each arthropod order,
263 family, and genus was used to generate scatter plots and correlation values (Figure 3). Regardless of
264 sample or taxonomic level, the two rounds of hybridization capture produced highly correlated (0.99 to
265 1.00) proportions assigned to each taxon. The degree of correlation between PCR amplification and
266 hybridization capture varied (0.43 to 0.93), but was consistently below that of hybridization capture round
267 one versus round two. Similarly, Bray-Curtis dissimilarity values calculated from the proportion by taxon
268 matrices at the arthropod order, family, and genus level were calculated (Figure 4). Likewise,
269 dissimilarity values were much lower for hybridization capture round one versus round two comparisons
270 (0.011 to 0.077), than those for PCR amplification versus hybridization capture comparisons (0.166 to
271 0.615).

272 **DISCUSSION**

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

284 The recovery of non-target DNA regions has been noted in previous use of capture enrichment in
285 resequencing studies where multiple regions of the genome need to be enriched and analyzed (33,34,37).
286 This occurrence is likely due to the capture of adjacent (“flanking”) DNA regions during the
287 hybridization process. The proportion of recovered sequences not matching targets ranges between 18 and
288 42% in these previous studies. In the present study, 31.8 to 47.6% of high-quality sequences produced
289 through hybridization capture recovered were not positively identified as *COI*. This rate in target
290 recovery, while similar to previous resequencing studies, differs fundamentally from past mitochondrial
291 enrichment experiments. Previously, hybridization baits were designed and implemented to target a
292 number of different gene regions for a narrow range of organisms (e.g., the entire mitochondrial genome
293 of one species). The present use of hybridization capture targets many versions of one small gene region
294 in order to retain sequence data from a taxonomically wide range of organisms. As such, the rate of non-
295 target sequence generation can be interpreted as either capture of non-*COI* sequences with some affinity
296 for the oligonucleotide baits used, or else sequencing error within the Illumina MiSeq resulting in
297 uninformative sequences. Any truly “flanking” sequences adjacent to the *COI* region targeted could still
298 be retained and identified as *COI* sequences.

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

314 It has been stated that PCR primer bias is likely to make conventional barcode copy abundance
315 calculations from HTS environmental DNA studies impossible (e.g., 47). The removal of this bias
316 facilitates a shift towards more meaningful interpretations of sequence numbers recovered from
317 environmental samples and decreases chances of false negatives for taxa with a small biomass or rare taxa
318 (15,17). For all benthic and Malaise samples, there was consistency in relative sequence frequencies
319 between hybridization capture attempts at the arthropod genus, family, and order levels (Figure 2, Table
320 2). In all samples, PCR amplification recovered a different community profile from hybridization capture
321 (Figures 2, 3; Table 1).

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

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

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

348 **ACKNOWLEDGEMENT**

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

350 **FUNDING**

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

355 REFERENCES

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

- 382 12. Shokralla,S., Porter,T.M., Gibson,J.F., Dobosz,R., Janzen,D.H., Hallwachs,W., Golding,G.B. and
383 Hajibabaei,M. (2015) Massively parallel multiplex DNA sequencing for specimen identification
384 using an Illumina MiSeq platform. *Scientific Reports*, 5, 9687.
- 385 13. Meier,R., Wong,W., Srivathsan,A. and Foo,M. (2015) \$1 DNA barcodes for reconstructing
386 complex phenomes and finding rare species in specimen-rich samples. *Cladistics*, 32, 100-110.
- 387 14. Hajibabaei,M., Shokralla,S., Zhou,X., Singer,G.A.C. and Baird,D.J. (2011) Environmental
388 barcoding: a next-generation sequencing approach for biomonitoring applications using river
389 benthos. *PLoS One*, 6, e17497.
- 390 15. Gibson,J., Shokralla,S., Porter,T.M., King,I., vanKonynenburg,S., Janzen,D.H., Hallwachs,W.
391 and Hajibabaei,M. (2014) Simultaneous assessment of the macrobiome and microbiome in a bulk
392 sample of tropical arthropods through DNA metasytematics. *Proceedings of the National
393 Academy of Science USA*, 111, 8007-8012.
- 394 16. Leray,M. and Knowlton,N. (2015) DNA barcoding and metabarcoding of standardized samples
395 reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Science
396 USA*, 112, 2076-2081.
- 397 17. Shokralla,S., Spall,J.L., Gibson,J.F. and Hajibabaei,M. (2012) Next-generation sequencing
398 technologies for environmental DNA research. *Molecular Ecology*, 21, 1794-1805.
- 399 18. Gibson,J.F., Skevington,J.H. and Kelso,S. (2010) Placement of Conopidae (Diptera) within
400 Schizophora based on mtDNA and nrDNA gene regions. *Molecular Phylogenetics and Evolution*,
401 56, 91-103.
- 402 19. Haran,J., Koutroumpa,F., Magnoux,E., Roques,A. and Roux,G. (2015) Ghost mtDNA haplotypes
403 generated by fortuitous NUMTs can deeply disturb infra-specific genetic diversity and
404 phylogeographic pattern. *Journal of Zoological Systematics and Evolutionary Research*, 53, 109-
405 115.
- 406 20. Suzuki,M.T. and Giovannoni,S.J. (1996) Bias caused by template annealing in the amplification
407 of mixtures of 16S genes by PCR *Applied and Environmental Microbiology*, 62, 625-630.
- 408 21. Polz,M.F. and Cavanaugh,C.M. (1998) Bias in template-to-product ratios in multitemplate PCR.
409 *Applied and Environmental Microbiology*, 64, 3724-3730.
- 410 22. Gonzalez,J.M., Portillo,M.C., Belda-Ferre,P. and Mira,A. (2012) Amplification by PCR
411 artificially reduces the proportion of the rare biosphere in microbial communities. *PLoS One*, 7,
412 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 vanKonyenburg, 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 processed 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., Kieneker, 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.

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

484

485 **Figure legends**

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

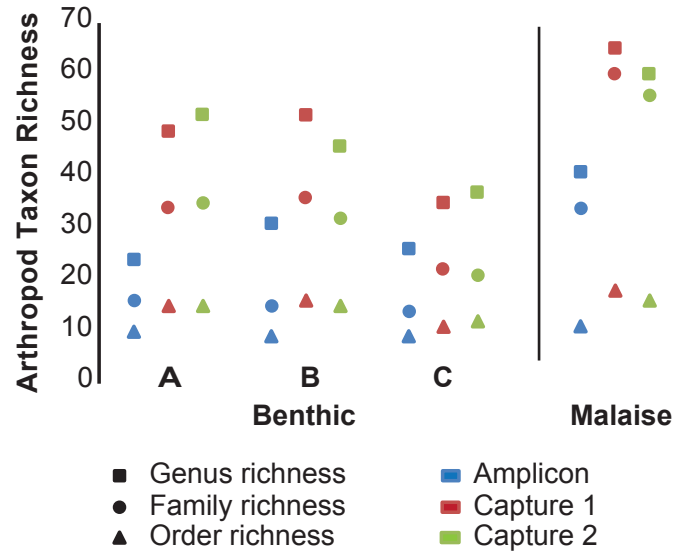
515 Table 1. Summary of number of DNA sequences produced, filtered, and used via four different samples

516 and two different methods.

Sample	Process	# raw DNA sequences	# sequences passing quality filter	# <i>COI</i> sequences	% high quality sequences used in <i>COI</i> 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
Benthic B	Amplicon	582,590	341,528	304,548	89.2
	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
Benthic C	Amplicon	510,396	411,362	373,668	90.8
	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
Malaise	Amplicon	2,559,402	2,342,372	2,075,555	88.6
	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

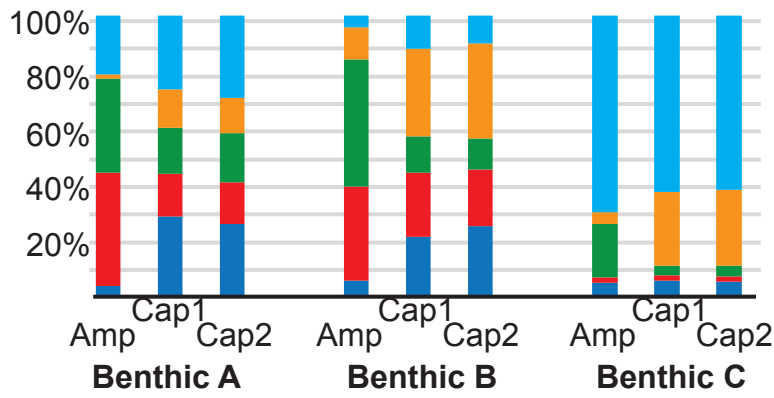
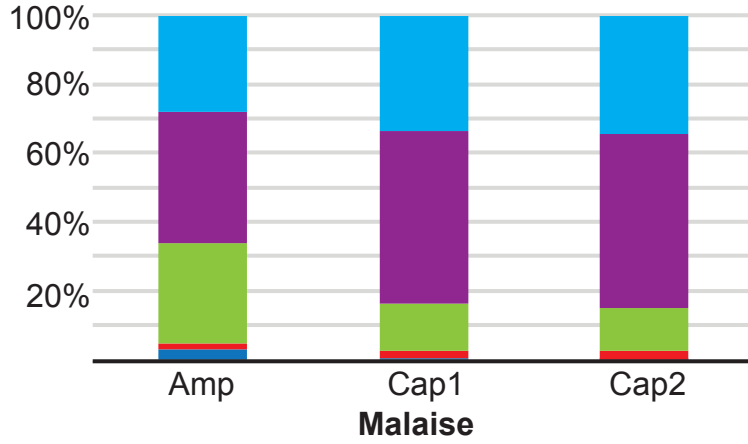
517

518



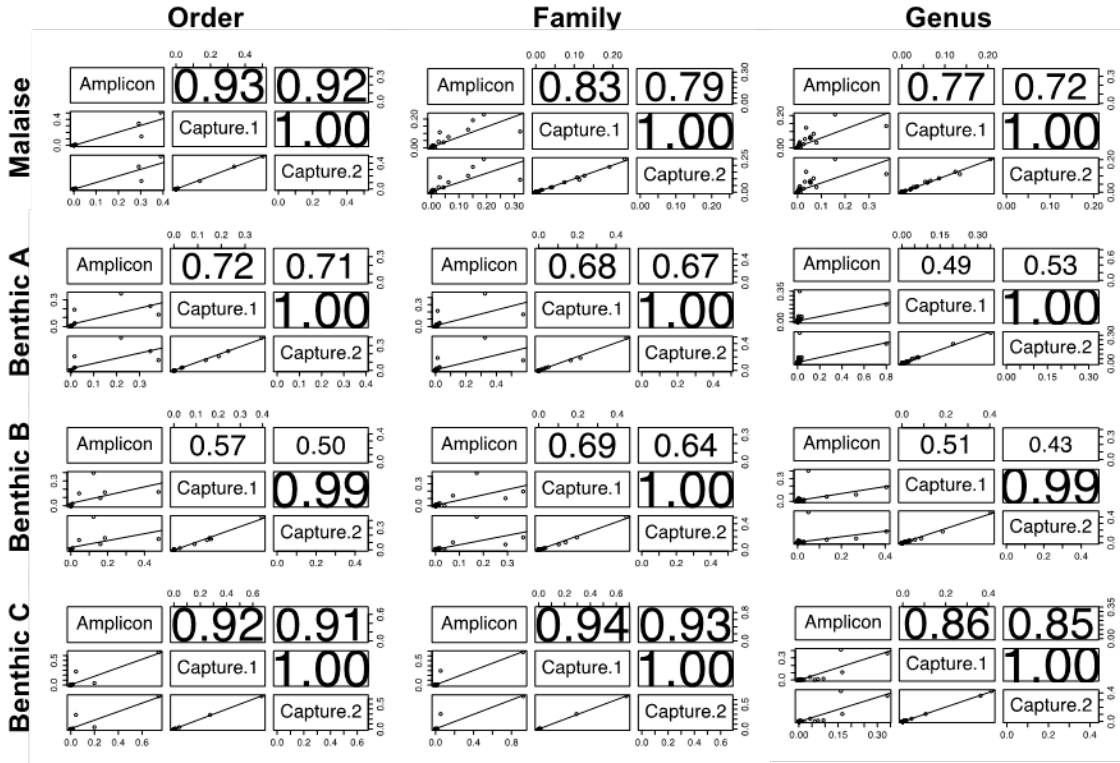
519

520



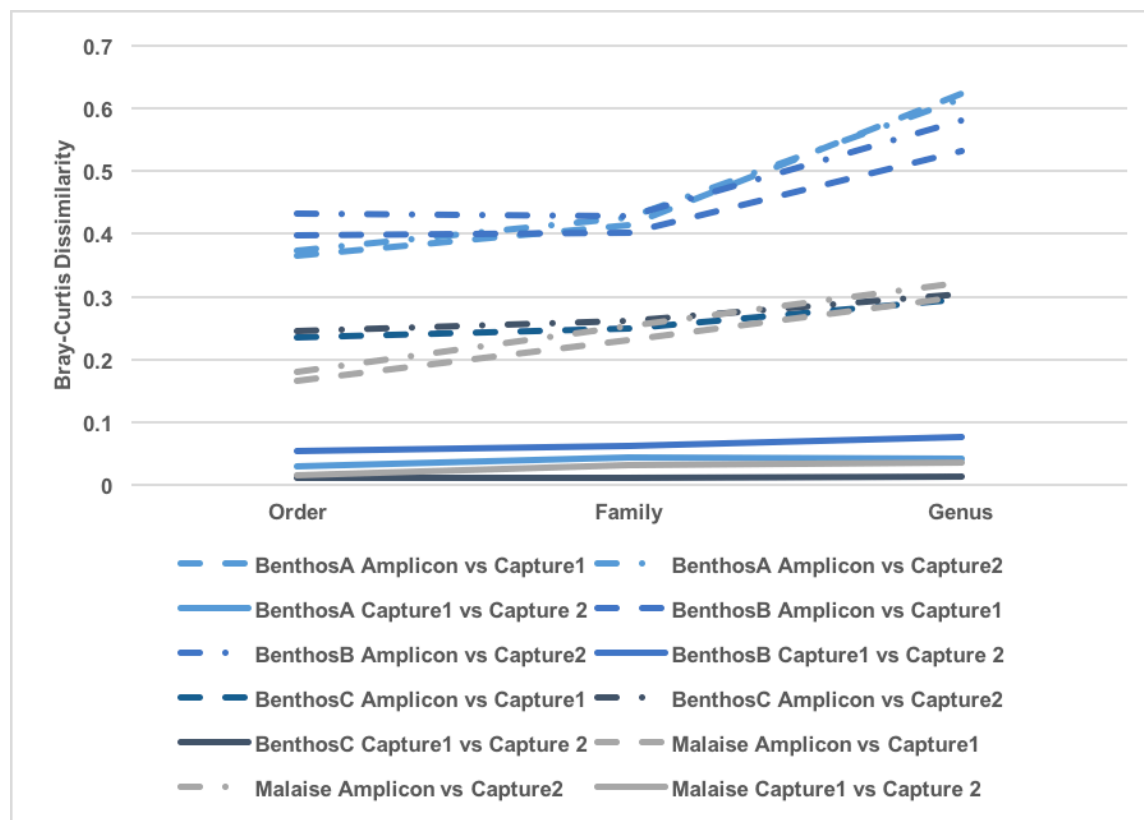
521

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	6	25	497	276,365
STREPSIPTERA				
THYSANOPTERA				
DERMAPTERA	1	1	34	21,751
ISOPODA	28	92	1193	702,654
EPHEMEROPTERA	15	47	2895	1,786,921
Blattaria	49	172	1,964	1,196,335
Dermaptera				
Isoptera				
Mecoptera				
Megaloptera				
Neuroptera				
Odonata				
PLECOPTERA	11	42	1,624	1,005,905
TRICHOPTERA	43	196	6,986	4,421,067

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

533 **Table S2.** Proportion of calculated total biomass and proportion of *COI* DNA sequences identified to insect,
534 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%
Arthropods	Araneae	0.090%	0.031%	0.002%	
	Collembola	0.020%		0.013%	0.008%

Decapoda			0.002%	
Diplostraca		0.012%	0.026%	0.027%
Podocopida		0.023%	0.037%	0.031%
Trombidiformes	0.080%			
Haplotaxida		0.026%	0.035%	0.025%
Rickettsiales		2.805%	0.028%	0.027%
Boraginales		0.003%		
Phyllodocida			0.002%	0.002%
Naviculales			0.002%	0.002%
Lumbriculida		0.005%	0.005%	0.006%
Basommatophora		0.003%	0.008%	0.006%
Scale		0.005%	5.000%	50.000%

535

536

537

538

539

540

541

542

543

544