Revealing the Complexities of Metabarcoding with a Diverse Arthropod Mock Community

Thomas W. A. Braukmann[1*], Natalia V. Ivanova[1], Sean W. J. Prosser[1], Vasco Elbrecht[1], Dirk Steinke[1,2], Sujeevan Ratnasingham[1], Jeremy R. deWaard[1,3], Jayme E. Sones[1], Evgeny V. Zakharov[1], and Paul D. N. Hebert[1,2]

[1] Centre for Biodiversity Genomics, University of Guelph, Guelph ON, Canada N1G 2W1

[2] Department of Integrative Biology, University of Guelph, Guelph ON, Canada N1G 2W1

[3] School of Environmental Sciences, Uinversity of Guelph, Guelph, ON, Canada N1G 2W1

* Corresponding author email: tbraukma@uoguelph.ca

Abstract

DNA metabarcoding is an attractive approach for monitoring biodiversity. However, it is subject to biases that often impede detection of all species in a sample. In particular, the proportion of sequences recovered from each species depends on its biomass, mitome copy number, and primer set employed for PCR. To examine these variables, we constructed a mock community of terrestrial arthropods comprised of 374 BINs, a species proxy. We used this community to examine how species recovery was impacted when amplicon pools were constructed in four ways. The first two protocols involved the construction of bulk DNA extracts from different body partitions (Bulk Abdomen, Bulk Leg). The other protocols involved the production of DNA extracts from single legs which were then merged prior to PCR (Composite Leg) or PCR-amplified separately (Single Leg) and then pooled. The amplicon generated by these four treatments were then sequenced on three platforms (Illumina MiSeq, Ion Torrent PGM and Ion Torrent S5). The choice of sequencing platform did not substantially influence species recovery, other variables did. As expected, the best recovery was obtained from the Single Leg treatment, but the Bulk Abdomen produced a more uniform read abundance than the Bulk Leg or Composite Leg samples. Primer choice also influenced species recovery. Our results reveal how variation in protocols can have substantive impacts on perceived diversity unless sequencing coverage is sufficient to reach an asymptote. Although metabarcoding is a powerful approach, further optimization of analytical protocols is crucial to obtain reproducible results and increase its cost-effectiveness.

Introduction

It is generally accepted that we have entered a period of unprecedented global biodiversity loss (Pimm et al. 2014, Vogel 2017). Halting it will require the capacity to quantify shifts in species composition rapidly and on a far larger scale than ever before so we can better understand and manage ecosystems (Ji et al. 2013; Cristescu 2014; Moriniere et al. 2016, Waldron et al 2017). As arthropods account for the majority of terrestrial biodiversity (Medeiros et al. 2013), they are an obvious target for bio-surveillance. Although they are easily collected in large numbers (Russo et al. 2011), the subsequent processing and identification of specimens has traditionally been a barrier to large-scale monitoring programs (Bassett et al. 2012). DNA barcoding, the use of short standardized gene regions to discriminate species, breaks this barrier by enabling non-taxonomists to identify specimens once a reference sequence library is established (Hebert et al. 2003; Hebert and Gregory 2005).

DNA barcode studies initially focused on developing the analytical protocols required to construct a specimen-based reference library (Hebert et al. 2003; Hebert et al. 2004). Although improved protocols have reduced costs, making it possible to analyze millions of single specimens (Hajibabaei et al. 2005; Ivanova et al. 2006; Hebert et al. 2018), they are still too expensive to support large-scale bio-monitoring programs. However, by coupling a DNA barcode reference library with the analytical capacity of high-throughput sequencers (HTS), DNA metabarcoding provides a path to rapid, low-cost assessments of species composition (Hajibabaei et al. 2011; Yu et al. 2012; Brandon-Mong et al. 2015; Moriniere et al. 2016). It achieves this goal by the amplification and sequence characterization of the barcode region from bulk DNA extracts which can then be assigned to operational taxonomic units (OTUs) that

43 can be queried against reference sequences to ascertain their source species (see Cristescu
44 2014). Studies have now employed this approach to assess species composition in communities
45 of aquatic and terrestrial arthropods (Ji et al. 2013; Beng et al. 2016; Elbrecht et al. 2017B),
46 vertebrates (Sato et al. 2017), diatoms (Vasselon et al. 2017), and fungi (Bellemain et al. 2012;
47 Aas et al. 2017; Tedersoo et al. 2018). Metabarcoding can reveal more species than
48 morphological approaches while requiring far less time (Ji et al. 2013; Yu et al. 2012; Vivien et
49 al. 2016; Brandon-Mong et al. 2015; Elbrecht et al. 2017A; Hebert et al. 2018; Shokralla et al.
50 2015; Elbrecht et al 2017B).

51 Despite the advantages of metabarcoding, several factors often complicate the recovery
52 of all species in a sample. Firstly, DNA templates derived from the species in a mixed sample are
53 often differentially amplified (Elbrecht and Leese 2015; Pinol et al. 2015; Elbrecht and Leese
54 2017; Tedersoo et al. 2018). Such bias can arise from either the DNA polymerase (Nichols et al.
55 2018; Pan et al. 2014, Dabney and Meyer 2012) or the primers (Clarke et al. 2014) employed for
56 PCR. Polymerase bias involves the differential amplification of templates as a result of their
57 variation in sequence motifs, GC content, or length (Nichols et al. 2018; Pan et al. 2014, Dabney
58 and Meyer 2012). Primer bias is due to either varying levels of primer mismatch or template
59 degradation (Clarke et al. 2014; Elbrecht and Leese 2015). The impact of primer mismatches
60 can often be reduced by either lowering annealing temperatures or by raising the degeneracy
61 of the primers (Clarke et al. 2014; Elbrecht and Leese 2017). However, these 'solutions' have a
62 downside; they often increase the amplification of non-target sequences such as bacterial
63 endosymbionts or mitochondrial pseudogenes, which is especially problematic for eDNA
64 samples (Smith et al. 2012; Song et al. 2008; Macher et al. 2018).

The capacity of metabarcoding to recover all species in a bulk sample is further complicated because the component species typically vary by several orders of magnitude in mass and hence in copy numbers of the target template. Unless other factors intervene, this variation in template number means that large-bodied species are more likely to be recovered (Brandon-Mong et al. 2015). Because of this effect (in addition to primer bias), efforts to infer species abundance from the read counts obtained in metabarcoding studies are at best weak (Elbrecht and Leese 2015; Pinol et al. 2014). Correction factors can improve such estimates (Thomas et al. 2015; Vasselon et al 2017), but any method based on the analysis of bulk DNA extracts will fail to accurately estimate species abundance.

In addition to factors complicating the recovery of sequences from all species in a bulk sample, sequence variation introduced during PCR, library preparation, and sequencing can make it difficult to assign sequences to their source species (Tedersoo et al. 2018). PCR error can be reduced by the use of high-fidelity polymerases (Lee et al. 2016; Potapov et al. 2017), but it is more difficult to escape complexities introduced by sequencing error because all second-generation sequencers have error rates (e.g. 1–2%) that are high enough to complicate the discrimination of closely-related species. Third-generation sequencers, such as Pacific Biosciences Sequel (e.g. Hebert et al. 2018), can produce much lower error rates, but they currently generate too few reads (circa 0.2 million/run) to reveal all species in a taxonomically diverse sample (Tedersoo et al. 2018). As a consequence, despite their high error rates, second-generation platforms (Illumina, Ion Torrent) are commonly used for metabarcoding as they produce many millions of reads per run (Mardis et al. 2013, Cristescu 2014). Illumina sequencers generate more reads (20–250 million/run) with lower error rates than Ion Torrent

platforms, but the latter instruments can deliver longer reads more rapidly (Mardis et al. 2013). It is unclear how severely the choice of HTS platform affects species recovery as their performance has rarely been compared in eukaryotes (Divolli et al. 2018). However, work on microbial communities found general agreement between platforms although Ion Torrent reads were lower quality and more length variable than those from Illumina (Salipante et al. 2014; Tessler et al. 2017).

To examine factors influencing the success in recovering species through metabarcode analysis, we assembled a mock community that included a single representative of 374 species. We subsequently used this community to examine the impacts of DNA source, extraction method, PCR protocol, target template, and sequencing platform on species recovery. Specifically, we compared results obtained by analyzing four amplicon pools on three sequencing platforms (Illumina MiSeq, Ion Torrent PGM, Ion Torrent S5). Two of these amplicon pools derived from the PCR of bulk DNA extracts (abdomen, leg) to test the impact of tissue type. Two other amplicon pools derived from DNA extracts of single legs that were analyzed by pooling prior to or after PCR. Finally, we examined species recovery for two amplicons of differing length on the S5. The overall analytical approach involved evaluation of the relationship between read depth and species recovery for these treatment variables.

Material and Methods

Assembly of Mock Community

We began the assembly of a mock community by obtaining COI sequences from 3,044 insects collected in Malaise traps deployed near Cambridge, Ontario, Canada. A DNA extract

was prepared from a single leg from each specimen employing a membrane-based protocol (Ivanova et al. 2006). The 658 bp barcode region of COI was amplified and then Sanger sequenced, to link a haplotype to each individual specimen. Amplicons were generated using a primer cocktail of C_LepFolF and C_LepFolR (Hernández-Triana LM et al. 2014) with initial denaturation at 94 °C for 2 min followed by 5 cycles of denaturation for 40 s at 94 °C, annealing for 40 s at 45 °C and extension for 1 min at 72 °C; then 35 cycles of denaturation for 40 s at 94 °C with annealing for 40 s at 51 °C and extension for 1 min at 72 °C; and a final extension for 5 min at 72 °C (Ivanova et al. 2006; Hebert et al. 2018). Most reactions generated a 709 bp amplicon comprised of 658 bp of COI plus 51 bp of forward and reverse primers. A few amplicons were slightly shorter as a result of deletions in the COI gene. Unpurified PCR products were diluted 1:4 with ddH_2O before 2 µl was used as the template for a cycle sequencing reaction. All products were sequenced following standard procedures on an ABI 3730xl DNA Analyzer (Applied Biosystems, Foster City, California, USA).

Because some specimens could not be identified to a species level, we employed the Barcode Index Number (BIN) system which examines patterns of sequence variation at COI to assign each specimen to a persistent species proxy (Ratnasingham and Hebert 2013). The overall analysis provided sequence records for 803 BINs. From this total, we selected 374 BINs that met two criteria. They showed >2% COI sequence divergence from their nearest-neighbor, and they were taxonomically distant. The resulting mock community included representatives of 10 orders and 104 insect families. Supplemental Table 1 provides a list of the taxa included in the mock community as well as details on vouchers, their body size (as estimated by abdominal mass), and the GC content of their COI. Following selection of the specimens for inclusion in the

mock community, fresh DNA extracts were made following the four protocols described below, and the amplicon pools generated from them were subsequently analyzed on three sequencing platforms.

Experimental design for metabarcode analysis

Species recovery was compared for amplicon pools generated by four protocols (Figure 1). Two involved the analysis of amplicons generated from bulk DNA extracts derived from two tissues (Bulk Abdomen, Bulk Leg). The other two treatments involved the initial extraction of DNA from individual legs. The resultant DNA extracts were either pooled prior to PCR to create the Composite Leg treatment or separately amplified and subsequently pooled to create the Single Leg treatment (Figure 1). Although the initial design called for the same specimens to be included in each mock community, this was not possible. The Composite Leg and Single Leg treatments did include the selected array of 374 BINs. However, five of their source specimens either lacked an abdomen or another leg for inclusion in the Bulk Abdomen or Bulk Leg treatments. As a result, five BINs, generally belonging to the same order as the excluded ones, were employed as replacements to maintain 374 BINs per treatment (BOLD:AAA2323, BOLD:AAA2632, BOLD:AAF4234, BOLD:AAP6354; BOLD:ABV1240). Further details on the treatments are available at the following DOIs: Bulk Abdomen and Bulk Leg: dx.doi.org/10.5883/DS-NGS375A; Composite Leg and Single Leg: dx.doi.org/10.5883/DS-NGS375B.

Bulk DNA extractions and PCR

151 DNA extracts for the two bulk samples (Bulk Abdomen, Bulk Leg) were generated with a
152 modified membrane-based protocol (Ivanova et al. 2006). Specifically, the bulk abdomens
153 (combined mass = 1,062.8 mg) and bulk legs (combined mass = 30.9 mg) were lysed overnight
154 in the same relative volume of insect lysis buffer (51.6 ml and 1.5 ml respectively), with 10
155 mg/ml of Proteinase K (Invitrogen). Following lysis, a 100 µl aliquot of each extract was mixed
156 with 200 µl of binding mix and transferred to an EconoSpin® column (Epoch Life Sciences)
157 before centrifugation at 5000 g for 2 min. The DNA extracts were then purified with three wash
158 steps. The first wash employed 300 µl of protein wash buffer before centrifugation at 5000 g for
159 2 min. Columns were then washed twice with 600 µl of wash buffer before being centrifuged at
160 5000 g for 4 minutes. Columns were transferred to clean tubes and spun dry at 10000 g for 4
161 min to remove any leftover buffer, then transferred to clean collection tubes and incubated for
162 30 min at 56°C to dry the membrane. DNA was subsequently eluted by adding 50 µl of 10 mM
163 Tris-HCl pH 8.0 followed by centrifugation at 10000 g for 5 min. All DNA extracts were
164 normalized to 3 ng/µl prior to PCR. All PCR reactions were composed of 5% trehalose (Fluka
165 Analytical), 1× Platinum Taq reaction buffer (Invitrogen), 2.5 mM $MgCl_2$ (Invitrogen), 0.1 µM of
166 each primer (Integrated DNA Technologies), 50 µM of each dNTP (KAPA Biosystems), 0.15 units
167 of Platinum Taq (Invitrogen), 1 µl of template, and HyClone® ultra-pure water (Thermo
168 Scientific) for a final volume of 6 µl.

169 Construction of HTS libraries

170 Two rounds of PCR were used to generate the amplicon libraries destined for sequence
171 characterization on the three platforms. Most first-round reactions employed a primer cocktail
172 targeting a 407 and 421 bp region of COI and will be referred to as the 407 bp amplicon

throughout the manuscript. The 407 bp region was amplified using MLepF1 (Hebert et al. 2004) and the 421 bp region with RonMWasp (Smith et al. 2012) as forward primers and LepR1 (Hebert et al. 2004) and HCO2198 (Folmer et al. 1994) as reverse primers (Table S2). An alternate first-round PCR targeted a 463 bp amplicon of COI; it was generated with another forward primer — AncientLepF3 (Prosser et al. 2016) (Table S2). In addition, the primers employed to generate amplicons for MiSeq analysis contained a Nextera transposase adapter (Table S2). All first-round PCRs were run under the same conditions with initial denaturation of 94 °C for 2 min, followed by 20 cycles of denaturation at 94°C for 40 s, annealing at 51°C for 1 min and extension at 72 °C for 1 min, with a final extension at 72°C of 5 min. Three technical PCR replicates were generated for three of the treatments – Bulk Abdomen, Bulk Leg, and Composite Leg.

Prior to the second PCR, first-round products were diluted 2x with dd H_20. Fusion primers were used to attach platform-specific unique molecular identifiers (UMIs) along with sequencing adaptors for Ion Torrent libraries and a flow cell bind for the MiSeq libraries (Table S2). The second PCR was run under the same conditions as the first round for reactions slated for analysis on the Ion Torrent platforms, but the samples for Illumina were amplified following manufacturer's specifications with initial denaturation at 94°C for 2 min, then 20 cycles of denaturation at 94°C for 40 with annealing at 61°C for 1 min and extension at 72 °C for 1 min, followed by a final extension at 72°C of 5 min. Supplementary Table S2 provides all primer sequences and details on samples indexing.

For each platform, the UMI-labelled reaction products were pooled prior to sequencing. The sequence libraries for the S5 were prepared on an Ion Chef™ (Thermo Fisher Scientific)

following manufacturer's instructions while those for the PGM were prepared using the Ion PGM™ Hi-Q™ View OT2 400 Kit and the Ion PGM™ Hi-Q™ Sequencing Kit (Thermo Fisher Scientific). The PGM libraries were sequenced on a 318 v2 chip while the S5 libraries were sequenced on a 530 chip at the Canadian Centre for DNA Barcoding. Illumina libraries were sequenced (paired end) using the 300 bp reagent kit v3 on an Illumina MiSeq in the Genomics Facility of the Advanced Analysis Centre at the University of Guelph.

Bioinformatics and analysis

Prior to uploading MiSeq runs, read libraries were paired using the QIIME (Caporaso et al. 2010) pair join script (join_paired_ends.py) with a minimum overlap of 20 bp and a maximum difference of 10%. All read libraries were uploaded to mBRAVE (http://mbrave.net/), an online platform for analyzing and visualizing metabarcoding data. The Quality Value (QV) of each sequence was evaluated and all records failing to meet any one of three quality standards were discarded: 1) mean QV<20; 2) >25% of bp with QV<20; 3) >5% of bp with QV<10. All reads were trimmed to be either 407 bp or 463 bp. Retained sequences were viewed as a match to a BIN in the custom Sanger reference library if their distance was <3% to any reference. All reads not matching a reference sequence were clustered at an OTU threshold of 2%. These OTUs were then queried against four other system libraries (insects, non-insect arthropods, non-arthropod invertebrates, bacteria). Standard analytical parameters were used for all treatments and sequencing platforms. All raw data is available in NCBI's Short Read Archive (SRP158933). The three replicates for the Bulk Abdomen, Bulk Leg, and Composite Leg treatments were pooled.

OTU tables for each run were merged in R ver 3.4.4 (R Core Team 2018). To compare BIN accumulation across all samples, we randomly subsampled each run at different read depths

for 10,000 replicates using a custom script (Supplemental material). To measure the BIN accumulation for each treatment, we compared the slopes between sequential points at eight read counts (10^2, 10^3, $10^{3.5}$ 10^4, $10^{4.5}$, 10^5, $10^{5.5}$,10^6). Sequential points with a slope of less than 0.01 were viewed as indicating that an asymptote had been achieved.

To compare the different treatments and sequencing platforms, we reduced the data set to the 369 shared BINs. Read distributions were visualized using the JAMP v0.44 package (https://github.com/VascoElbrecht/JAMP) in R to produce a heat map using the "OTU_heatmap" function. Read distributions across BINs were compared using density graphs generated with ggplot2 v2.2.1 (Wickham 2009). The relative abundances of all BINs comprising greater than 0.01% of the overall reads were used to estimate Simpson's index, Pielou's mean evenness, and Renyi's entropy implemented in the R package vegan v2.5-1 (Oksanen et al. 2018). Compositional dissimilarity between replicates and treatments were examined using a dendrogram based on the Bray-Curtis index and calculated with vegan. The values for the Bray-Curtis index were also used to generate a non-metric multidimensional scaling (NMDS) with vegan.

The relationships between read counts and body size, as measured by abdominal mass, and the read count and GC content of the COI amplicon were examined using Kendall Tau correlations in R ver 3.4.4 (R Core Team 2018). An analysis of similarity (ANOSIM) with 999 permutations was used to compare species recovery among treatment types, and sequencing platforms and between the two amplicons with the R package vegan v2.5-1 (Oksanen et al. 2018). All custom scripts are available as supplementary materials.

238 The relationships between the read count for each BIN and primer mismatches were
239 investigated for the 407 bp and 463 bp amplicons. The number of mismatches were quantified
240 by counting the number of nucleotide substitutions between the primer sequence and the
241 template DNA for each BIN. Information on the DNA sequence for the forward primer binding
242 sites was available from the Sanger reads for all 369 BINs. Calculation of mismatches was
243 straightforward for the 463 bp amplicon as it involved a single forward primer. As the 407 bp
244 amplicon was generated with two different forward primers, total mismatches were quantified
245 based upon the forward primer with the best match to the template for each BIN. The same
246 two reverse primers were employed to generate the 407 bp and 463 bp amplicons, requiring a
247 similar approach, but with the complication that DNA sequence information for template DNA
248 was not available from the Sanger sequence (as it was based on amplicons generated with the
249 same reverse primers). As a result, a new reverse primer was employed to extend the sequence
250 in a 3' direction, an approach which delivered the desired sequence information for 203 of the
251 369 BINs. As a consequence, it was possible to examine the relationship between read counts
252 and the number of mismatches between template and forward primer for all 369 BINs and for
253 the total mismatch count for the forward and reverse primers for the 203 BINs with template
254 sequences for both regions.

255 **Results**

256 Run quality

257 We first compared the output and quality of the reads from the HTS platforms. The S5
258 and MiSeq generated a similar number of reads (~ 1 million per replicate), while the PGM
259 generated substantially fewer. About 60-65% of the MiSeq reads were filtered during merging

of the paired-end reads, but subsequent filtering was minimal (< 1%). The PGM and S5 encountered a similar loss of reads as 45–50% of the raw reads were filtered (Table 1). The MiSeq reads showed more length consistency and higher quality than those from both Ion platforms, reflecting their near consistent QV versus the decline towards the 3' end of the PGM and S5 reads (Figure S1).

Read depth

Rarefaction curves were calculated for each of the four treatments and their technical replicates to ascertain if read depths were sufficient to recover all BINs (Figure 2). Although BIN recovery was high in all cases, the Single Leg treatment reached it with far fewer reads of the 407 bp amplicon than the other treatments ($10^{4\text{-}4.5}$ versus $10^{4.5\text{-}5}$ –Table S3). There was evidence of variation among platforms as the PGM needed more reads to achieve an asymptote than the S5 or MiSeq. BIN accumulation curves for the other treatments were similar, but the Bulk Abdomen showed a small, but consistent outperformance of the Bulk Leg and Composite Leg treatments. The target amplicon also had a substantial impact as just $10^{3.5}$ reads of the 463 bp amplicon were required for the Single Leg treatment to reach its asymptote (Table S3). The technical replicates showed little divergence on all platforms; they had similar BIN recovery, similar mean read counts per BIN, and similar coefficients of variation (Table S1). Pielou's evenness, Simpson's Index, Inverse Simpson's Index, Renyi's diversity, and Shannon Indices were also similar across treatments on all platforms (Table 2; Figure S2). Finally, density plots were similar among technical replicates for all treatments and platforms suggesting that different HTS platforms produced similar results for the different treatments (Figure 1; Figure S3).

282

283

284 BIN recovery

285 When the criterion for BIN recovery was set at one or more reads, all platforms recovered
286 >98% of the BINs but only the Single Leg treatment recovered all of them (Figure 4). Differences
287 in recovery success among treatments were greater when the criterion for recovery was set at
288 >0.01% of the reads. Under this criterion, the Single Leg treatment recovered >92.5% of the
289 BINs versus 83-89% for the Bulk Abdomen treatment and about 76-83% for the Composite Leg
290 and Bulk Leg treatments (Table 1). The greater evenness in read count for the Single Leg
291 treatment was striking; it led to lower coefficients of variation, higher diversity indices, and
292 Pielou's evenness (Table 2; Figure 4; Figure S2). Density plots of read abundance also
293 demonstrated much higher evenness for the Single Leg treatment, especially for the 407 bp
294 amplicon on the MiSeq and for the 463 bp amplicon on the S5 (Figure S3). These differences
295 were also reflected in BIN recovery, Pielou's evenness, and diversity indices (Table 2; Table S2).

296 BIN abundances

297 Because a single specimen of each BIN was included in the mock community, the
298 proportion of sequences from each should, in the absence of bias, be similar across sequencing
299 platforms, amplicons, and treatments. In practice, the relative abundances of the BINs varied
300 markedly. A single-link dendrogram based on Bray-Curtis dissimilarity values indicated that
301 samples clustered first by treatment, next by amplicon length, and finally by sequencing
302 platform (Figure 3A). An analysis of similarity using Bray-Curtis distances affirmed significant

303 differences in BIN abundances by treatment type (p = 0.001, R = 1), amplicon length (p = 0.027,
304 R = 0.17), but not by sequencing platform (p = 0.13, R = 0.037) (Figure 3B; Figure S5).

305 Primer mismatches and read count

306 Examination of the relationship between the read count for each of the 369 BINs and its
307 number of mismatches from the forward primer revealed a strong negative relationship. BINs
308 with a high mismatch count were typically represented by few reads. For example, very few
309 sequences were recovered from the only BIN belonging to the order Dermaptera and this was
310 associated with a high Mismatch Index from the forward primers for the 407 bp and 463 bp
311 amplicons. BIN recovery was substantially higher for the 463 bp amplicon than for the 407 bp
312 amplicon (Table 2; Figure S2). Its superior performance was associated with the fact that its
313 forward primer showed a better match to the DNA extracts. Just 18 BINs had >3 mismatches for
314 the forward primer used to amplify the 463 bp amplicon versus 62 for the forward primer for
315 the 407 bp amplicon (Table S1 and Table S4). The impact of these mismatches was clear; mean
316 read depth and relative abundance of BINs declined after two mismatches for the Bulk
317 Abdomen, Bulk Leg, and Composite Leg treatments and after four mismatches for the Single
318 Leg. When we examined the impact of forward and reverse primer mismatches for a subset of
319 203 BINs, mean read depth and relative abundance showed a significant decline after four
320 mismatches for the Bulk Abdomen, Bulk Leg, and Composite Leg treatments and after seven for
321 the Single Leg. Kruskal-Wallis tests showed that read depth declined significantly with an
322 increasing number of primer mismatches for the forward primers for both the 463 bp and 407
323 bp amplicons ($p < 0.0001$) and for the summed primer mismatches (5' + 3') for the subset of
324 203 BINs ($p < 0.0001$).

325

326 Impacts of biomass and nucleotide composition on read count

327 Other factors were also responsible for some of the variation in read counts among BINs.
328 There was, for example, a weak negative correlation between the GC content of an amplicon
329 and its read count, although all values were low ($r^2 < 0.1$) excepting the Single Leg treatment on
330 the MiSeq ($r^2 = 0.32$). A weak positive correlation ($r^2 = 0.24 – 0.28$) was also apparent between
331 the abdominal mass of a BIN and its read count on all platforms.

332 Non-Target Sequences

333 Every run recovered some sequences with substantial sequence divergence from the
334 Sanger reference library (Table 1). The incidence of these non-target sequences for the 407 bp
335 amplicon was slightly lower (4 – 6%) on the PGM and S5 platforms than on the MiSeq (8 – 10%).
336 Interestingly, the 463 bp amplicon had substantially more non-target reads (15 – 17%). Many of
337 the non-target reads were chimeras (8 – 81 %; Table 1). After their exclusion, most sequences
338 assigned to an OTU did not find a match to a sequence in the supplemental libraries. Of those
339 that did, similar numbers matched to a known bacterial sequence or to another arthropod.

340 Taxonomic Bias

341 There was also evidence of taxonomic bias in the read counts for BINs between the two
342 amplicons. For example, Orthoptera, Lepidoptera, and Diptera dominated the 407 bp
343 sequences from the Bulk Abdomen and Bulk Leg treatments while Lepidoptera, Mecoptera,
344 Diptera, and Coleoptera dominated the 463 bp amplicon (Table S5). The 463 bp amplicon also
345 showed more variation among treatments than the 407 bp amplicon (Table S5). Few sequences

346 were recovered for Dermaptera, especially for the 407 bp amplicon, likely reflecting its
347 possession of 5 mismatches from the forward primer (Table S1). Among the bulk samples,
348 relative abundance differed among treatments. For example, the relative abundance of
349 Lepidoptera and Mecoptera was lower, while Diptera and Orthoptera were higher in the
350 Composite Leg than in the Bulk Leg and Bulk Abdomen treatments. The proportion of read
351 counts for Trichoptera showed particularly large variation, being 5–25X higher for the Bulk Leg
352 than the Bulk Abdomen and Composite Leg treatments across all platforms and for both
353 amplicons.

354 **Discussion**

355 Metabarcoding is a powerful tool for characterizing biodiversity patterns (Cristescu 2014),
356 but data interpretation is complicated by several factors. PCR amplification bias and variation in
357 the copy number of template DNA from the source specimens not only make it impossible to
358 estimate abundances, but can impede the recovery of all species (Yu et al. 2012; Li et al. 2013;
359 Beng et al. 2016). Although prior studies have revealed these complexities, there has been
360 limited evaluation of the strength of their influence on interpretations of taxon diversity. To
361 address this gap, the present study has examined the impacts of diverse factors including
362 source DNA, PCR primers, sequencing platform, and sequencing depth on species recovery from
363 a mock community of arthropods.

364 Sequencing Depth

365 Variation in sequencing depth (read count) can directly impact taxon recovery and hence
366 perceived diversity patterns. This is particularly true for comparisons among datasets with

differing read counts (Leray et al. 2015; Leray and Knowlton 2017; Elbrecht et al. 2017). Low read depth typically means that some rare species with little biomass in a community will be overlooked, leading to underestimation of alpha diversity. When taxon counts are incomplete, comparisons among sites are also compromised (Bellemain et al. 2012; Sickle et al. 2015; Yamamoto et al. 2017), producing overestimates of beta diversity (Sickle et al. 2015). Both rarefaction and species accumulation curves are valuable for ascertaining if sequencing depth has been adequate. When sequences have been recovered from all species, the slope of the rarefaction curve is zero, providing a simple criterion for gauging the adequacy of read coverage (Lanzen et al. 2017). In real world situations, the true species count is unknown and increased sampling effort nearly always raises the species count, meaning there is no asymptote. Although taxon richness was fixed in our study, we employed a slope of 0.01 as the criterion for assessing when taxon diversity had achieved an asymptote as this approach can be employed in studies on natural communities. Under this criterion, there were substantial differences among the four treatments and between primer sets.

Analysis of BIN accumulation curves indicated that read depth was sufficient for all four treatments to achieve a slope of 0.01. However, the Single Leg treatment reached this value with much lower read depth than the bulk samples due to its relative protection from the impacts of PCR bias (Nichols et al. 2017; Pan et al. 2014, Dabney and Meyer 2012; Elbrecht and Leese 2015). Interestingly, the other three treatments showed similar BIN accumulation curves on all three sequence platforms, suggesting shared factors are constraining BIN recovery.

Sequencing Platforms

388 The three sequencing platforms generated similar estimates of BIN diversity. However,
389 results from the MiSeq had advantages over those from the PGM and S5. Its paired end
390 protocol consistently recovered sequences for the full 407 bp amplicon, which can be a
391 requirement for clustering algorithms and is also useful for haplotype analysis (Elbrecht et al.
392 2018), while those from the other platforms were often truncated. Its reads also possessed
393 fewer indels than those from PGM and S5. Finally, the MiSeq reads had consistently higher QV
394 across the amplicon. Because these factors simplified data analysis (Mardis et al. 2013; Edgar et
395 al. 2013) and sequencing costs were similar, the MiSeq is currently the best platform for
396 metabarcoding (Mardis 2013). However, because it cannot analyze amplicons longer than 500
397 bp, third-generation sequencing platforms (Tedersoo et al. 2018; Hebert et al. 2018; Wilkinson
398 et al. 2017) will be an attractive option if their current limitations on read number (Pacific
399 Biosciences) and quality (Oxford Nanopore) are overcome.

400 Impacts of Analytical Protocols

401 Our four treatments made it possible to compare the impact of targeting different tissues,
402 employing different DNA extraction regimes, and using different PCR protocols. Despite their
403 similar tissue input and DNA extraction regime, the Single Leg treatment achieved asymptotic
404 diversity much more rapidly than the Composite Leg treatment, indicating how separate PCR
405 reactions reduce amplification bias. By contrast, BIN accumulation curves for the Composite Leg
406 treatment were similar to those for the Bulk Leg and Bulk Abdomen, indicating that DNA
407 extraction was equally effective whether carried out on single specimens or on bulk samples.
408 Comparison of the results for the bulk/composite samples did reveal more nuanced differences
409 as they showed the number of reads for particular taxa varied among these three treatments

410 despite similar BIN recovery profiles. These differences likely stem from differential
411 leg/abdomen mass ratios among species which varied the mitochondrial copy number for the
412 component species among treatments. Certainly, mitochondrial copy number varies among
413 tissues and among species (Veltri et al. 1990; Cole 2016). Future efforts to explore this
414 relationship and its importance to metabarcoding studies must quantify copy number
415 differences between tissues and species. In the absence of such information, copy number bias
416 can be reduced by partitioning the specimens in a bulk sample into a few size fractions
417 (Elbrecht et al. 2017A; Vivien et al. 2016).

418 Variation in read counts for the taxa in any bulk sample are also influenced by primer
419 bias. In this study, these effects were indicated by the linkage between the read counts for each
420 BIN and the number of mismatches between its COI sequence and the primer set. Although
421 degenerate primers (Yu et al. 2012; Elbrecht and Leese 2017; Moriniere et al. 2016) and
422 improved primer sets (Clarke et al. 2014; Leray and Knowlton 2017; Elbrecht and Leese 2017)
423 can reduce such bias, they cannot conquer the problem unless all target species possess
424 identical sequences for the primer binding sites, a condition that will never be satisfied for a
425 large assemblage. However, efforts to target highly conserved regions can improve the
426 situation. For example, the BIN accumulation curve reached its asymptote with much lower
427 read coverage for the 463 bp than the 407 bp amplicon. More effort is needed to develop
428 better primer sets by testing their performance on a breadth of taxa and to minimize
429 mismatches by sorting samples into taxonomic groups (Moriniere et al. 2016; Bellemaine et al.
430 2012; Cristescu 2014; Tedersoo et al. 2015). Efforts to minimize specimen bias should include
431 mass and taxonomic sorting to reduce differences in template DNA quantity between

specimens (Elbrecht et al. 2017A; Moriniere et al. 2016; Vivien et al. 2016). Currently, the only means to circumvent biases is to process specimens individually (e.g. Single Leg treatment), which is so time consuming and costly that it is difficult to implement for large biodiversity surveys (Ji. et al 2013).

BIN recovery

Most of the 369 BINs in the template pools were recovered in all four treatments. However, this outcome shifts if recovery success is defined as those BINs comprising at least 0.01% of the read count, a criterion often employed to minimize the impacts of sequencing errors, chimeras, and contaminants (Leray and Knowlton 2017). Under this criterion, BIN recovery was high (> 92.5%) for the Single Leg treatment, but substantially lower (76% – 89%) for the other three. Interestingly, the Bulk Abdomen treatment showed higher BIN recovery than the Bulk Leg and Composite Leg treatments, perhaps reflecting more similar mitochondrial copy numbers among abdomens than legs (Veltri et al. 1990; Cole 2016). As expected, BIN recovery was higher for the 463 bp than the 407 bp amplicon.

False positives and negatives

Up to 26 of all BINs failed to achieve a 0.01% read abundance in the bulk and composite treatments meaning they would often be excluded during analysis, creating false negatives that would underestimate alpha diversity. As in other metabarcoding studies (Vivien et al. 2016; Ficetola et al 2014; Brandon-Mong et al. 2015; Port et al. 2015), false positives were also encountered, likely reflecting eDNA associated with specimens, contamination during sample processing (Port et al. 2015) or NUMTs (Song et al. 2008). Their impact can be reduced by

453 employing curated reference libraries to both recognize sequences derived from known species
454 and to exclude paralogs and pseudogenes (Hebert et al. 2003; Landi et al. 2014; Zimmerman et
455 al. 2014; Braukmann et al. 2017; Bergsten et al. 2014). In addition, the use of negative controls
456 is an effective way to evaluate the incidence of contaminants introduced during sample
457 processing (Port et al. 2015).

458 Conclusions

459 This study has examined the complexities encountered in evaluating the species
460 composition of a mock community comprised of 374 species of arthropods. Some results were
461 reassuring. Similar measures of overall taxon diversity were obtained from different sequencing
462 platforms, from different tissues, from different DNA extraction protocols, and from different
463 PCR primers. However, this congruence needs to be qualified. Firstly, this study has shown that
464 the analytical effort required to obtain comprehensive information on species composition
465 through the analysis of bulk samples is far higher than that required to obtain the same
466 information through specimen-based protocols. For example, the Sanger sequencing of 369
467 specimens would deliver precise information on the species composition and abundance of
468 each sample. By comparison, the recovery of a complete species list by sequencing a merged
469 pool of amplicons following separate extraction and PCR (e.g. Single Leg treatment) required
470 60,000 reads. When samples were pooled prior to PCR, the recovery of a near-complete species
471 list required at least 250,000 reads, and the proportion of the taxa recovered were represented
472 by so few reads (<0.01%) that they could be excluded during data cleansing. It is worth
473 emphasizing that many natural communities present greater analytical complexity than the
474 mock assemblage examined in this study – they include more species and the abundances of

475 these species show great variation. Given these complications, it is clear that community
476 characterization through metabarcoding will often require both intensive sequencing effort and
477 improved approaches to discriminate between those low frequency reads that are spurious and
478 those that derive from rare species.

479 **Acknowledgements**

480 We thank the lab and collections staff at the Centre for Biodiversity Genomics for aiding
481 in acquiring and processing the specimens in this study and Suzanne Bateson for aid with
482 graphics. We are also grateful to Jenna Quinn and other staff at the rare Charitable Research
483 Reserve for facilitating the collection of specimens. This study was enabled by support from the
484 Ontario Ministry of Research, Innovation and Science and from the Canada First Research
485 Excellence Fund to the 'Food From Thought' research program.

References

Aas AB, Davey ML, and H Kauserud. (2017). ITS all right mama: investigating the formation of chimeric sequences in the ITS2 region by DNA metabarcoding analyses of fungal mock communities of different complexities. *Molecular Ecology Resources,* 17, 730-741 DOI: 10.1111/1755-0998.12622

Bassett Y, Cizek L, Cuénoud P, Didham RK, Guilhaumon F, Missa O, Novotny V, Ødegard F, Roslin T, Schmidt J, Tishechkin AK, Winchester NN, Roubik DW, Aberlenc H-P, Bail J, Barrios H, Bridle JR, Castaño-Meneses G, Corbara B, Curletti G, da Rocha WD, De Bakker D, Delabie JHC, Dejean A, Fagan LL, Floren A, Kitching RL, Medianero E, Miller SE, de Oliveira EG, Orivel J, Pollet M, Rapp M, Ribeiro SP, Roisin Y, Schmidt JB, Sørensen L, and M Leponce. (2012). Arthropod diversity in a tropical forest. *Science*, 338, 1481-1484

Bellemain E, Davey ML, Kauserud H, Epp LS, Boessonkool S, Coissac E, Geml J, Edwards M, Willerslev E, Gussarova G, Taberlet P, and C Brochmann. (2012). Fungal palaeodiversity revealed using high-throughput metabarcoding of ancient DNA from arctic permafrost. *Environmental Microbiology*, 15, 1176-1189 DOI: 10.1111/1462.29220.12020

Beng KC, Tomlinson KW, Shen XH, Surget-Goba Y, Hughes AC, Corlett RT, and JWF Slik. (2016). The utility of DNA metabarcoding for studying the response of arthropod diversity and composition to land-use change in the tropics. *Scientific Reports*, 6, 24965 DOI: 10.1038/srep24965

Bergsten J, Bilton DT, Fujisawa T, Elliott M, Monaghan MT, Balke M, Hendrich L, Geijer J, Herrmann J, Foster GN, Ribera I, Nilsson AN, Barraclough TG, and AP Vogler. (2014). The effect of geographical scale of sampling on DNA barcoding. *Systematic Biology,* 61, 851-869 DOI: 10.1093/sysbio/sys037

Brandon-Mong GJ, Gan HM, Sing KW, Lee PS, Lim PE, and JJ Wilson. (2015). DNA metabarcoding of insects and allies: an evaluation of primers and pipelines. *Bulletin of Entomological Research,* 105, 717-727 DOI: 10.1017/S0007485315000681

Braukmann TWA, Kuzmina ML, Sills J, Zakharov EV, and PDN Hebert. (2017). Testing the efficacy of DNA barcodes for identifying the vascular plants of Canada. *PLoS ONE,* 12, E0169515 DOI: 10.1371/journal.pone.0169515

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, and R Knight. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7, 335-336 DOI: 10.1038/nmeth.f.303

521 Clarke LJ, Soubrier J, Weyrich LS, and A Cooper. (2014). Environmental metabarcodes for
522 insects: in silico PCR reveals potential for taxonomic bias. *Molecular Ecology Resources*, 14,
523 1160-1170 DOI: 10.1111/1755-0998.12265

524 Cole LW. (2016). The evolution of per-cell organelle number. *Frontiers in Cell and*
525 *Developmental Biology*, 4, 85 DOI: 10.3389/fcell.2016.00085

526 Cristescu ME. (2014). From barcoding single individuals to metabarcoding biological
527 communities: towards an integrative approach to the study of global biodiversity. *Trends in*
528 *Ecology and Evolution*, 29, 566-571

529 Dabney J and M Meyer. (2012). Length and GC-biases during sequencing library amplification: a
530 comparison of various polymerase-buffer systems with ancient and modern DNA sequencing
531 libraries. *Biotechniques*, 52, 87-94 DOI: 10.2144/000113809

532 Divoll TJ, Brown VA, Kinne J, McCracken GF, and JM O'Keefe. (2018). Disparities in second-
533 generation DNA metabarcoding results exposed with accessible and repeatable workflows.
534 *Molecular Ecology Resources*, 18, 590-601

535 Edgar RC. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads.
536 *Nature Methods*, 10, 996-998 DOI: 10.1038/nmeth.2604

537 Elbrecht V and F Leese. (2015). Can DNA-based ecosystem assessments quantify species
538 abundance? testing primer bias and biomass-sequence relationships with an innovative
539 metabarcoding protocol. *PLoS ONE*, 10, e0130324 DOI: 10.1371/journal.pone.0130324

540 Elbrecht V, Taberlet P, Dejean T, Valentini A, Usseglio-Polatera P, Beisel J-N, Coissac E, Boyer F,
541 and F Leese. (2016). Testing the potential of a ribosomal 16S marker for DNA metabarcoding
542 of insects. *PeerJ*, 4, e1966 DOI: 10.7717/PeerJ.1966

543 Elbrecht V, and F Leese. (2017). Validation and development of freshwater invertebrate
544 metabarcoding COI primers for Environmental Impact Assessment. *Frontiers in*
545 *Environmental Science*, 5, 1–11. DOI: 10.3389/fenvs.2017.00011

546 Elbrecht V, Peinert B, and F Leese. (2017A). Sorting things out: Assessing effects of unequal
547 specimen biomass on DNA metabarcoding. *Ecology and Evolution*, 7, 6918-6926 DOI:
548 10.1002/ece3.3192

549 Elbrecht V, Vamos EE, Meissner K, Aroviita J, and F Leese. (2017B). Assessing strengths and
550 weaknesses of DNA metabarcoding-based macroinvertebrate identification four routine
551 stream monitoring. *Methods in Ecology and Evolution*, 8, 1265-1275 DOI: 10.1111/2041-
552 210X.12789

553 Elbrecht V, Varnos EE, Steinke D, and F Leese. (2018). Estimating intraspecific genetic diversity
554 from community DNA metabarcoding data. *PeerJ*, 6, e4644 DOI: 10.7717/PeerJ.4644

555 Ficetola GF, Pansu J, Bonin A, Cossiac E, Giguet-Covex C, De Barba M, Gielly L, Lopes CM, Boyer
556 F, Raye FPG, and P Taberlet. (2015). Replication levels, false presences, and the estimation of

the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, 15, 543-556 DOI: 10.1111/1755-0998.12338

Folmer O, Black M, Hoeh W, Lutz R, and R Vrijenhoek. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, 3, 294–299

Hajibabaei M, Ivanova NV, Ratnasingham S, Dooh RT, Kirk SL, Mackie PM, and PDN Hebert. (2005). Critical factors for assembling a high volume of DNA barcodes. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 360, 1959-1967

Hajibabaei M, Shokralla S, Zhou S, Singer GAC, and DJ Baird. (2011). Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE*, 6, e17497 DOI: 10.1371/journal.pone.0017497

Hebert PDN, Cywinska A, Ball SL, and JR deWaard. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Science*, 270, 313-321 DOI: 10.1098/rspb.2002.2218

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences of the United States of America*, 101, 14812-14817

Hebert PDN and TR Gregory. (2005). The promise of DNA barcoding for taxonomy. *Systematic Biology*, 54, 852-859 DOI: 10.1080/10635150500354886

Hebert PDN, Braukmann TWA, Prosser SWJ, Ratnasingham S, deWaard JR, Ivanova NV, Janzen DH, Hallwach W, Naik S, Sones JE, and EV Zakharov. (2018). A Sequel to Sanger: Amplicon sequencing that scales. *BMC Bioinformatics*, 19, 219 DOI: 10.1186/s12864-018-4611-3

Hernández-Triana LM, Prosser SW, Rodríguez-Perez MA, Chaverri LG, Hebert PDN, and TR Gregory. (2014). Recovery of DNA barcodes from blackfly museum specimens (Diptera: Simuliidae) using primer sets that target a variety of sequence lengths. *Molecular Ecology Resources*, 14, 508–18 DOI: 10.1111/1755-0998.12208

Ivanova NV, deWaard JR, and PDN Hebert. 2006. An inexpensive, automation friendly protocol for recovering high quality DNA. *Molecular Ecology Notes*, 6, 998-1002

Ji Y, Ashton L, Pedley SM, Edwards DP, Tang Y, Nakamura A, Kiching R, Dolman PM, Woodcock P, Edwards FA, Larsen TH, Hsu WW, Benedick S, Hamer KC, Wilcove DS, Bruce C, Wang X, Levi T, Lott M, Emerson BC, and DW Yu. (2013). Reliable, verifiable, and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, 16, 1245-1257

Landi M, Dimech M, Arculeo M, Biondo G, Martins R, Carneiro M, Carvalho GR, Brutto SL, and FO Costa. (2014). DNA barcoding for species assignment: The case of Mediterranean marine fishes. *PLoS ONE*, 9, e106135 DOI: 0.1371/journal.pone.0106135

593 Lanzen A, Lejang K, Jonassen I, Thompson EM, and C Troedsson. (2017). DNA extraction
594 replicates improve diversity and compositional dissimilarity in metabarcoding of eukaryotes
595 in marine sediments. *PLoS ONE*, 12, e0179443 DOI: 10.1371/journal.pone.0179443

596 Lee DF, Lu J, Chang S, Loparo JJ, and XS Xie. (2016). Mapping DNA polymerase error by single
597 molecule sequencing**.** *Nucleic Acids Research*, 44, e118

598 Leray M, and N Knowlton. (2015). DNA barcoding and metabarcoding of standardized samples
599 reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences*
600 *of the United States of America*, 112, 2076-2081

601 Leray M, and N Knowlton. (2017). Random sampling causes the low reproducibility of rare
602 eukaryotic OTUs in Illumina COI metabarcoding. *PeerJ*, 5, e3006

603 Macher JN, Vivancos A, Piggott JJ, Centeno FC, Matthaei CD, and F Leese. (2018). Comparison of
604 environmental DNA and bulk-sample metabarcoding using highly degenerate COI primers.
605 *Molecular Ecology Resources* DOI: 10.1111/1755-0998.12940

606 Mardis ER. (2013). Next-generation sequencing platforms. *Annual Review of Analytical*
607 *Chemistry*, 6, 287-303 DOI: 10.1146/annurev-anchem-062012-092628

608 Medeiros MJ, Eiben JA, Haines WP, Kaholoaa RL, King CBA, Krushekncky PD, Magnacca KN,
609 Rubinoff D, Starr F, and K Starr. (2013). The importance of insect monitoring to conservation
610 actions in Hawaii. *Proceedings of the Hawaiian Entomological Society*, 45, 149-166

611 Moriniere J, de Araujo BC, Lam AW, Hausmann A, Balke M, Schmidt S, Hendrich L, Docxkal D,
612 Fartmann B, Arvidsson S, and G Haszpruar. (2016). Species identification in Malaise trap
613 samples by DNA barcoding based on NGS technologies and a scoring matrix. *PLoS ONE*, 11,
614 e0155497 DOI: 10.1371/journal.pone.0155497

615 Nichols RV, Vollmers C, Newsom LA, Wang Y, Heintzman PD, Leighton M, Green RE, and B
616 Shapiro. (2018). Minimizing polymerase biases in metabarcoding. *Molecular Ecology*
617 *Resources*, 18, 927-939

618 Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB,
619 Simpson GL, Solymos P, Stevens MHH, Szoecs E and H Wagner. (2018). Vegan: Community
620 Ecology Package. R package version 2.5-1. https://CRAN.R-project.org/package=vegan

621 Pan W, Byrne-Steele M, Wang C, Lu S, Clemmons S, Zahorchak RJ, J Han. (2014). DNA
622 polymerase preference determines PCR priming efficiency. *BMC Biotechnology*, 14, 10 DOI:
623 10.1186/1472-6750-14-10

624 Pimm SL, Jenkins CN, Abell R, Brooks TM, Gittleman JL, Joppa LN, Raven PH, Roberts M, and JO
625 Sexton. (2014). The biodiversity of species and their rates of extinction, distribution, and
626 protection. *Science*, 344, 1246752 DOI:10.1126/science.1246752

627 Piñol J, Mir G, Gomez-Polo P, and N Agustí. (2015). Universal and blocking primer mismatches
628 limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of
629 arthropods. *Molecular Ecology Resources*, 15, 819-830 DOI: 10.1111 /1755-0998.12355

630 Port JA, O'Donnell JL, Romero-Maraccini OC, Leary PR, Litvin SY, Nickols KJ, Yamahara KM, and
631 RP Kelly. (2016). Assessing vertebrate biodiversity in a kelp forest ecosystem using
632 environmental DNA. *Molecular Ecology*, 25, 527-541 DOI: 10.1111/mec.13481

633 Potapov V and JL Ong. (2017). Examining sources of error in PCR by single molecule sequencing.
634 *PLoS One*, 12, e0169774

635 Prosser SWJ, deWaard JR, Miller SE, and PDN Hebert. (2016). DNA barcodes from century-old
636 type specimens using next-generation sequencing. *Molecular Ecology Resources,* 16: 487-
637 497 DOI: 10.1111/1755-0998.12474

638 R Core Team (2018). R: A language and environment for statistical computing. R Foundation for
639 Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

640 Ratnasingham S and PDN Hebert. (2013). A DNA-based registry for all animal species: the
641 barcode index number (BIN) system. *PLoS ONE*, 8, e66213 DOI:
642 10.1371/journal.pone.0066213

643 Russo L, Stehouwer R, Heberling JM, and K Shea. (2011). The composite insect trap: an
644 innovative combination trap for biologically diverse sampling. *PLoS ONE*, 26, e21079 DOI:
645 10.1371/journal.pone.0021079

646 Salipante SJ, Kawashima T, Rosenthal C, Hoogestraat DR, Cummings LA, Sengupta DJ, Harkins
647 TT, Cookson BT, and NG Hoffman. (2014). Performance comparison of Illumina and ion
648 torrent next-generation sequencing platforms for 16S rRNA-based bacterial community
649 profiling. *Applied and Environmental Microbiology*, 80, 7583–7591. DOI:
650 10.1128/AEM.02206-14

651 Sato H, Sogo Y, Doi H, and H Yamanaka. (2017). Usefulness and limitations of sample pooling for
652 environmental DNA metabarcoding of freshwater fish communities. *Scientific Reports*, 7,
653 14860 DOI: 10.1038/s41598-017-14978-6

654 Shokralla S, Porter TM, Gibson JF, Dobosz R, Janzen DH, Hallwachs W, Golding GB, and M
655 Hajibabaei. (2015). Massively parallel multiplex DNA sequencing foe specimen identification
656 using an Illumina MiSeq platform. *Scientific Reports*, 5, 9687 DOI: 10.1038/srep09687

657 Sickle W, Ankenbrand MJ, Grimmer G, Holzschuh A, Hartel S, Lanzen J, Steffan-Dewenter I, and
658 A Keller. (2015). Increased efficiency in identifying mixed pollen samples by meta-barcoding
659 with a dual-indexing approach. *BMC Ecology*, 15, 20 DOI: 10.1186/s12898-015-0051-y

660 Smith MA, Bertrand C, Crosby K, Eveleigh ES, Fernandez-Triana J, Fisher BL, Gibbs J, Hajibabaei
661 M, Hallwachs W, Hind K, Hrcek J, Huang D-W, Janda M, Janzen DH, Li Y, Miller SE, Packer L,
662 Quicke D, Ratnasingham S, Rodriguez J, Rougerie R, Shaw MR, Sheffield C, Stahlhut JK,

663 Steinke D, Whitfield J, Wood M, and X Zhou. (2012). Wolbachia and DNA barcoding insects:
664 Patterns, potential, and problems. *PLoS ONE*, 7, e36514 DOI: 10.1371/journal.pone.0036514

665 Song H, Buhay JE, Whiting MF, and KA Crandall. (2008). Many species in one: DNA barcoding
666 overestimates the number of species when nuclear mitochondrial pseudogenes are
667 coamplified. *Proceedings of the National Academy of Sciences of the United States of*
668 *America*, 105, 13486-13491 DOI: 10.1073/pnas.0803076105

669 Tedersoo L, Ansian S, Bahram M, Polme S, Riit T, Liiv I, Koljalg U, Kisand V, Nilsson H, Hildebrand
670 F, Bork P, and K Abarenkov. (2015). Shotgun metagenomes and multiple primer pair barcode
671 combinations of amplicons reveal biases in metabarcoding analyses of fungi. *MycoKeys*, 10,
672 1-43 DOI: 10.3897/mycokeys.10.4852

673 Tedersoo L, Tooming-Klunderud A, and S Anslan. (2018). PacBio metabarcoding of Fungi and
674 other eukaryotes: errors, biases, and perspectives. *New Phytologist*, 217, 1370-1385

675 Tessler M, Neumann JS, Afshinnekoo E, Pineda M, Hersch R, Velho LFM, Segovia BT, Lansac-
676 Toha FA, Lemke M, DeSalle R, Masone CE, and MR Brugler. (2017). Large-scale differences in
677 microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Scientific*
678 *Reports* 7, 6589

679 Thomas AC, Deagle BE, Eveson JP, Harsch CH, and AW Trites. (2015). Quantitative DNA
680 metabarcoding: improved estimates of species proportional biomass using correction factors
681 derived from control material. *Molecular Ecology Resources*, 16, 714–726. DOI:
682 10.1111/1755-0998.12490

683 Vasselon V, Bouchez A, Rimet F, Jacquet S, Trobajo R, Corniquel M, Tapolczai K, and I Domaizon.
684 (2017). Avoiding quantification bias in metabarcoding: Application of a cell biovolume
685 correction factor in diatom molecular biomonitoring. *Methods in Ecology and Evolution*, 9,
686 1060-1069 DOI: 10.1111/2041-210X.12960

687 Veltri KL, Espiritu M, and G Singh. (1990). Distinct genomic copy number in mitochondria of
688 different mammalian organs. *Journal of Cellular Physiology*, 143, 160–164 DOI:
689 10.1002/jcp.1041430122

690 Vivien R, Lejzerowicz F, and J Pawlowski. (2016). Next-generation sequencing of aquatic
691 oligochaetes: Comparison of experimental communities. *PLoS ONE*, 11, e0148644 DOI:
692 10.1371/journal.pone.0148644

693 Vogel G. (2017). Where have all the insects gone? *Science*, 356, 569-579 DOI:
694 10.1126/science.356.6338.576

695 Waldron A, Miller DC, Redding D, Mooers A, Kuhn TS, Nibbelink N, Roberts JT, Tobias JA, and JL
696 Gittleman. (2017). Reductions in global biodiversity loss predicted from conservation
697 spending. *Nature*, 551, 364-367 DOI:10.1038/nature24295

698 Wickham H. (2009). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York

699 Wilkinson MJ, Szabo C, Ford CS, Yarom Y, Croxford AE, Camp A, and P Gooding. (2017).
700 Replacing Sanger with Next Generation Sequencing to improve coverage and quality of
701 reference DNA barcodes for plants. *Scientific Reports*, 7, 46040 DOI: 10.1038/srep46040

702 Yamamoto S, Masuda R, Sato Y, Sado T, Araki H, Kondoh M, Minamoto T, and M Miya. (2017).
703 Environmental DNA metabarcoding reveals local fish communities in a species-rich coastal
704 sea. *Scientific Reports*, 7, 40368 DOI: 10.1038/srep40368

705 Yu DW, Ji Y, Emerson BC, Wang X, Ye C, Yang C, and Z Ding. (2012). Biodiversity soup:
706 metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods*
707 *in Ecology and Evolution*, 3, 613-623

708 Zimmerman J, Abarca N, Enk N, Skibbe O, Kusber W-H, and R Jahn. (2014). Taxonomic reference
709 libraries for environmental barcoding: a best practice example from diatom research. *PLoS*
710 *ONE*, 9, e114758 DOI: 10.1371/journal.pone.01147

Figure 1 – Protocol employed to examine species recovery from the mock community. Four amplicon pools were examined. Two derived from bulk DNA extracts (Bulk Abdomen, Bulk Leg). The others derived from DNA extracts from single legs that were either pooled (Composite Leg) or kept separate (Single Leg) prior to PCR. All four amplicon pools were sequenced on three platforms (Illumina MiSeq, Ion Torrent S5, and Ion Torrent PGM). There were three technical replicates for each treatment except Single Leg.

Figure 2 – Rarefaction curves showing BIN recovery versus the number of sequences analyzed for the four amplicon pools (Bulk Abdomen, Bulk Leg, Composite Leg, Single Leg) on the three sequencing platforms. Two amplicon lengths (407 bp, 463 bp) were analyzed on the S5, but just one (407 bp) on the other platforms.

Figure 3A – Bray-Curtis Dissimilarity dendrogram for the four amplicon pools (BA = Bulk abdomen, BL = Bulk leg, CL = Composite leg, SL = Single Leg). Replicates are numbered 1-3 while P is the result from pooling the replicates. The 463 bp amplicon is indicated with an asterisk (*). 3B – Non-metric multidimensional scaling (NMDS) ordinations using Bray-Curtis dissimilarity for the four amplicon pools. Coloured ellipses represent 95% confidence intervals for the BIN composition of the different treatments using ordiellipse (Oksanen et al. 2012). The shapes within each ellipse represent replicates for the four combinations of sequencing platform-amplicon length for three treatments. No replicates were available for the Single Leg treatment, so it has just four points.

Figure 4 – Heat map showing the relative log abundance of the 369 BINs in each treatment for the four amplicon pools. This heat map was created using the JAMP package (https://github.com/VascoElbrecht/JAMP). Technical replicates are indicated with numbers while *in silico* pooled results are designated by the letter P.

Table 1 – Summary of run results for all treatments. mBRAVE filtering and BIN recovery including false positives are indicated for the four amplicon pools (BA = Bulk Abdomen, BL = Bulk Leg, CL = Composite Leg, SL = Single Leg). Replicates are numbered 1-3 and pooled replicates are denoted by a P. BINs (Barcode Index Number) for not reference library matches were only counted if their relative abundance was greater than 0.01%. All results are based on the analysis of a 407 bp amplicon except those marked with a * which are based on a 463 bp amplicon.

Table 2 – Values for selected diversity indices (Shannon-Weaver, Simpson, Inverse Simpsons, Pielou's Evenness) for the four amplicon pools (BA = Bulk Abdomen, BL = Bulk Leg, CL = Composite Leg, SL = Single Leg). Replicates are numbered 1-3 while P is the result from pooling the replicates. All results are based on the analysis of a 407 bp amplicon except those marked with a * which are based on a 463 bp amplicon.

Supplementary Figures and Tables

Figure S1A – Phred or quality (QV) scores across read length for 407 bp amplicon on the Illumina MiSeq, Ion Torrent PGM, and Ion Torrent S5 as well as for 463 bp amplicon on the Ion Torrent S5. S1B – Histogram of read lengths for the three platforms and two amplicon lengths on the Ion Torrent S5.

Figure S2 – Renyi's diversity graphs using the pooled replicates for the Bulk Abdomen, Bulk Leg, and Composite Leg treatments and the single replicate for the Single Leg treatment.

Figure S3 – Density plots based on the relative abundance for the 369 BINs common across all treatments. The results from Bulk Abdomen, Bulk Leg, Composite Leg, and Single Leg PCR are represented by blue, green, red, and black lines respectively. Each panel represents a different sequencing platform: A) 463 bp on the Ion Torrent S5, B) 407 bp on the Ion Torrent S5, C) 407 bp on the Illumina MiSeq, and D) 407 bp on the Ion Torrent PGM.

Figure S4 – Jaccard Similarity dendrogram for the four amplicon pools (BA = Bulk Abdomen, BL = Bulk Leg, CL = Composite Leg, SL = Single Leg). Replicates are numbered 1-3 while P is the result from pooling the replicates. The 463 bp amplicon is indicated with an asterisk (*).

Figure S5 – Non-metric multidimensional scaling (NMDS) ordinations using Bray-Curtis dissimilarity for the four amplicon pools (BA = Bulk Abdomen, BL = Bulk Leg, CL = Composite Leg, SL = Single Leg). Coloured ellipses represent 95% confidence intervals for the BIN composition of the different treatments using ordiellipse (Oksanen et al. 2012). No replicates were available for the Single Leg treatment, so it has just four points.

Table S1 – Taxonomy and results for the mock communities used in this study. A) includes the read depth for each BIN while B) shows the relative abundance of each BIN in the different mock communities for the four amplicon pools (BA = Bulk Abdomen, BL = Bulk Leg, CL = Composite Leg, SL = Single Leg). Replicates are numbered 1-3 while P is the result from pooling the replicates. The 463 bp amplicon is indicated with an asterisk (*).

Table S2 – Primer sequences for the different platforms. A) primer sequences for Ion Torrent PGM and S5 and B) primers for the Illumina MiSeq including the COI primer (red) and unique molecular identifiers (UMIs; green). The Nextera Transposase adapters have two components: an adaptor sequence (yellow) and a sequencing primer (purple). Sequencing adaptors (Ion Torrent PGM and S5) and flow cell adapters (Illumina) are shown in blue.

Table S3 – Slopes from rarefaction curve at the prior 1, 2, 3, 5, and 10 points for the four amplicon pools (BA = Bulk Abdomen, BL = Bulk Leg, CL = Composite Leg, SL = Single Leg). Replicates are numbered 1-3 while P is the result from pooling the replicates. The 463 bp amplicon is indicated with an asterisk (*).

Table S4 – Summary table of primer mismatches for the 463 bp and 407 bp amplicons for A) each BIN and B) mean read depth per primer mismatch for the four amplicon pools (BA = Bulk Abdomen, BL = Bulk Leg, CL = Composite Leg, SL = Single Leg). Replicates are numbered

783 1-3 while P is the result from pooling the replicates. The 463 bp amplicon is indicated with an
784 asterisk (*).

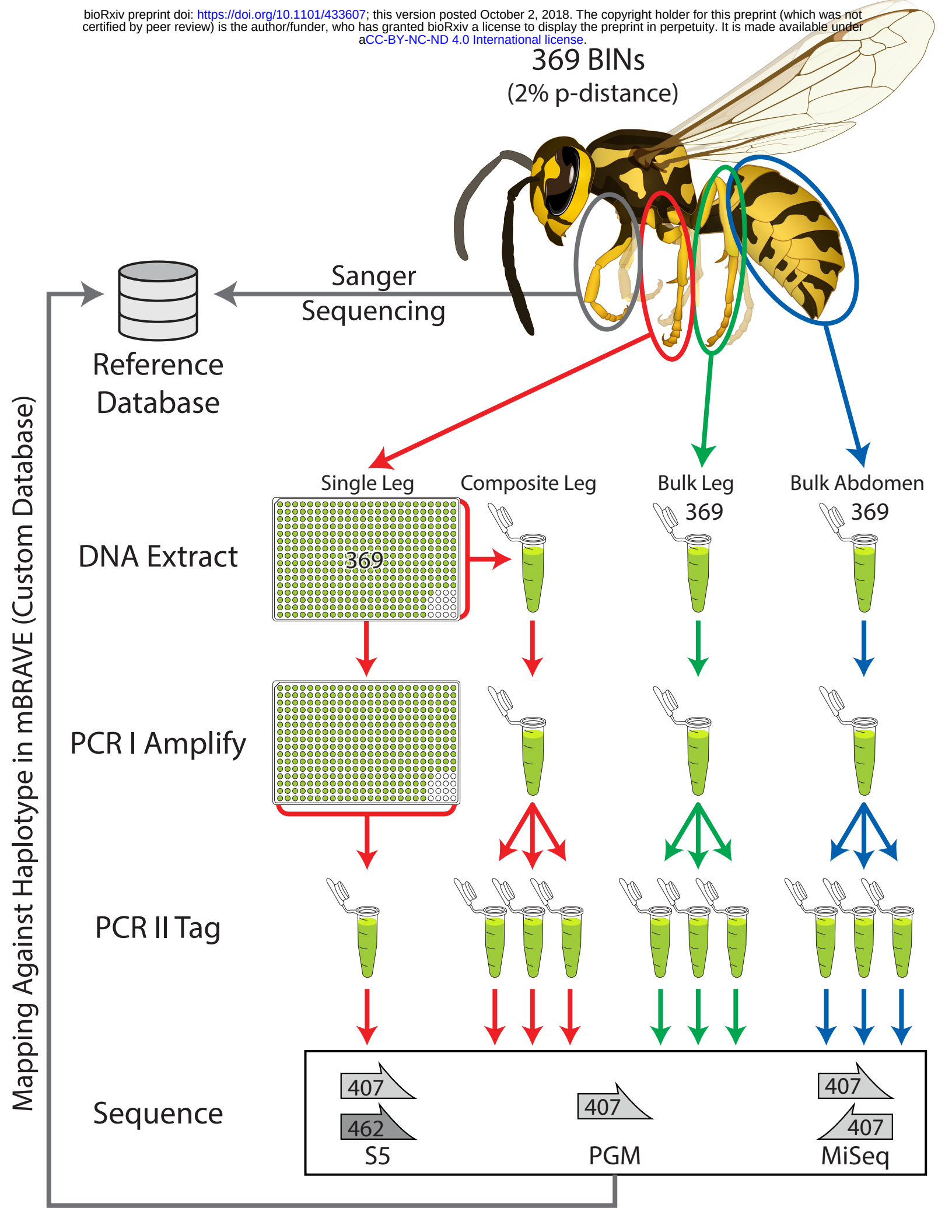
785 Tables S5 – Read depth and relative abundance for each treatment and replicate by A) Order
786 and B) Family for the four amplicon pools (BA = Bulk Abdomen, BL = Bulk Leg, CL =
787 Composite Leg, SL = Single Leg). Replicates are numbered 1-3 while P is the result from
788 pooling the replicates. The 463 bp amplicon is indicated with an asterisk (*).

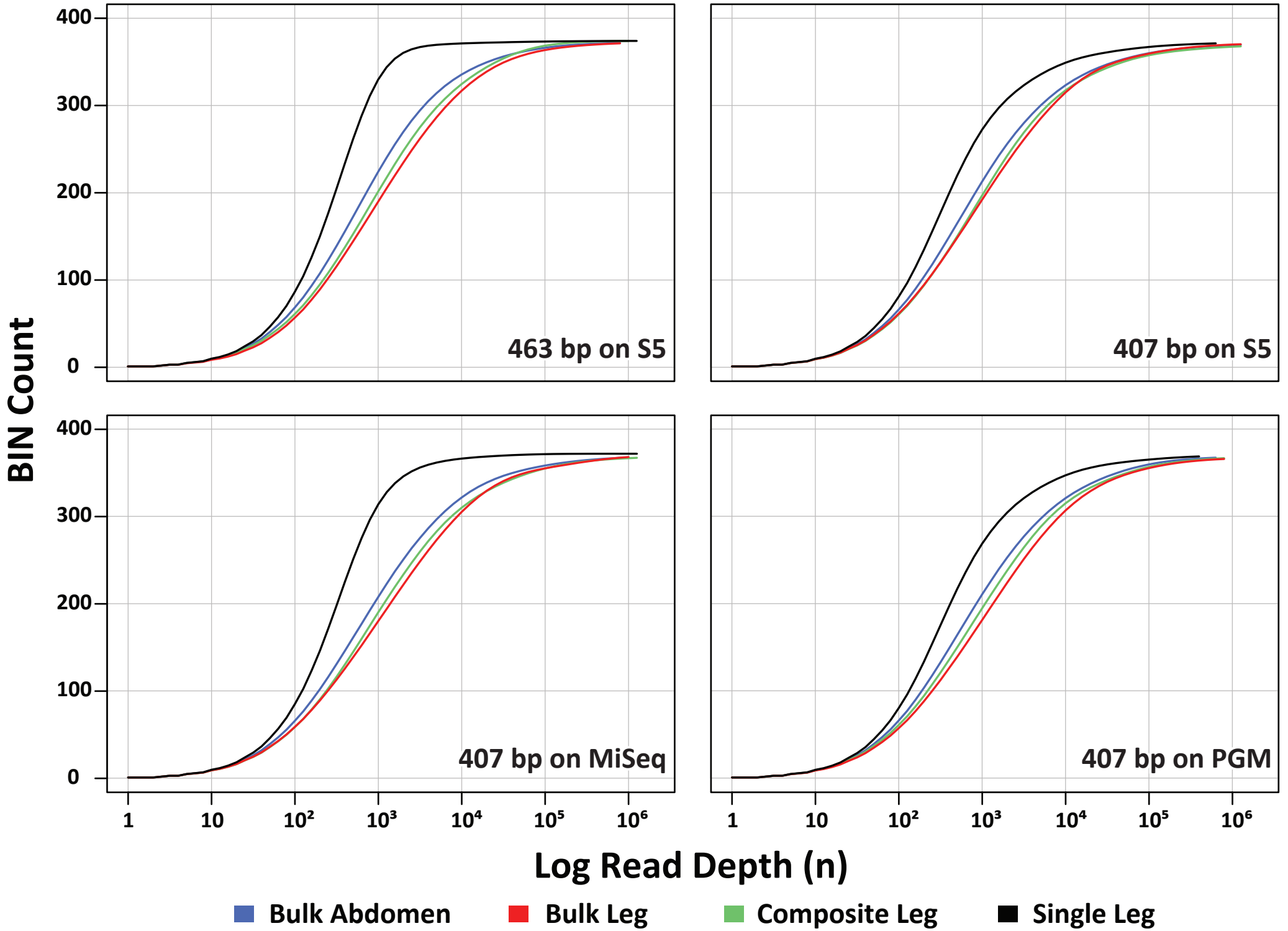
789

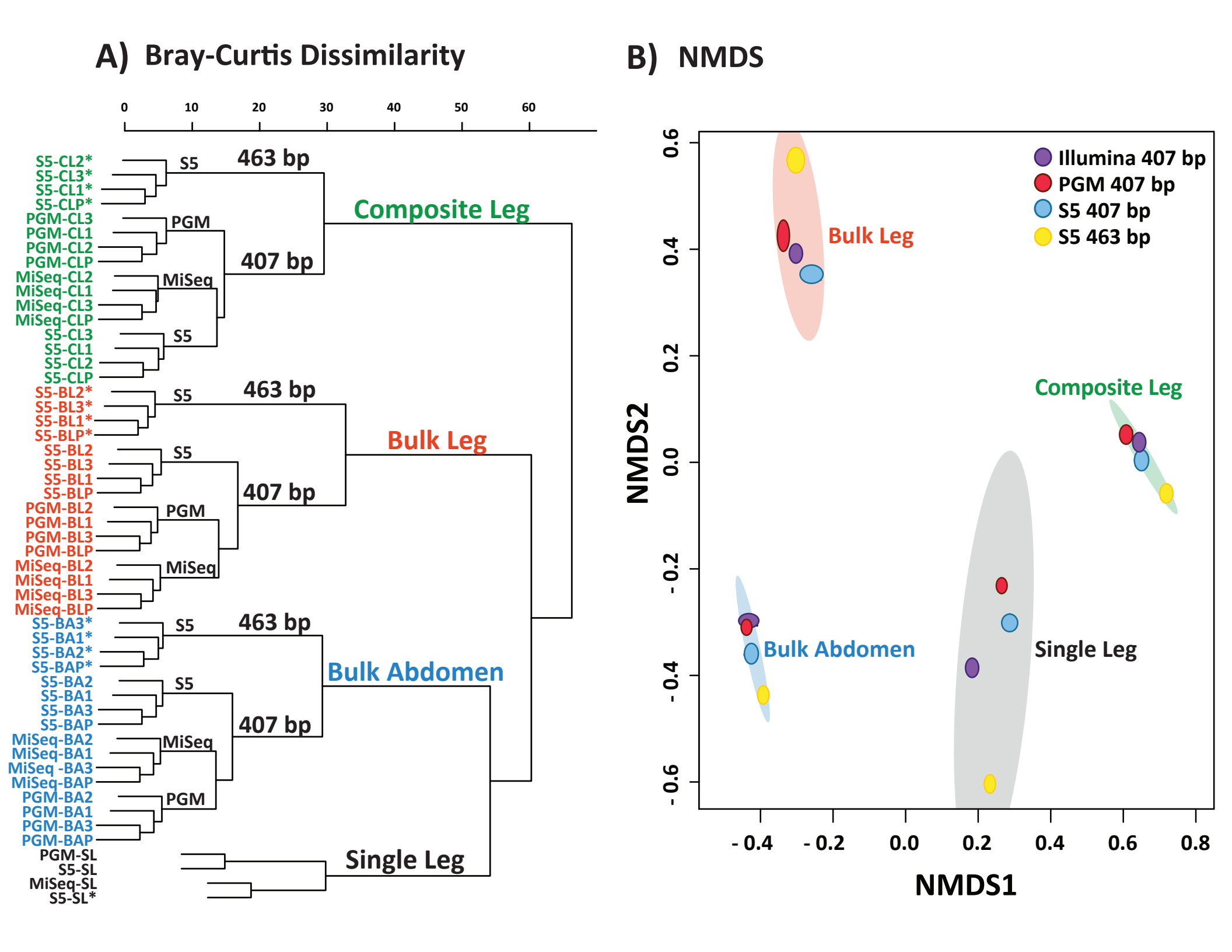
790 **Other Supplementary material**

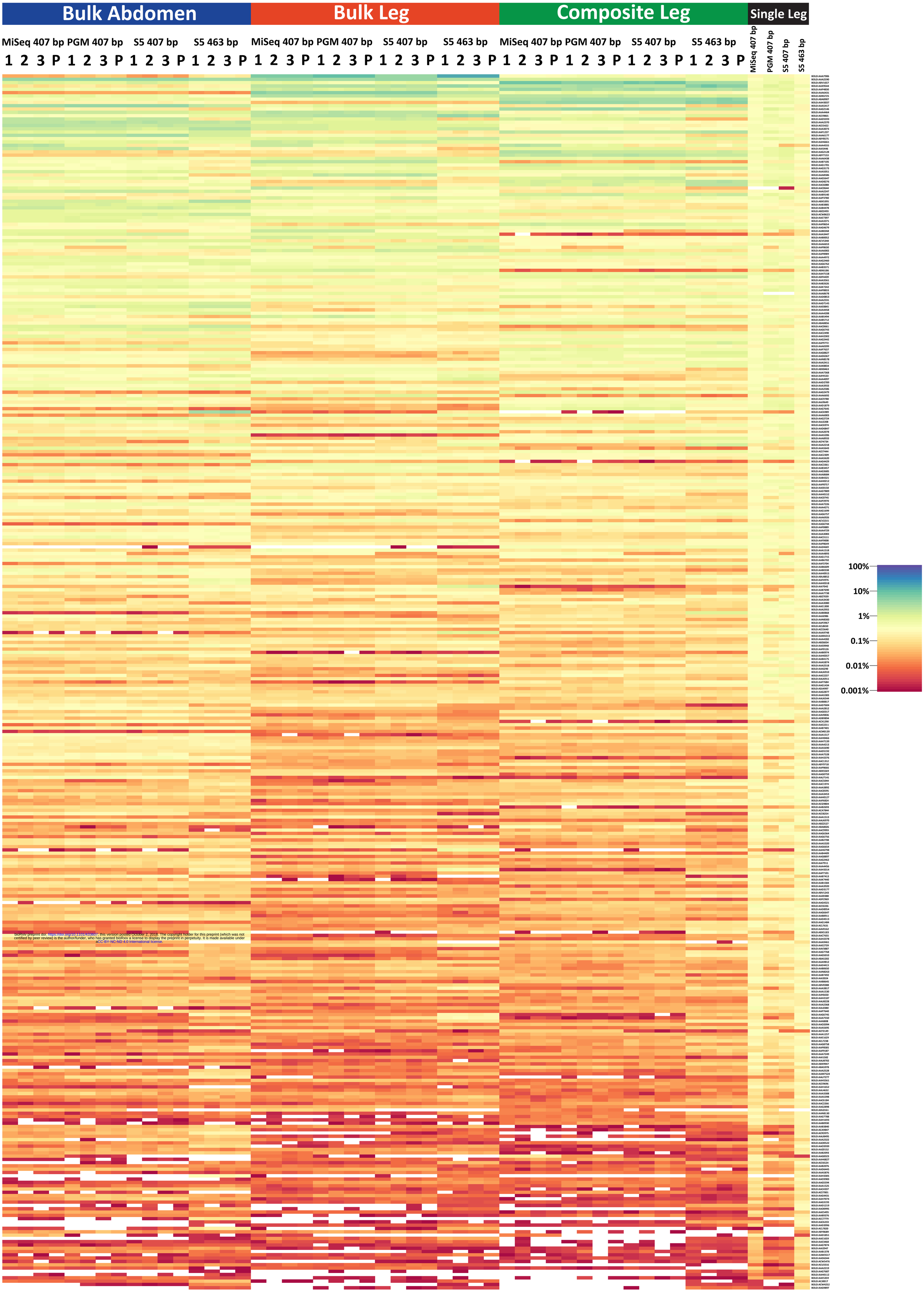
791 MBRAVE_OTUmerger_MER.R – R script for combining mBRAVE OTU tables into single file.

792 OTU_subsampler_MER.R – R Script for generating rarefraction curves from an OTU table.




BIN Count
400
300
200
100
0
463 bp on S5
407 bp on S5
407 bp on MiSeq
407 bp on PGM
1
10
10²
10³
10⁴
10⁵
10⁶
Log Read Depth (n)
Bulk Abdomen
Bulk Leg
Composite Leg
Single Leg


A) Bray-Curtis Dissimilarity
0
10
20
30
40
50
60
S5-CL2*
S5-CL3*
S5-CL1*
S5-CLP*
PGM-CL3
PGM-CL1
PGM-CL2
PGM-CLP
MiSeq-CL2
MiSeq-CL1
MiSeq-CL3
MiSeq-CLP
S5-CL3
S5-CL1
S5-CL2
S5-CLP
S5-BL2*
S5-BL3*
S5-BL1*
S5-BLP*
S5-BL2
S5-BL3
S5-BL1
S5-BLP
PGM-BL2
PGM-BL1
PGM-BL3
PGM-BLP
MiSeq-BL2
MiSeq-BL1
MiSeq-BL3
MiSeq-BLP
S5-BA3*
S5-BA1*
S5-BA2*
S5-BAP*
S5-BA2
S5-BA1
S5-BA3
S5-BAP
MiSeq-BA2
MiSeq-BA1
MiSeq -BA3
MiSeq-BAP
PGM-BA2
PGM-BA1
PGM-BA3
PGM-BAP
PGM-SL
S5-SL
MiSeq-SL
S5-SL*
S5
PGM
MiSeq
S5
463 bp
407 bp
Composite Leg
S5
PGM
MiSeq
S5
463 bp
407 bp
Bulk Leg
S5
PGM
MiSeq
S5
463 bp
407 bp
Bulk Abdomen
Single Leg
B) NMDS
Illumina 407 bp
PGM 407 bp
S5 407 bp
S5 463 bp
Bulk Leg
Composite Leg
Bulk Abdomen
Single Leg
NMDS2
0.6
0.4
0.2
0.0
- 0.2
- 0.4
- 0.6
- 0.4
- 0.2
0.0
0.2
0.4
0.6
0.8
NMDS1



Platform	Treatment (Replicate #)	uploaded reads (n)	post filter reads (n)	mean length (bp)	mean QV	mean GC%	Reads matching reference library	BINs recovered	BINS > 0.01% RA	BINs > 0.01% RA	chimeric reads (n)	Bacterial reads (n)	Bacterial BINs	Insect reads (n)	Insect BINs	non-insect reads	non-insect BINs	unmatched reads (n)	umatched read OTUs
MiSeq	BA1	357146	356697	405	36.13	0.319	323704	364	317	57	26356	37	0	951	5	1	0	5648	13
	BA2	403622	403052	405	36.18	0.319	368224	362	315	59	27648	49	0	1183	4	0	0	5948	17
	BA3	506268	505585	405	36.17	0.319	460032	364	318	56	36377	63	0	1319	4	13	0	7781	20
	BAP	1267036	1265534	405	36.16	0.319	1151960	369	316	58	91560	149	0	3453	4	14	0	18198	14
	BL1	422755	422421	405	36.12	0.320	382545	360	293	81	32008	46	0	1139	3	9	0	6674	18
	BL2	373342	373126	405	35.99	0.319	337850	362	296	78	27706	34	0	1079	5	17	0	6440	20
	BL3	546462	546060	405	36.25	0.321	495255	364	290	84	40309	43	0	1491	4	5	0	8957	23
	BLP	1342560	1341607	405	36.14	0.320	1215650	370	297	77	101231	123	0	3709	3	31	0	20863	22
	CL1	521445	521181	405	36.26	0.319	472713	361	296	78	28023	51	0	1518	8	25	0	8851	19
	CL2	442923	442543	405	36.12	0.320	398739	366	296	78	35139	21	0	1122	4	1	0	7521	20
	CL3	545824	545602	405	36.24	0.319	492799	362	304	70	42018	33	0	1508	5	5	0	9239	16
	CLP	1510192	1509326	405	36.27	0.319	1364250	369	303	71	116265	105	0	4132	5	31	0	24542	20
	SL	1542853	1540592	405	36.23	0.315	1513470	372	365	9	2338	521	0	7586	10	133	0	16543	6
PGM	BA1	473837	259691	381	27.62	0.321	243861	366	316	58	6707	32	0	961	7	2	0	8128	16
	BA2	448934	258358	381	27.72	0.321	242366	362	312	62	6814	62	0	1066	4	4	0	8046	11
	BA3	534795	326968	382	27.69	0.321	306818	362	310	64	9406	63	0	1246	8	3	0	9432	6
	BAP	1457566	845017	382	27.68	0.321	793045	369	314	60	24707	157	0	3273	6	9	0	23826	9
	BL1	591907	328155	379	27.63	0.324	307657	362	298	76	9672	87	1	1194	4	10	0	9535	9
	BL2	476860	272357	380	27.61	0.323	256701	363	287	87	6835	69	0	1056	7	20	0	7676	10
	BL3	603124	328797	380	27.7	0.324	307758	362	292	82	10234	48	0	1156	6	6	0	9595	12
	BLP	1671711	929309	379	27.65	0.324	872116	367	286	88	28241	204	0	3406	4	36	0	25306	12
	CL1	532057	300890	384	27.89	0.321	285294	362	303	71	7485	45	0	1167	7	20	0	6879	5
	CL2	519479	291023	383	27.86	0.321	277352	363	311	63	6161	39	0	885	8	1	0	6585	6
	CL3	453785	256149	384	27.83	0.321	245042	360	312	62	4616	33	0	769	6	1	0	5688	3
	CLP	1505320	848052	384	27.86	0.321	807688	368	313	61	19978	117	0	2821	7	22	0	17436	4
	SL	787631	438045	381	27.76	0.319	420427	370	347	27	3207	96	0	2423	14	15	0	11877	6
S5	BA1	1032904	435627	396	27.01	0.322	416248	365	317	57	9616	111	0	1478	4	6	0	8168	10
	BA2	935020	408907	396	27.02	0.322	390537	364	317	57	8652	194	1	1628	3	3	0	7893	11
	BA3	1140310	537677	396	27.18	0.322	512138	366	316	58	13650	129	1	1821	3	10	0	9929	12
	BAP	3108238	1382211	396	27.08	0.322	1318940	370	317	57	34739	434	1	4889	4	19	0	23129	9
	BL1	1145406	489294	393	27.03	0.321	460637	365	304	70	14549	227	1	1640	3	5	0	12236	12
	BL2	925531	401390	394	27.06	0.321	379942	363	309	65	10055	167	1	1584	4	12	0	9630	13
	BL3	1186802	491140	393	26.98	0.322	461824	364	305	69	15258	161	1	1495	3	8	0	12394	14
	BLP	3257559	1381824	393	27.02	0.321	1302400	371	310	64	42257	555	1	4719	2	25	0	31865	13
	CL1	1188057	547727	396	27.26	0.321	521824	364	304	70	14012	128	0	1706	9	17	0	10040	8
	CL2	1029459	479655	396	27.22	0.320	459412	366	310	64	10286	140	1	1128	5	5	0	8684	7
	CL3	991378	474744	396	27.32	0.320	456177	363	311	63	8722	151	1	1203	3	3	0	8488	8
	CLP	3208890	1502126	396	27.27	0.320	1437430	369	311	63	35561	419	0	3991	6	25	0	24700	6
	SL	1735346	769426	394	27.12	0.319	743673	372	348	26	4778	350	1	3262	7	24	0	17339	10
	BA1*	906390	401586	447	27.27	0.323	336480	372	331	43	48317	1095	4	1293	5	44	0	14357	11
	BA2*	817662	379557	448	27.3	0.323	315568	368	327	47	48099	814	4	1452	8	37	0	13587	8
	BA3*	858365	383034	448	27.22	0.324	322595	370	331	43	45058	975	3	1194	7	12	0	13200	10
	BAP*	2582420	1164177	448	27.27	0.323	974643	372	331	43	148719	2884	4	3939	5	93	0	33899	7
	BL1*	790492	366040	450	27.32	0.326	310157	369	305	69	41399	878	4	1190	7	20	0	12396	14
	BL2*	865993	393835	450	27.3	0.327	333924	369	306	68	44241	848	4	1143	6	26	0	13653	15
	BL3*	857754	383518	449	27.32	0.326	324214	370	306	68	44200	770	5	1116	5	14	0	13204	14
	BLP*	2514240	1143393	450	27.31	0.326	968295	372	308	66	135142	2496	3	3449	7	60	0	33951	17
	CL1*	965017	468482	449	27.4	0.324	388336	372	312	62	58208	359	3	2057	10	8	0	19214	36
	CL2*	860367	431266	449	27.41	0.323	363098	374	312	62	49241	445	2	1830	10	5	0	16647	29
	CL3*	806183	403854	449	27.38	0.324	341946	373	311	63	44404	505	3	1579	11	33	0	15117	27
	CLP*	2631840	1303332	449	27.4	0.324	1093380	374	312	62	156996	1609	3	5466	8	46	0	45835	35
	SL*	3301111	1549636	449	27.41	0.323	1477380	374	371	3	23821	8227	5	9886	14	1877	3	28444	4

Platform	Treatment (Replicate #)	Pielou's Evenness	Simpson's	InvSimpson	Shannon Weaver
MiSeq	BA1	0.84	0.99	83.39	4.96
	BA2	0.84	0.99	83.89	4.96
	BA3	0.84	0.99	82.55	4.95
	BAP	0.84	0.99	83.38	4.96
	BL1	0.80	0.98	55.48	4.68
	BL2	0.79	0.98	55.51	4.66
	BL3	0.79	0.98	52.16	4.65
	BLP	0.79	0.98	54.23	4.67
	CL1	0.79	0.98	47.39	4.65
	CL2	0.79	0.98	48.17	4.64
	CL3	0.80	0.98	49.44	4.67
	CLP	0.79	0.98	48.47	4.66
	SL	0.98	1.00	294.42	5.76
PGM	BA1	0.84	0.99	80.52	4.97
	BA2	0.84	0.99	80.15	4.96
	BA3	0.85	0.99	81.80	4.97
	BAP	0.84	0.99	81.04	4.97
	BL1	0.78	0.98	40.80	4.58
	BL2	0.78	0.98	41.45	4.58
	BL3	0.78	0.98	41.43	4.59
	BLP	0.78	0.98	41.26	4.59
	CL1	0.81	0.98	58.74	4.74
	CL2	0.81	0.98	60.84	4.76
	CL3	0.82	0.98	63.78	4.79
	CLP	0.81	0.98	61.12	4.77
	SL	0.94	1.00	212.46	5.53
S5	BA1	0.84	0.99	78.96	4.96
	BA2	0.84	0.99	79.87	4.97
	BA3	0.84	0.99	80.04	4.97
	BAP	0.84	0.99	79.83	4.97
	BL1	0.81	0.98	64.51	4.79
	BL2	0.81	0.98	64.35	4.78
	BL3	0.81	0.98	63.59	4.78
	BLP	0.81	0.98	64.27	4.79
	CL1	0.80	0.98	51.54	4.71
	CL2	0.80	0.98	54.07	4.73
	CL3	0.81	0.98	55.41	4.76
	CLP	0.80	0.98	53.67	4.74
	SL	0.94	1.00	217.39	5.54
	BA1*	0.86	0.99	95.11	5.07
	BA2*	0.86	0.99	95.49	5.07
	BA3*	0.86	0.99	92.22	5.05
	BAP*	0.86	0.99	94.57	5.07
	BL1*	0.75	0.95	22.20	4.45
	BL2*	0.75	0.95	21.39	4.43
	BL3*	0.76	0.96	22.70	4.46
	BLP*	0.75	0.95	22.09	4.45
	CL1*	0.81	0.98	60.80	4.78
	CL2*	0.80	0.98	51.09	4.72
	CL3*	0.81	0.98	60.73	4.79
	CLP*	0.81	0.98	58.16	4.77
	SL*	0.99	1.00	319.57	5.82
Index Maximum		1.00	1.00	369.00	5.91