




RESOURCE ARTICLE

Metabarcoding a diverse arthropod mock community

Thomas W. A. Braukmann¹  | Natalia V. Ivanova¹ | Sean W. J. Prosser¹ |
 Vasco Elbrecht¹  | Dirk Steinke^{1,2}  | Sujeevan Ratnasingham¹ |
 Jeremy R. de Waard^{1,3} | Jayme E. Sones¹ | Evgeny V. Zakharov¹ | Paul D. N. Hebert^{1,2}

¹Centre for Biodiversity
Genomics, University of Guelph, Guelph,
Ontario, Canada

²Department of Integrative
Biology, University of Guelph, Guelph,
Ontario, Canada

³School of Environmental
Sciences, University of Guelph, Guelph,
Ontario, Canada

Correspondence

Thomas W. A. Braukmann, Centre for
Biodiversity Genomics, University of
Guelph, Guelph, ON, Canada.
Email: tbraukma@uoguelph.ca

Funding information

Ministry of Research, Innovation and
Science

Abstract

Although DNA metabarcoding is an attractive approach for monitoring biodiversity, it is often difficult to detect all the species present in a bulk sample. In particular, sequence recovery for a given species depends on its biomass and mitome copy number as well as the primer set employed for PCR. To examine these variables, we constructed a mock community of terrestrial arthropods comprised of 374 species. 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 segments (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 amplicons 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, although the MiSeq delivered the highest sequence quality. As expected, species recovery was most efficient from the Single Leg treatment because amplicon abundance varied little among taxa. Among the three treatments where PCR occurred after pooling, the Bulk Abdomen treatment produced a more uniform read abundance than the Bulk Leg or Composite Leg treatment. Primer choice also influenced species recovery and evenness. Our results reveal how variation in protocols can have substantial impacts on perceived diversity unless sequencing coverage is sufficient to reach an asymptote.

KEYWORDS

community ecology, DNA barcoding, ecological genetics, environmental DNA

1 | INTRODUCTION

It is generally accepted we have entered a period of unprecedented biodiversity loss (Pimm et al., 2014; Vogel, 2017). Evaluating the scope and regional variation in this loss will require the capacity

to quantify shifts in species composition rapidly and on a far larger scale than ever before to better understand and manage ecosystems (Cristescu, 2014; Ji et al., 2013; 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

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2019 The Authors. *Molecular Ecology Resources* Published by John Wiley & Sons Ltd

bio-surveillance. Although they are easily collected in large numbers (Russo, Stehouwer, Heberling, & Shea, 2011), the subsequent processing and identification of specimens has traditionally been a barrier to large-scale monitoring programmes (Bassett et al., 2012). DNA barcoding, the use of short standardized gene regions to discriminate species, breaks this barrier by enabling nontaxonomists to identify specimens once a reference sequence library is established (Hebert, Cywinska, Ball, & de Waard, 2003; Hebert & Gregory, 2005).

DNA barcode studies initially focused on developing the analytical protocols to construct a specimen-based reference library (Hebert et al., 2003; Hebert, Penton, Burns, Janzen, & Hallwachs, 2004). Although improved protocols have reduced costs, leading to the analysis of millions of single specimens (Hajibabaei et al., 2005; Hebert et al., 2018; Ivanova, deWaard, & Hebert, 2006), this approach is too expensive to support large-scale bio-monitoring programmes. 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 (Brandon-Mong et al., 2015; Hajibabaei, Shokralla, Zhou, Singer, & Baird, 2011; Moriniere et al., 2016; Yu et al., 2012). It achieves this goal by generating amplicons internal to the barcode region from bulk DNA extracts which are then sequenced and assigned to operational taxonomic units (OTUs) that are queried against reference sequences to ascertain their source species (see Cristescu, 2014). Studies have now employed this approach to assess species composition in communities of aquatic and terrestrial arthropods (Beng et al., 2016; Elbrecht, Vamos, Meissner, Aroviita, & Leese, 2017; Ji et al., 2013), vertebrates (Sato, Sogo, Doi, & Yamanaka, 2017), diatoms (Vasselon et al., 2017) and fungi (Aas, Davey, & Kausrud, 2017; Bellemain et al., 2012; Tedersoo, Tooming-Klunderud, & Anslan, 2018). Such metabarcoding analysis routinely reveals more species than morphological approaches while requiring far less time (Brandon-Mong et al., 2015; Elbrecht, Peinert, & Leese, 2017; Elbrecht, Vamos et al., 2017; Hebert et al., 2018; Ji et al., 2013; Shokralla et al., 2015; Vivien, Lejzerowicz, & Pawlowski, 2016; Yu et al., 2012).

Despite the advantages of metabarcoding, several factors often complicate the recovery of all species in a sample. First, DNA templates derived from the species in a mixed sample are often differentially amplified (Elbrecht & Leese, ; Piñol, Mir, Gomez-Polo, & Agustí, 2015; Tedersoo et al., 2018). Such bias can arise from either the DNA polymerase (Dabney & Meyer, 2012; Nichols et al., 2018; Pan et al., 2014) or the PCR primers (Clarke, Soubrier, Weyrich, & Cooper, 2014). Polymerase bias involves the differential amplification of templates as a result of variation in their sequence motifs, GC content or length (Dabney & Meyer, 2012; Nichols et al., 2018; Pan et al., 2014). Primer bias arises due to either varying levels of primer mismatch or template degradation (Clarke et al., 2014; Elbrecht & Leese, 2015). The impact of primer mismatches can often be reduced either by lowering annealing temperatures or by raising the degeneracy of the primers (Clarke et al., 2014; Elbrecht & Leese, 2017). However, these "solutions" have a downside; they often increase the amplification of

nontarget sequences such as bacterial endosymbionts or mitochondrial pseudogenes, which is especially problematic for eDNA studies (Macher et al., 2018; Smith et al., 2012; Song, Buhay, Whiting, & Crandall, 2008).

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; Elbrecht, Peinert et al., 2017). Because of this effect (in addition to primer bias), efforts to infer species abundance from read counts obtained in metabarcoding studies are at best weak (Elbrecht & Leese, 2015; Piñol et al., 2015). Correction factors can improve such estimates (Thomas, Deagle, Eveson, Harsch, & Trites, 2015; Vasselon et al., 2017), but any method based on the analysis of bulk DNA extracts will fail to accurately determine 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, Lu, Chang, Loparo, & Xie, 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 platforms, such as Pacific Biosciences Sequel (e.g. Hebert et al., 2018), can produce sequences with much lower error rates, but they currently generate too few reads (~0.3 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 (Cristescu, 2014; Mardis, 2013). Illumina sequencers generate more reads (20 million–10 billion/run) with lower error rates than Ion Torrent platforms, but the latter instruments can deliver longer reads and can generate results 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, Brown, Kinne, McCracken, & O'Keefe, 2018). However, work on microbial communities found general agreement between platforms although reads from Ion Torrent platforms were lower quality and more length variable than those from Illumina (Salipante et al., 2014; Tessler et al., 2017). In cases where speed is critical, Ion Torrent platforms have an advantage because of their short run times.

To explore factors affecting the reliability of metabarcoding, we targeted the 658 bp barcode region of the cytochrome *c* oxidase I gene (COI). This gene region has three advantages for metabarcoding studies. First, reference sequences are available for more than 500,000 animal species, far more than any other gene region (Andújar, Arribas, Yu, Vogler, & Emerson, 2018; Porter & Hajibabaei, 2018). Second, because it is protein-coding, pseudogenes can often

be detected because of the presence of frameshift mutations or stop codons. Third, COI has a more rapid rate of evolution than other candidate gene regions, an important advantage in discriminating closely related species, especially given the short amplicons often employed in metabarcoding. Some recent studies have suggested that 16S RNA would be a better target region for metabarcoding studies because of its more conserved priming sites (Deagle, Jarman, Cossiac, Pompanon, & Taberlet, 2014; Elbrecht et al., 2016). This proposal overlooks three disadvantages: (a) it lacks a comprehensive reference database, (b) its slower rate of evolution means that sister species will often share the same sequence (Andújar et al., 2018), and (c) no diagnostic sequence changes are available to recognize pseudogenes. Based on these three weaknesses, it is clear that COI is the better gene region for metabarcoding studies and that effort should be directed towards primer redesign in those cases where current primer sets are ineffective for the group under study (Elbrecht & Leese, 2017).

To examine factors influencing the success in recovering species through metabarcoding analysis of COI, we assembled a mock community that included single representatives of 374 insect species. We subsequently used this community to examine the impacts of DNA source, extraction method, PCR protocol, amplicon template and sequencing platform on species recovery. In particular, we examined whether tissue type (abdomens and legs) influences success in the recovery of community composition or whether certain tissues are more prone to false positives. We also wanted to ascertain whether sample processing (bulk vs. individual) affected species recovery. Furthermore, we compared the major HTS platforms to determine whether different sequencing technologies introduced a bias. Specifically, we compared results obtained by analysing read abundance, evenness and species recovery for 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 and leg) to test the impact of tissue type. The other two amplicon pools derived from DNA extracts of single legs that were analysed by pooling prior to or after PCR. Finally, we examined species recovery and evenness 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 on three sequencing platforms.

2 | MATERIALS AND METHODS

2.1 | 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 the primer

cocktail of C_LepFolF/C_LepFolR (Hernández-Triana 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 (Hebert et al., 2018; Ivanova et al., 2006). Unpurified PCR products were diluted 1:4 with ddH₂O before 2 µl was used as the template for a cycle sequencing reaction (Hebert et al., 2018). All products were sequenced in the forward and reverse directions 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 & Hebert, 2013). The overall analysis provided sequence records for 803 BINs. From this total, we selected 374 BINs showing > 2% COI sequence divergence from their nearest neighbour under the Kimura 2-parameter model (Kimura, 1980). The resulting mock community included representatives of 10 orders and 104 insect families. Supporting Information Table S1 lists the taxa in the mock community and provides details on vouchers, their body size (as estimated by abdominal mass) and the GC content of their COI barcode. Following selection of the specimens for inclusion in the mock community, DNA extraction and PCR utilized the protocols described below, and the resultant amplicon pools were analysed on three sequencing platforms.

2.2 | Experimental design for metabarcoding analysis

Species recovery was compared for amplicon pools that resulted from four DNA extraction/PCR protocols (Figure 1). Two involved the analysis of amplicons generated from bulk DNA extracts derived from two tissues (Bulk Abdomen and 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 and BOLD:AAP6354; BOLD:ABV1240). Due to the complexity of our mock community and our desire to evaluate several variables (tissue type, PCR protocol, PCR amplicon and sequencing platform), we did not evaluate multiple biological replicates.

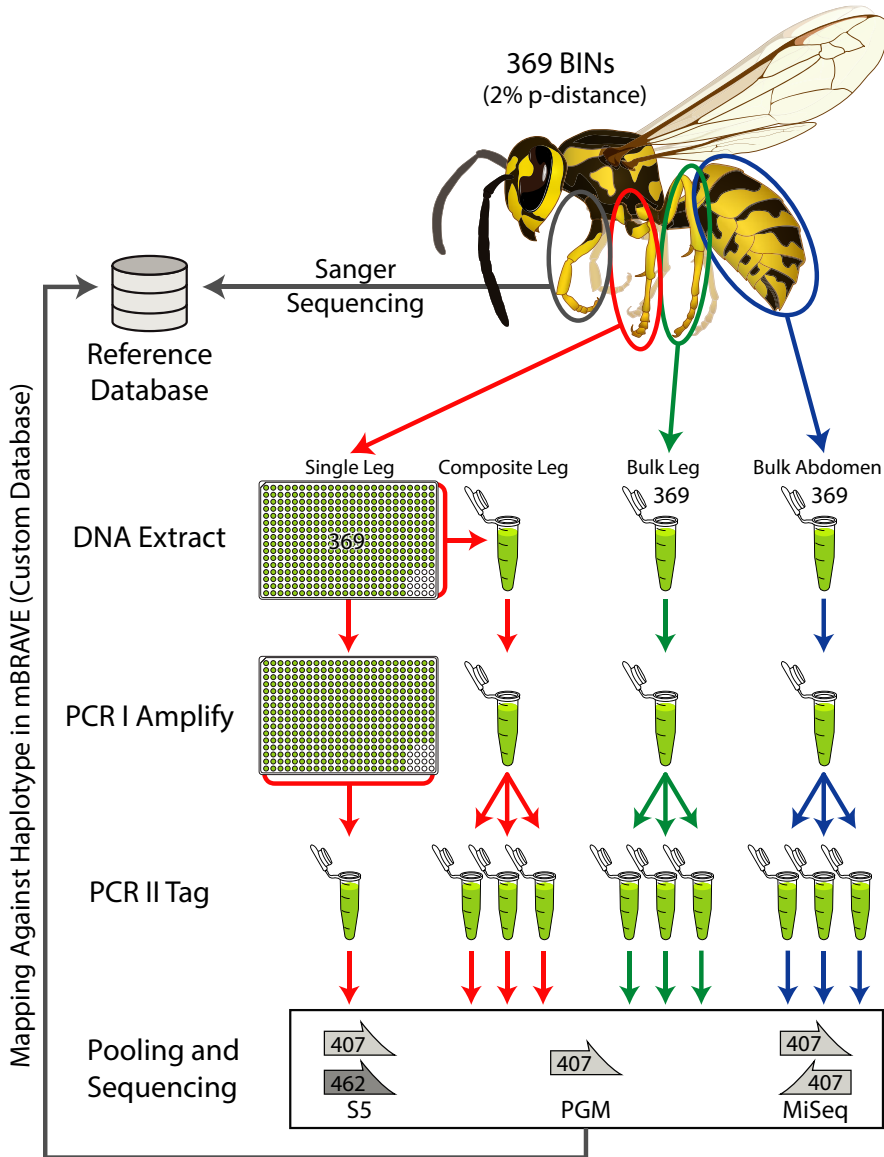


FIGURE 1 Protocol employed to examine species recovery from the mock community. Four amplicon pools were examined. Two derived from bulk DNA extracts (Bulk Abdomen and Bulk Leg). The others derived from DNA extracts from single legs that were either pooled (Composite Leg) or kept separate (Single Leg) 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 [Colour figure can be viewed at wileyonlinelibrary.com]

2.3 | Bulk DNA extractions and PCR

Prior to mock community assembly, the dry abdominal weight of each specimen was measured as a proxy for biomass. DNA extracts for the two bulk samples (Bulk Abdomen and Bulk Leg) were generated with a modified silica membrane-based protocol (Ivanova et al., 2006). Specifically, the bulk abdomens (combined mass = 1,062.8 mg) and bulk legs (combined mass = 30.9 mg) were lysed overnight in the same relative volume of insect lysis buffer (51.6 ml and 1.5 ml, respectively) with 10 mg/ml of Proteinase K (Invitrogen). Following lysis, a 100- μ l aliquot of each lysate was mixed with 200 μ l of binding mix and transferred to an EconoSpin® column (Epoch Life Sciences) before centrifugation at 5,000 g for 2 min. The DNA extracts were then purified with three wash steps. The first wash employed 300 μ l of protein wash buffer before centrifugation at 5,000 g for 2 min. Columns were then washed twice with 600 μ l of wash buffer before being centrifuged at 5,000 g for 4 min. Columns were transferred to clean tubes and spun dry at 10,000 g for 4 min to remove any residual

buffer, then transferred to clean collection tubes and incubated for 30 min at 56°C to dry the membrane. DNA was subsequently eluted by adding 50 μ l of 10 mM Tris-HCl pH 8.0 followed by centrifugation at 10,000 g for 5 min. All DNA extracts were normalized to 3 ng/ μ l prior to PCR. All PCR reactions were composed of 5% trehalose (Fluka Analytical), 1 \times Platinum Taq reaction buffer (Invitrogen), 2.5 mM MgCl₂ (Invitrogen), 0.1 μ M of each primer (Integrated DNA Technologies), 50 μ M of each dNTP (KAPA Biosystems), 0.15 units of Platinum Taq (Invitrogen), 1 μ l of template and HyClone® ultra-pure water (Thermo Scientific) for a final volume of 6 μ l.

2.4 | Construction of HTS libraries

Two rounds of PCR were used to generate the amplicon libraries destined for sequence characterization on the three platforms. Most first-round reactions employed a primer cocktail targeting a 407 or 421 bp region of COI, subsequently collectively referred to as the 407 bp amplicon. The 407 bp amplicon was generated

using MLepF1 (Hebert et al., 2004) while the 421 bp region was generated by using RonMWASPdegen (Smith et al., 2012) as forward primers; LepR1 (Hebert et al., 2004) and HCO2198 (Folmer, Black, Hoeh, Lutz, & Vrijenhoek, 1994) were used as reverse primers (Supporting Information Table S2). These primers have demonstrated high effectiveness in recovering this segment of the DNA barcode region from diverse lineages of arthropods for Sanger-based sequencing. They have the advantage of generating an amplicon that is long enough to provide good taxonomic resolution, but short enough to allow characterization on second-generation HTS platforms. An alternate first-round PCR targeted a 463 bp amplicon of COI; it was generated with a different forward primer—AncientLepF3 (Prosser, deWaard, Miller, & Hebert, 2016) (Supporting Information Table S2) and sequenced on the S5 platform. Both amplicons employed in this study are longer than those often used in environmental DNA studies, but they have the advantage of maximizing taxonomic resolution and carry no disadvantage when dealing with nondegraded DNA.

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 2 × with dd H₂O. 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 (Supporting Information 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. Supporting Information Table S2 provides all primer sequences and details on sample indexing. For both PCRs, negative controls were used and checked for the lack of a PCR product by visually inspecting an agarose gel.

For each platform, the UMI-labelled reaction products were pooled prior to sequencing. The two Ion Torrent platforms, the PGM and the S5, differ in their workflows, chemistries and read output. As the S5 is the newer platform, it has a higher read output and generates longer reads (up to 600 bp). 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.

2.5 | Bioinformatics and analysis

All read libraries were uploaded to mBRAVE (Multiplex Barcode Research and Visualization Environment) an online platform for analysing and visualizing metabarcoding data (<http://mbrave.net/>). 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%. The quality value (QV) of each sequence was evaluated, and all records failing to meet any one of three quality standards were discarded: (a) mean QV < 20; (2) >25% of bp with QV < 20; and (3) >5% of bp with QV < 10. All reads were trimmed to 407 or 463 bp following a 30 bp trim at the 5' end to remove the forward primer. Reads shorter than 300 bp were discarded for the 407 bp amplicons while a 350 bp threshold was used for the 463 bp amplicon. Retained sequences were viewed as matching a BIN in the custom Sanger library if their distance was < 3% to any reference, a commonly employed threshold (Edgar, 2013). Any reads not matching the Sanger reference library were subsequently queried against four other reference libraries (bacteria, noninsect arthropods, nonarthropod invertebrates and insects). All reads not matching any reference sequence were clustered at an OTU threshold of 2%. Standard analytical parameters were used for all treatments and sequencing platforms. The three replicates for the Bulk Abdomen, Bulk Leg and Composite Leg treatments were also pooled for comparison with the technical replicates.

OTU tables for each run were merged in R v3.4.4 (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 count intervals (10^2 , 10^3 , $10^{3.5}$, 10^4 , $10^{4.5}$, 10^5 , $10^{5.5}$ and 10^6). Sequential points with a slope of < 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 > 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 was 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 nonmetric multidimensional scaling (NMDS) with `vegan`.

The relationships between read counts and body size, as measured by abdominal mass, and between read count and GC content of the COI amplicon were examined using Kendall Tau correlations in R v3.4.4 (R Core Team, 2018). An analysis of similarity (ANOSIM) with 999 permutations was used to compare species recovery among treatment types, sequencing platforms and between the two

TABLE 1 Summary of run results for all treatments. mBRAVE filtering and BIN recovery including false positives are indicated for the four amplicon pools [Colour table can be viewed at wileyonlinelibrary.com]

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)	unmatched 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
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
	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	

Note. 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 reads not matching the direct reference library 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.

amplicons with the R package VEGAN v2.5-1 (Oksanen et al., 2018). All custom scripts are available as supplementary materials.

The relationship between the read count for each BIN and primer mismatches were investigated for the 407 and 463 bp amplicons. The number of mismatches was quantified by counting the number of nucleotide substitutions between the primer sequence and the template DNA for each BIN. Information on the DNA sequence for the forward primer binding sites was available from the Sanger reads for all 369 BINs. Calculation of mismatches was straightforward for the 463 bp amplicon as it involved a single forward primer. As the 407 bp amplicon was generated with two different forward primers, mismatches were quantified based upon the forward primer with the best match to the template for each BIN. The same two reverse primers were employed to generate the 407 and 463 bp amplicons, but DNA sequence information for template DNA was not available from the Sanger sequence (as it was based on amplicons generated with the same reverse primer). As a result, an alternate reverse primer, C1-N-2395d (Simon et al., 1994), was employed to extend each sequence in the 3' direction, an approach which delivered the desired sequence information for 203 of the 369 BINs. As a consequence, it was possible to examine the relationship between read counts and the number of mismatches between template and forward primer for all 369 BINs and the total mismatch count for the forward and reverse primers for the 203 BINs with template sequences for both regions.

3 | RESULTS

3.1 | Run quality

We first compared the output and quality of the reads from the HTS platforms. The S5 and MiSeq generated a similar number of reads

(~1 million per replicate), while the PGM generated substantially fewer (~450,000 per replicate). 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 Torrent platforms, reflecting their near consistent QV versus the decline towards the 3' end of the PGM and S5 reads (Supporting Information Figure S1).

3.2 | 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; Supporting Information Figure S2). 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 ($10^{4-4.5}$ vs. $10^{4.5-5}$ – Supporting Information 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 versus 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 (Supporting Information 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 (Supporting Information Table S1). Pielou's evenness, Simpson's Index, Inverse Simpson's Index, Renyi's diversity and Shannon Indices were also similar across

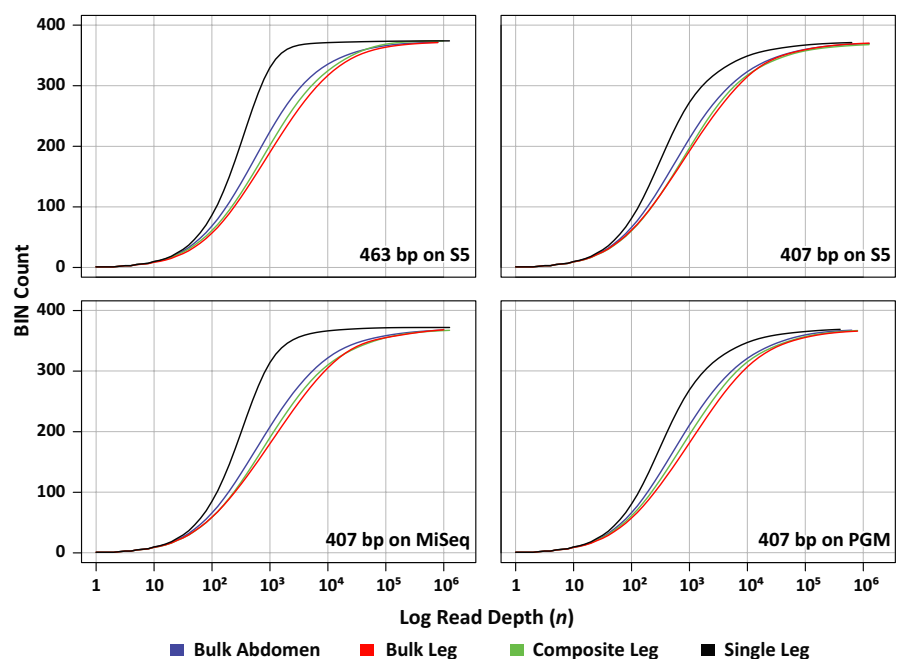


FIGURE 2 Rarefaction curves showing BIN recovery versus the number of sequences analysed for the four amplicon pools (Bulk Abdomen, Bulk Leg, Composite Leg and Single Leg) on the three sequencing platforms. Two amplicon lengths (407 and 463 bp) were analysed on the S5, but just one (407 bp) on the other platforms [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Values for selected diversity indices (Shannon–Weaver, Simpson, Inverse Simpsons and Pielou's Evenness) for the four amplicon pools [Colour table can be viewed at wileyonlinelibrary.com]

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

Note. 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.

treatments on all platforms (Table 2; Supporting Information Figure S3). Finally, density plots were congruent among technical replicates for all treatments and platforms indicating that different HTS platforms produced similar results (Figure 1; Supporting Information Figure S4).

3.3 | BIN recovery

When the criterion for BIN recovery was set at one or more reads, all platforms recovered > 98% of the BINs, but only the Single Leg treatment recovered all of them (Figure 3). Differences in recovery success among treatments were greater when the criterion for recovery was set at > 0.01% of the reads. Under this criterion, the Single Leg treatment recovered > 92.5% of the BINs versus 83%–89% for the Bulk Abdomen treatment and 76%–83% for the Composite Leg and Bulk Leg treatments (Table 1). The greater evenness in read count for the Single Leg treatment was striking; it led to lower coefficients of variation, higher diversity indices and Pielou's evenness (Table 2; Figure 3; Supporting Information Figure S3). Density plots of read abundance also demonstrated much higher evenness for the Single Leg treatment, especially for the 407 bp amplicon on the MiSeq and for the 463 bp amplicon on the S5 (Supporting Information Figure S4). These differences were also reflected in BIN recovery, Pielou's evenness and diversity indices (Table 2; Supporting Information Table S2).

3.4 | BIN abundances

Because a single specimen of each BIN was included in the mock community, the proportion of sequences from each should, in the absence of bias, be similar across sequencing platforms, amplicons and treatments. In practice, the relative abundances of the BINs varied markedly. Perceived abundance of the 369 taxa based on their read counts varied more than 11,000-fold for the Bulk Abdomen, Bulk Leg and Composite Leg treatments, and 4,000-fold for the Single Legs. A single-link dendrogram based on Bray–Curtis dissimilarity values indicated that samples clustered first by treatment, next by amplicon length and finally by sequencing platform (Figure 4a). An analysis of similarity using Bray–Curtis distances affirmed significant differences in BIN abundances by treatment type ($p = 0.001$, $R = 1$), amplicon length ($p = 0.027$, $R = 0.17$), but not by sequencing platform ($p = 0.13$, $R = 0.037$) (Figure 4b; Supporting Information Figure S5).

3.5 | Primer mismatches and read count

Examination of the relationship between the read count for each of the 369 BINs and its number of mismatches from the forward primer revealed a strong negative relationship. BINs with a high mismatch count were typically represented by few reads. For example, very few reads were recovered from the only BIN in the order Dermoptera and this was associated with a high mismatch Index from the forward primers for both the 407 and 463 bp amplicons.

Considering all taxa, BIN recovery was substantially higher for the 463 bp amplicon than for the 407 bp amplicon (Table 2; Supporting Information Figure S3) reflecting the fact that its forward primer better matched the template DNA (18 BINs had > 3 mismatches) versus the forward primer for the 407 bp amplicon (62 BINs with > 3 mismatches) (Supporting Information Tables S1 and S4). The impact of these mismatches was clear; read count and relative abundance of BINs declined after two mismatches for the Bulk Abdomen, Bulk Leg and Composite Leg treatments and after four mismatches for the Single Leg. Examination of the joint impact of forward and reverse primer mismatches for 203 BINs similarly showed a significant decline in read count and relative abundance after four mismatches for the Bulk Abdomen, Bulk Leg and Composite Leg treatments and after seven mismatches for the Single Leg (Supporting Information Figure S6). Kruskal–Wallis tests showed that read depth declined significantly with an increasing number of primer mismatches for the forward primers for both the 407 and 463 bp amplicons ($p < 0.0001$) and for the summed primer mismatches ($5' + 3'$) for the subset of 203 BINs ($p < 0.0001$).

3.6 | Impacts of biomass and nucleotide composition on read count

Other factors also explained some of the variation in read counts among BINs. There was, for example, a weak negative correlation ($r^2 < 0.10$) between the GC content of an amplicon and its read count, excepting the Single Leg treatment on the MiSeq where it was higher ($r^2 = 0.32$) (Supporting Information Figure S7). A weak positive correlation ($r^2 = 0.24$ – 0.28) was also apparent between the abdominal mass of a BIN and its read count on all platforms (Supporting Information Figure S8).

3.7 | Nontarget sequences

Each run recovered some sequences with substantial sequence divergence from the Sanger reference library (Table 1). The incidence of nontarget sequences for the 407 bp amplicon was slightly lower (4%–6%) on the PGM and S5 platforms than on the MiSeq (8%–10%). Interestingly, the 463 bp amplicon had substantially more nontarget reads (15%–17%). After excluding the relatively few chimeras (1.5%–12.5%), half the reads from the nontarget sequences failed to match any sequence in the supplemental libraries. Of those that did find a match, most were arthropods.

3.8 | Taxonomic bias

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

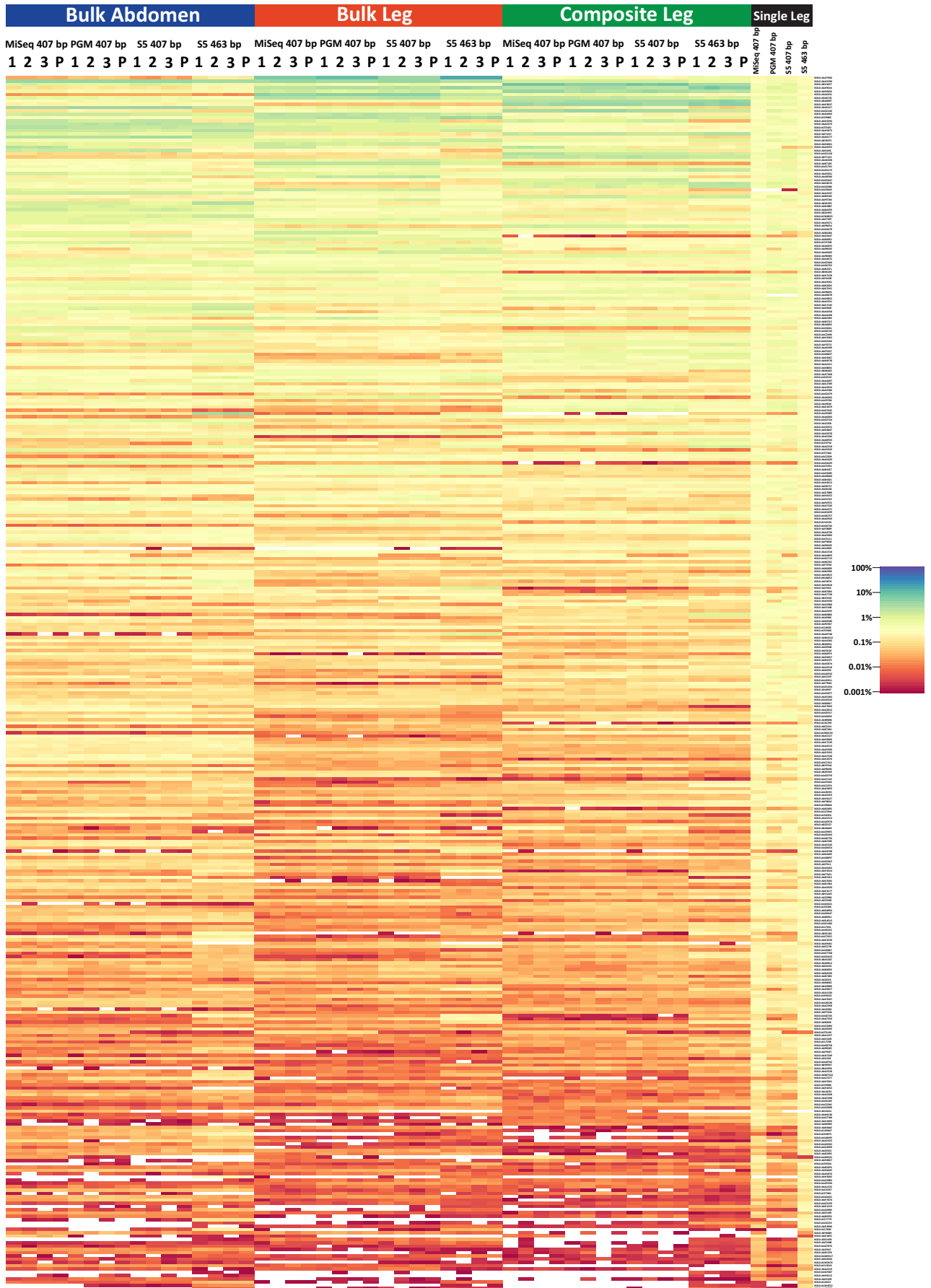


FIGURE 3 Heat map showing the relative log abundance of the 369 BINs in each treatment for the four amplicon pools. This heat map was created using the JAMP package (<https://github.com/VascoElbrecht/JAMP>). Technical replicates are indicated with numbers while in silico pooled results are designated by the letter P [Colour figure can be viewed at wileyonlinelibrary.com]

amplicon (Supporting Information Table S5). Among the bulk samples, relative abundance differed among treatments. For example, the relative abundance of Lepidoptera and Mecoptera was lower, while Diptera and Orthoptera were higher in the Composite Leg than in the Bulk Leg and Bulk Abdomen treatments. The proportion of read counts for Trichoptera showed particularly large variation, being 5–25X higher for the Bulk Leg than the Bulk Abdomen and Composite Leg treatments across all platforms and for both amplicons.

4 | DISCUSSION

Metabarcoding is a powerful tool for characterizing biodiversity patterns (Cristescu, 2014), but data interpretation is complicated

by several factors. PCR amplification bias and variation in the copy number of template DNA from the source specimens not only make it impossible to estimate abundances, but can impede the recovery of all species (Beng et al., 2016; Elbrecht & Leese, 2015; Ji et al., 2013; Yu et al., 2012). Although prior studies have revealed these complexities, there has been limited evaluation of the strength of their influence on interpretations of taxon diversity. To address this gap, the present study examined the impact of diverse factors including source DNA, PCR primers, sequencing platform and sequencing depth on species recovery from a diverse assemblage of insects. Both primer sets demonstrated their effectiveness for metabarcoding as they recovered > 98% of the 374 species in a taxonomically diverse mock community (10 orders, 104 families). However, it did require substantial sequence coverage to recover these species

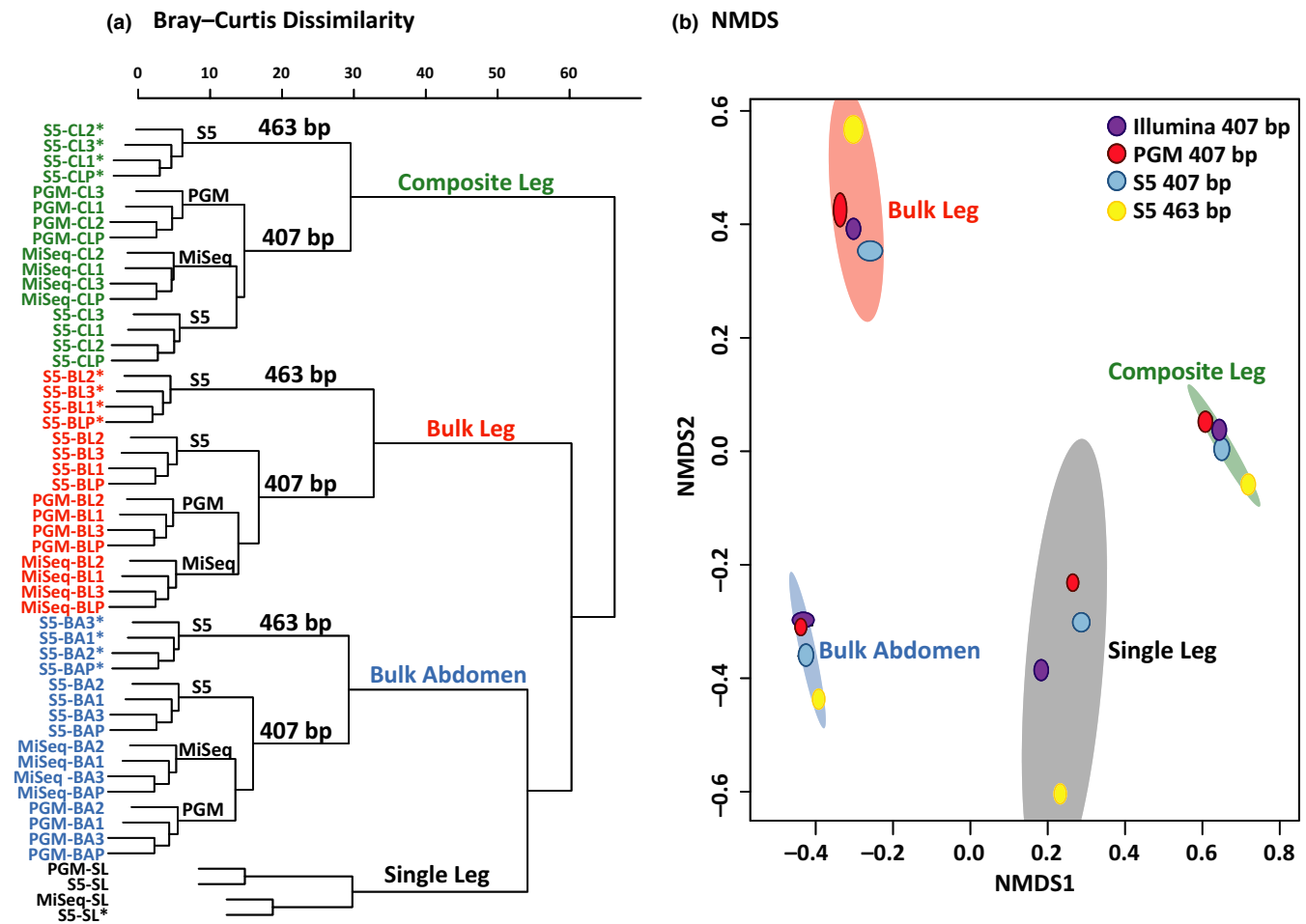


FIGURE 4 (a) Bray-Curtis dissimilarity dendrogram for the four amplicon pools (BA = Bulk Abdomen, BL = Bulk Leg, CL = Composite Leg and SL = Single Leg). Replicates are numbered 1–3 while P is the result from pooling the replicates. The 463 bp amplicon is indicated with an asterisk (*). (b) Nonmetric 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 [Colour figure can be viewed at wileyonlinelibrary.com]

because of their varied amplicon abundance. Variation in body size of the species and shifts in the GC composition of their COI templates partially accounted for the divergence in amplicon abundance, but the mismatch count between primers and template DNA had a greater effect. While new primer sets can be designed to reduce such mismatches, it will never be possible to entirely escape them in any large assemblage of phylogenetically diverse species.

4.1 | Sequencing depth

Because of diversity in amplicon abundance among taxa, sequencing depth has a strong impact on taxon recovery and hence perceived diversity patterns (Leray & Knowlton, 2015; Leray & Knowlton, 2017). Species with low representation in the amplicon pool, because of primer–template mismatches or low template concentrations linked to rarity or small body size, are likely to be missed unless sequencing depth is very high. When species are overlooked, alpha diversity is underestimated, and beta diversity is exaggerated (Bellemain et al., 2012; Sickle et al., 2015; Sickle et al., 2015; Yamamoto et al., 2017). There is a simple way to assess whether sequencing effort has been adequate; the slope of the rarefaction curve is zero when all species have been recovered (Lanzen, Lejang, Jonassen, Thompson, & Troedsson, 2017). Technical replicates are also useful because every replicate should include the same OTUs when sequence coverage is adequate. Although taxon richness was fixed in our study, we employed a slope for the rarefaction curve of < 0.01 as the criterion to decide whether taxon diversity had achieved an asymptote. This criterion has the advantage of being applicable in situations where the species count is unknown, as is regularly the case in nature. Based upon this criterion, we detected up to 100-fold differences in the level of sequencing that was required to achieve an asymptote among the four treatments and two amplicons examined in this study.

BIN accumulation curves indicated that the read depth employed in this study allowed all four treatments to meet 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. 2018; Pan et al., 2014, Dabney & Meyer, 2012; Elbrecht & Leese, 2015). Interestingly, the other three treatments showed similar BIN accumulation curves on all three sequencers, suggesting shared factors constrain BIN recovery.

4.2 | Sequencing platforms

The three sequencing platforms generated similar estimates of BIN diversity. However, results from the MiSeq had advantages over those from the PGM and S5 for both clustering algorithms and for haplotype analysis (Elbrecht, Varnos, Steinke, & Leese, 2018) reflecting its delivery of full-length, higher fidelity reads. In particular, the paired-end protocol consistently recovered sequences for the full 407 bp amplicon, while those from the PGM and S5 were often truncated and possessed more indels. Finally, the MiSeq reads had consistently higher mean QV. Because these factors simplified data analysis (Edgar et al., 2013; Mardis et al., 2013) and sequencing costs

were similar, the MiSeq is currently the best platform for metabarcoding (Mardis et al., 2013).

4.3 | Impacts of analytical protocols

Our four treatments made it possible to compare the impact of targeting different tissues, employing different DNA extraction regimes and using different PCR protocols. Despite their similar tissue input and DNA extraction regime, the Single Leg treatment achieved asymptotic diversity much more rapidly than the Composite Leg treatment, indicating how separate PCR reactions reduce amplification bias. By contrast, BIN accumulation curves and diversity indices for the Composite Leg treatment were similar to those for the Bulk Leg and Bulk Abdomen, indicating that DNA extraction was equally effective whether carried out on single specimens or on bulk samples (Table 2). The comparison of the results for the bulk/composite samples did reveal more nuanced differences as the number of reads for particular taxa varied among these three treatments despite similar BIN recovery profiles (Figure 3). These differences likely stem from differential leg/abdomen mass ratios among species which led to varied mitochondrial copy numbers for the component species among treatments. Certainly, mitochondrial copy number varies among tissues and among species (Cole, 2016; Veltri, Espiritu, & Singh, 1990). Future efforts to explore this relationship and its importance to metabarcoding studies should quantify copy number differences between tissues and species. In the absence of such information, copy number bias due to biomass or species differences can be reduced by partitioning the specimens in a bulk sample into size fractions (Elbrecht, Peinert et al., 2017; Vivien et al., 2016).

Variation in read counts for the taxa in any bulk sample is, as already noted, strongly influenced by primer–template mismatches. Although degenerate primers (Elbrecht & Leese, 2017; Moriniere et al., 2016; Yu et al., 2012) and improved primer sets (Clarke et al., 2014; Elbrecht & Leese, 2017; Leray & Knowlton, 2017) can reduce such bias, it cannot be avoided unless all target species possess identical sequences for the primer binding sites, a condition that will never be satisfied for a large assemblage. However, efforts to target highly conserved regions can improve the situation. For example, the BIN accumulation curve for the 463 bp amplicon reached its asymptote with much lower read coverage than for the 407 bp. Further effort to develop primers that maximize primer–template matches for diverse taxa will reduce the sequencing effort needed to recover all taxa. So too will strategies that minimize mismatches by partitioning bulk samples into major taxonomic groups (Bellemain et al., 2012; Cristescu, 2014; Moriniere et al., 2016; Tedersoo et al., 2015). Lowering the variation in recovery success linked to size differences among species can be achieved by partitioning the species present in bulk samples into subsets with similar size (Elbrecht, Peinert et al., 2017; Moriniere et al., 2016; Vivien et al., 2016). Currently, the only means to fully escape the varied factors influencing sequence recovery is to

process specimens individually through the entire analytical chain from DNA extraction to sequencing but this approach is too costly for large biodiversity surveys (Ji et al., 2013).

4.4 | BIN recovery

Although sequences were recovered from most of the BINs in each treatment, this outcome shifted when recovery success was defined as only those BINs comprising $> 0.01\%$ of the read count, a criterion often employed to exclude low-frequency sequences that are chimeras, contaminants or sequencing errors (Leray & Knowlton, 2017). Under this criterion, BIN recovery was substantially higher ($>92.5\%$) for the Single Leg treatment than for the other three (76%–89%). 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 (Figure 3) (Cole, 2016; Veltri et al., 1990). As expected, BIN recovery was more efficient for the 463 bp than the 407 bp amplicon because of its higher primer–template correspondence. There was also less taxonomic bias (Table 2) for the 463 bp than for 407 bp amplicon in three treatments (Bulk Abdomen, Composite Leg and Single Leg).

4.5 | False positives, negatives and unmatched OTUs

Although most BINs were recovered in each treatment, some comprised $< 0.01\%$ of the counts, creating false negatives that would underestimate alpha diversity. As in other metabarcoding studies (Vivien et al., 2016; Ficetola et al., 2015; Brandon-Mong et al., 2015; Port et al., 2016), false positives were also encountered, likely reflecting eDNA associated with specimens or contamination during sample processing (Port et al. 2016) or NUMTs (Song et al., 2008). Their impact can be reduced by employing curated reference libraries to discriminate sequences that derive from known species versus those that represent pseudogenes (Bergsten et al., 2014; Braukmann, Kuzmina, Sills, Zakharov, & Hebert, 2017; Hebert et al., 2003; Landi et al., 2014; Zimmerman et al., 2014). As well, negative controls make it possible to identify reads that derive from contamination events during sample processing (Port et al. 2016).

Although we expected the Bulk Abdomen treatment to generate more nontarget sequences than the others, reflecting template DNA from the digestive tract, this was not the case. In retrospect, it seems likely that DNA molecules from this source were too degraded to be recovered via the 407 bp and 463 bp amplicons. Certainly, most metabarcoding studies on gut contents or faecal samples have employed shorter amplicons (Hajibabaei, Spall, Shokralla, & Konynenburg, 2012; Kartzin et al. 2015; Linard, Arribas, Andújar, Crampton-Platt, & Vogler, 2016). Because legs have a much higher surface area to volume ratio than abdomens,

they may bind more eDNA, leading to their slightly higher recovery of nontarget DNA.

Approximately 1%–4% of the filtered reads did not match any sequence in the reference library. This varied slightly by platform with the MiSeq showing fewer unmatched reads (1.07%–1.73%) than the S5 (1.64%–2.50%) and PGM (2.06%–3.11%). Similarly, the 407 bp amplicon on the S5 had fewer nonmatching reads (1.64%–2.50%) than the 463 bp amplicon (1.84%–4.10%) on this platform. Some of the unmatched reads are undoubtedly derived from NUMTs, PCR errors and sequencing errors. Although we did not evaluate the incidence of pseudogenes, they likely represent some of the highly divergent unidentified OTUs (Leray et al., 2015). Their frequency among terrestrial arthropods needs to be further explored through specimen-based analysis. The Miseq produced the highest quality reads, suggesting that the higher incidence of unmatched reads on the Ion Torrent platforms reflect, in part, sequencing errors, especially the PGM which has a steep decline in quality towards the 3' end (Supporting Information Figure S1). PCR errors introduced by the DNA polymerase can be exacerbated by sequencing errors (Dabney & Meyer, 2012; Nichols et al., 2018; Pan et al., 2014). Because long amplicons improve taxonomic resolution, their use should be standard for metabarcoding studies unless template DNA is degraded.

5 | FUTURE METHODS AND CONCLUSIONS

This study has established that current PCR-based protocols for metabarcoding can recover most species in a diverse assemblage of insects when sequencing depth is adequate. Specifically, it required from 100,000 to 500,000 reads (300x–1,000x average read depth) to recover $> 95\%$ of the 374 species from bulk DNA extracts. Although the choice of sequencing platform had little impact on final results, the higher quality of sequences generated by Illumina MiSeq simplified data analysis. Given current analytical costs (Supporting Information Table S6), there is a clear justification to search for protocols that make it possible to reveal the species in any assemblage with limited sequencing effort. The importance of employing primer sets that minimize mismatches with template DNA was evidenced in this study by the fact that a tenth as many sequences were required to recover 95% of the species for the 463 bp than the 407 bp amplicon. Given such impacts, further effort to design primers which minimize mismatches for DNA extracts from diverse taxonomic assemblages (e.g. zooplankton, insects) are important. Success should be facilitated because 3rd generation sequencers can analyse longer amplicons, permitting the use of primer sets that target regions of COI where sequences are constrained because they code for amino acids that bind substrates or cofactors. However, because sequence variation is inevitable in any diverse taxonomic assemblage, the copy number of sequences in the amplicon pool will diverge from their abundance in the original DNA extract.

To escape such bias, PCR-free protocols have been proposed (Liu et al., 2016, Tang et al., 2014). In their simplest implementation, they involve the random analysis of genomic fragments, followed by the exclusion of reads that fail to match a sequence in the reference library. This approach means that many sequences do not contribute to a species assignment. For example, if analysis focused on COI-5', approximately 1 in every 2,000 sequences would be retained because it represents just 5% of the mitochondrial genome and mitochondrial DNA represents just 1% of the total DNA in a cell. Hence, to recover a single barcode sequence from 374 species, presuming an identical number of COI templates for each taxon and sampling one copy per species, 750,000 sequences would be needed. In practice, far more sequences would be required to overcome the impacts of random sampling and body size. To recover 95% of the species in the mix, assuming random sampling and equal template count for each species would require an average of ~ 3 sequences per species, based on $N \sum_{i=0}^k \frac{1}{N-i-1}$, a variation of the coupon-collector's problem (Motwani & Raghavan, 1995) where N is the number of species and k is the recovered subset, raising the required number of sequences to 2.25 million. Further sequencing would be needed to compensate for the variation in template numbers linked to body size variation. The species of arthropods examined in this study varied 7,500-fold in body mass. Assuming that mitochondrial copy number and metabolic rate scale in a similar fashion with body mass^{0.66} (Burgess et al., 2017), the largest species examined in this study should have possessed about 360 times more target template than the smallest species. Given of this difference, some 810 million sequences would be needed to recover 95% of the species in the assemblage. More conservatively, if specimens with the lowest 5% body mass are excluded, there is a 178-fold difference between the smallest and largest species. The largest species will have 30 times more target template, requiring 67.5 million reads to recover the species in the assemblage. Since the simplest implementation of PCR-free approaches is so inefficient, more advanced protocols employ baits to enrich for the target gene region (Dowle, Pochon, C. Banks, Shearer, & Wood, 2015), but they can create interpretational complexities when capture efficiencies vary among taxonomic groups.

It is important to emphasize that no current metabarcoding protocol allows the estimation of species abundances. PCR-based methods fail because of distortions in the amplicon pool introduced by variation in primer binding. PCR-free methods can reveal the abundance of each template in the total DNA extract, but this count of COI molecules cannot predict species abundance because a high value might derive from a few adults or many juveniles. Given these barriers to abundance assessment through metabarcoding, it is worth noting that specimen-based analysis can deliver this information with very limited sequencing effort. For example, just 374 Sanger reads or 3,000 reads of a UMI-tagged DNA pool on the Sequel platform (Hebert et al., 2018) would have revealed the presence and equal abundance of each species in the current sample. Although neither PCR-based nor PCR-free metabarcoding can deliver accurate information on species abundance, the two approaches can deliver

complementary information on species composition. Because the amplicon pool generated by PCR is influenced by variation in primer binding, the impact of variation in body size is diminished since small species whose COI template closely matches primer sequences will be well represented in the amplicon pool. By contrast, PCR-free methods can aid the recovery of large species that can be overlooked in PCR-based studies because of poor amplification.

In summary, this study has established that PCR-based metabarcoding provides a cost-effective way to recover information on the species composition of insect communities because current COI primer sets are broadly effective, and a well-provisioned reference library is available. The optimal solution may be different for other taxonomic groups, especially those where primers fail to amplify many species or where amplification bias is extreme. However, in such cases, it remains important to try to overcome these barriers rather than simply capitulating. It is also worth emphasizing that many natural communities possess greater complexity than the assemblage examined in this study; they include more species and the abundances of these species show great variation. Given these complications, community characterization through metabarcoding will often require both intensive sequencing and improved informatics support to recognize sequences that reflect rare species rather than analytical artefacts.

ACKNOWLEDGEMENTS

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

AUTHOR CONTRIBUTIONS

The study was conceived and designed by PDNH, EZ, TWAB, DS, SR, JRD, NI, and SP; The research was performed by NI, JS, SP, and TWAB; The analysis of data was performed by TWAB and VE; Analytical tools were contributed by SR and VE; The manuscript was written by TWAB and PDNH with input and revisions from DS, SR, VE, EZ, NI, SP, and JS.

DATA ACCESSIBILITY

Further details on the treatments are available at the following DOIs: Bulk Abdomen and Bulk Leg: <https://doi.org/10.5883/DS-NGS375A>; Composite Leg and Single Leg: <https://doi.org/10.5883/DS-NGS375B>. All raw HTS sequence data is available in NCBI's Short Read Archive (SRP158933).

ORCID

Thomas W. A. Braukmann  <https://orcid.org/0000-0002-2452-3776>

Vasco Elbrecht  <https://orcid.org/0000-0003-4672-7099>

Dirk Steinke  <http://orcid.org/0000-0002-8992-575X>

REFERENCES

- Aas, A. B., Davey, M. L., & Kausserud, H. (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. <https://doi.org/10.1111/1755-0998.12622>
- Andújar, C., Arribas, P., Yu, D. W., Vogler, A. P., & Emerson, B. C. (2018). Why the COI barcode should be the community DNA metabarcode for the metazoa. *Molecular Ecology*, *27*, 3968–3975. <https://doi.org/10.1111/mec.14844>
- Bassett, Y., Cizek, L., Cuénoud, P., Didham, R. K., Guilhaumon, F., Missa, O., ... Leponce, M. (2012). Arthropod diversity in a tropical forest. *Science*, *338*, 1481–1484. <https://doi.org/10.1126/science.1226727>
- Bellemain, E., Davey, M. L., Kausserud, H., Epp, L. S., Boessonkool, S., Coissac, E., ... Brochmann, C. (2012). Fungal palaeodiversity revealed using high-throughput metabarcoding of ancient DNA from arctic permafrost. *Environmental Microbiology*, *15*, 1176–1189. <https://doi.org/10.1111/1462.29220.12020>
- Beng, K. C., Tomlinson, K. W., Shen, X. H., Surget-Groba, Y., Hughes, A. C., Corlett, R. T., & Slik, J. W. F. (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. <https://doi.org/10.1038/srep24965>
- Bergsten, J., Bilton, D. T., Fujisawa, T., Elliott, M., Monaghan, M. T., Balke, M., ... Vogler, A. P. (2014). The effect of geographical scale of sampling on DNA barcoding. *Systematic Biology*, *61*, 851–869. <https://doi.org/10.1093/sysbio/sys037>
- Brandon-Mong, G. J., Gan, H. M., Sing, K. W., Lee, P. S., Lim, P. E., & Wilson, J. J. (2015). DNA metabarcoding of insects and allies: An evaluation of primers and pipelines. *Bulletin of Entomological Research*, *105*, 717–727. <https://doi.org/10.1017/S0007485315000681>
- Braukmann, T. W. A., Kuzmina, M. L., Sills, J., Zakharov, E. V., & Hebert, P. D. N. (2017). Testing the efficacy of DNA barcodes for identifying the vascular plants of Canada. *PLoS ONE*, *12*, E0169515. <https://doi.org/10.1371/journal.pone.0169515>
- Burgess, S. C., Ryan, W. H., Blackstone, N. W., Edmunds, P. J., Hoogenboom, M. O., Levitan, D. R., & Wulff, J. L. (2017). Metabolic scaling in modular animals. *Invertebrate Biology*, *136*, 456–472. <https://doi.org/10.1111/ivb.12199>
- Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., ... Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*, 335–336. <https://doi.org/10.1038/nmeth.f.303>
- Clarke, L. J., Soubrier, J., Weyrich, L. S., & Cooper, A. (2014). Environmental metabarcodes for insects: In silico PCR reveals potential for taxonomic bias. *Molecular Ecology Resources*, *14*, 1160–1170. <https://doi.org/10.1111/1755-0998.12265>
- Cole, L. W. (2016). The evolution of per-cell organelle number. *Frontiers in Cell and Developmental Biology*, *4*, 85. <https://doi.org/10.3389/fcell.2016.00085>
- R Core Team (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>
- Cristescu, M. E. (2014). From barcoding single individuals to metabarcoding biological communities: Towards an integrative approach to the study of global biodiversity. *Trends in Ecology and Evolution*, *29*, 566–571. <https://doi.org/10.1016/j.tree.2014.08.001>
- Dabney, J., & Meyer, M. (2012). Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *BioTechniques*, *52*, 87–94. <https://doi.org/10.2144/000113809>
- Deagle, B. E., Jarman, S. N., Cossiac, E., Pompanon, F., & Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: Not a perfect match. *Biology Letters*, *10*, 20140562. <https://doi.org/10.1098/rsbl.2014.0562>
- Divoll, T. J., Brown, V. A., Kinne, J., McCracken, G. F., & O'Keefe, J. M. (2018). Disparities in second-generation DNA metabarcoding results exposed with accessible and repeatable workflows. *Molecular Ecology Resources*, *18*, 590–601. <https://doi.org/10.1111/1755-0998.12770>
- Dowle, E. J., Pochon, X., C. Banks, J., Shearer, K., & Wood, S. A. (2015). Targeted gene enrichment and high-throughput sequencing for environmental biomonitoring: A case study using freshwater macroinvertebrates. *Molecular Ecology*, *16*, 1240–1254. <https://doi.org/10.1111/1755-0998.12488>
- Edgar, R. C. (2013). UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, *10*, 996–998. <https://doi.org/10.1038/nmeth.2604>
- Elbrecht, V., & Leese, F. (2015). Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass-sequence relationships with an innovative metabarcoding protocol. *PLoS ONE*, *10*, e0130324. <https://doi.org/10.1371/journal.pone.0130324>
- Elbrecht, V., & Leese, F. (2017). Validation and development of freshwater invertebrate metabarcoding COI primers for Environmental Impact Assessment. *Frontiers in Environmental Science*, *5*, 11. <https://doi.org/10.3389/fenvs.2017.00011>
- Elbrecht, V., Peinert, B., & Leese, F. (2017). Sorting things out: Assessing effects of unequal specimen biomass on DNA metabarcoding. *Ecology and Evolution*, *7*, 6918–6926. <https://doi.org/10.1002/ece3.3192>
- Elbrecht, V., Taberlet, P., Dejean, T., Valentini, A., Usseglio-Polatera, P., Beisel, J.-N., ... Leese, F. (2016). Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects. *PeerJ*, *4*, e1966. <https://doi.org/10.7717/PeerJ.1966>
- Elbrecht, V., Vamos, E. E., Meissner, K., Aroviita, J., & Leese, F. (2017). Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. *Methods in Ecology and Evolution*, *8*, 1265–1275. <https://doi.org/10.1111/2041-210X.12789>
- Elbrecht, V., Varnos, E. E., Steinke, D., & Leese, F. (2018). Estimating intra-specific genetic diversity from community DNA metabarcoding data. *PeerJ*, *6*, e4644. <https://doi.org/10.7717/PeerJ.4644>
- Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., ... Taberlet, P. (2015). Replication levels, false presences, and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, *15*, 543–556. <https://doi.org/10.1111/1755-0998.12338>
- Folmer, O., Black, M., Hoeh, W., Lutz, R., & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, *3*, 294–299.
- Hajibabaei, M., deWaard, J. R., Ivanova, N. V., Ratnasingham, S., Dooh, R. T., Kirk, S. L., ... Hebert, P. D. N. (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. <https://doi.org/10.1098/rstb.2005.1727>
- Hajibabaei, M., Shokralla, S., Zhou, S., Singer, G. A. C., & Baird, D. J. (2011). Environmental barcoding: A next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE*, *6*, e17497. <https://doi.org/10.1371/journal.pone.0017497>

- Hajibabaei, M., Spall, J. L., Shokralla, S., & van Konynenburg, S. (2012). Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. *BMC Ecology*, 12, 28. <https://doi.org/10.1186/1472-6785-12-28>
- Hebert, P. D. N., Braukmann, T. W. A., Prosser, S. W. J., Ratnasingham, S., deWaard, J. R., Ivanova, N. V., ... Zakharov, E. V. (2018). A sequel to sanger: Amplicon sequencing that scales. *BMC Genomics*, 19, 219. <https://doi.org/10.1186/s12864-018-4611-3>
- Hebert, P. D. N., Cywinska, A., Ball, S. L., & de Waard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Science*, 270, 313–321. <https://doi.org/10.1098/rspb.2002.2218>
- Hebert, P. D. N., & Gregory, T. R. (2005). The promise of DNA barcoding for taxonomy. *Systematic Biology*, 54, 852–859. <https://doi.org/10.1080/10635150500354886>
- Hebert, P. D. N., Penton, E. H., Burns, J. M., Janzen, D. H., & Hallwachs, W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences of the United States of America*, 101, 14812–14817. <https://doi.org/10.1073/pnas.0406166101>
- Hernández-Triana, L. M., Prosser, S. W., Rodríguez-Perez, M. A., Chaverri, L. G., Hebert, P. D. N., & Gregory, T. R. (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, 508–518. <https://doi.org/10.1111/1755-0998.12208>
- Ivanova, N. V., de Waard, J. R., & Hebert, P. D. N. (2006). An inexpensive, automation friendly protocol for recovering high quality DNA. *Molecular Ecology Notes*, 6, 998–1002.
- Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., ... Yu, D. W. (2013). Reliable, verifiable, and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, 16, 1245–1257. <https://doi.org/10.1111/ele.12162>
- Kartzinel, T. R., Chen, P. A., Coverdale, T. C., Erickson, D. L., Kress, W. J., Kuzmina, M. L., ... Pringle, R. M. (2015). DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. *Proceedings of the National Academy of Sciences of the United States of America*, 112, 8019–8024. <https://doi.org/10.1073/pnas.1503283112>
- Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16, 111–120.
- Landi, M., Dimech, M., Arculeo, M., Biondo, G., Martins, R., Carneiro, M., ... Costa, F. O. (2014). DNA barcoding for species assignment: The case of Mediterranean marine fishes. *PLoS ONE*, 9, e106135. <https://doi.org/10.1371/journal.pone.0106135>
- Lanzen, A., Lejang, K., Jonassen, I., Thompson, E. M., & Troedsson, C. (2017). DNA extraction replicates improve diversity and compositional dissimilarity in metabarcoding of eukaryotes in marine sediments. *PLoS ONE*, 12, e0179443. <https://doi.org/10.1371/journal.pone.0179443>
- Lee, D. F., Lu, J., Chang, S., Loparo, J. J., & Xie, X. S. (2016). Mapping DNA polymerase error by single molecule sequencing. *Nucleic Acids Research*, 44, e118.
- Leray, M., & Knowlton, N. (2015). DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 112, 2076–2081. <https://doi.org/10.1073/pnas.1424997112>
- Leray, M., & Knowlton, N. (2017). Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina COI metabarcoding. *PeerJ*, 5, e3006. <https://doi.org/10.7717/peerj.3006>
- Linard, B., Arribas, P., Andújar, C., Crampton-Platt, A., & Vogler, A. P. (2016). Lessons from genome skimming of arthropod-preserving ethanol. *Molecular Ecology Resources*, 16, 1365–1377. <https://doi.org/10.1111/1755-0998.12539>
- Liu, S., Wang, X., Xie, L., Tan, M., Li, Z., Su, X., ... Zhou, X. (2016). Mitochondrial capture enriches mito-DNA 100 fold, enabling PCR-free mitogenomics biodiversity analysis. *Molecular Ecology Resources*, 16, 470–479. <https://doi.org/10.1111/1755-0998.12472>
- Macher, J. N., Vivancos, A., Piggott, J. J., Centeno, F. C., Matthaei, C. D., & Leese, F. (2018). Comparison of environmental DNA and bulk-sample metabarcoding using highly degenerate COI primers. *Molecular Ecology Resources*, 18, 1456–1468. <https://doi.org/10.1111/1755-0998.12940>
- Mardis, E. R. (2013). Next-generation sequencing platforms. *Annual Review of Analytical Chemistry*, 6, 287–303. <https://doi.org/10.1146/annurev-anchem-062012-092628>
- Medeiros, M. J., Eiben, J. A., Haines, W. P., Kaholoaa, R. L., King, C. B. A., Krushekkncky, P. D., ... Starr, K. (2013). The importance of insect monitoring to conservation actions in Hawaii. *Proceedings of the Hawaiian Entomological Society*, 45, 149–166.
- Moriniere, J., de Araujo, B. C., Lam, A. W., Hausmann, A., Balke, M., Schmidt, S., ... Haszprunar, G. (2016). Species identification in Malaise trap samples by DNA barcoding based on NGS technologies and a scoring matrix. *PLoS ONE*, 11, e0155497. <https://doi.org/10.1371/journal.pone.0155497>
- Motwani, R., & Raghavan, P. (1995). 3.6. The Coupon Collector's problem. In *Randomized algorithms* (pp. 57–63). Cambridge, UK: Cambridge University Press.
- Nichols, R. V., Vollmers, C., Newsom, L. A., Wang, Y., Heintzman, P. D., Leighton, M., ... Shapiro, B. (2018). Minimizing polymerase biases in metabarcoding. *Molecular Ecology Resources*, 18, 927–939. <https://doi.org/10.1111/1755-0998.12895>
- Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., ... Wagner, H. (2018). *Vegan: Community ecology package. R package version 2.5-1*. Retrieved from <https://CRAN.R-project.org/package=vegan>
- Pan, W., Byrne-Steele, M., Wang, C., Lu, S., Clemmons, S., Zahorchak, R. J., & Han, J. (2014). DNA polymerase preference determines PCR priming efficiency. *BMC Biotechnology*, 14, 10. <https://doi.org/10.1186/1472-6750-14-10>
- Pimm, S. L., Jenkins, C. N., Abell, R., Brooks, T. M., Gittleman, J. L., Joppa, L. N., ... Sexton, J. O. (2014). The biodiversity of species and their rates of extinction, distribution, and protection. *Science*, 344, 1246752. <https://doi.org/10.1126/science.1246752>
- Piñol, J., Mir, G., Gomez-Polo, P., & Agustí, N. (2015). Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular Ecology Resources*, 15, 819–830. <https://doi.org/10.1111/1755-0998.12355>
- Port, J. A., O'Donnell, J. L., Romero-Maraccini, O. C., Leary, P. R., Litvin, S. Y., Nickols, K. J., ... Kelly, R. P. (2016). Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. *Molecular Ecology*, 25, 527–541. <https://doi.org/10.1111/mec.13481>
- Porter, T. M., & Hajibabaei, M. (2018). Over 2.5 million COI sequences in GenBank and growing. *PLoS ONE*, 13, e0200177. <https://doi.org/10.1371/journal.pone.0200177>
- Potapov, V., & Ong, J. L. (2017). Examining sources of error in PCR by single molecule sequencing. *PLoS ONE*, 12, e0169774. <https://doi.org/10.1371/journal.pone.0169774>
- Prosser, S. W. J., deWaard, J. R., Miller, S. E., & Hebert, P. D. N. (2016). DNA barcodes from century-old type specimens using next-generation sequencing. *Molecular Ecology Resources*, 16, 487–497. <https://doi.org/10.1111/1755-0998.12474>
- Ratnasingham, S., & Hebert, P. D. N. (2013). A DNA-based registry for all animal species: The Barcode Index Number (BIN) system. *PLoS ONE*, 8, e66213. <https://doi.org/10.1371/journal.pone.0066213>
- Russo, L., Stehouwer, R., Heberling, J. M., & Shea, K. (2011). The composite insect trap: An innovative combination trap for biologically

- diverse sampling. *PLoS ONE*, 26, e21079. <https://doi.org/10.1371/journal.pone.0021079>
- Salipante, S. J., Kawashima, T., Rosenthal, C., Hoogestraat, D. R., Cummings, L. A., Sengupta, D. J., ... Hoffman, N. G. (2014). Performance comparison of Illumina and Ion Torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. *Applied and Environmental Microbiology*, 80, 7583–7591. <https://doi.org/10.1128/AEM.02206-14>
- Sato, H., Sogo, Y., Doi, H., & Yamanaka, H. (2017). Usefulness and limitations of sample pooling for environmental DNA metabarcoding of freshwater fish communities. *Scientific Reports*, 7, 14860. <https://doi.org/10.1038/s41598-017-14978-6>
- Shokralla, S., Porter, T. M., Gibson, J. F., Dobosz, R., Janzen, D. H., Hallwachs, W., ... Hajibabaei, M. (2015). Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific Reports*, 5, 9687. <https://doi.org/10.1038/srep09687>
- Sickle, W., Ankenbrand, M. J., Grimmer, G., Holzschuh, A., Hartel, S., Lanzen, J., ... Keller, A. (2015). Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach. *BMC Ecology*, 15, 20. <https://doi.org/10.1186/s12898-015-0051-y>
- Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H., & Flook, P. (1994). Evolution, weighting and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Annals of the Entomological Society of America*, 87, 651–701. <https://doi.org/10.1093/aesa/87.6.651>
- Smith, M. A., Bertrand, C., Crosby, K., Eveleigh, E. S., Fernandez-Triana, J., Fisher, B. L., ... Zhou, X. (2012). *Wolbachia* and DNA barcoding insects: Patterns, potential, and problems. *PLoS ONE*, 7, e36514. <https://doi.org/10.1371/journal.pone.0036514>
- Song, H., Buhay, J. E., Whiting, M. F., & Crandall, K. A. (2008). Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the National Academy of Sciences of the United States of America*, 105, 13486–13491. <https://doi.org/10.1073/pnas.0803076105>
- Tang, M., Tan, M., Meng, G., Yang, S., Su, X., Liu, S., ... Zhou, X. (2014). Multiplex sequencing of pooled mitochondrial genomes - a crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Research*, 42, e166. <https://doi.org/10.1093/nar/gku917>
- Tedersoo, L., Anslan, S., Bahram, M., Polme, S., Riit, T., Liiv, I., ... Abarenkov, K. (2015). Shotgun metagenomes and multiple primer pair barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. *Mycology*, 10, 1–43. <https://doi.org/10.3897/mycokeys.10.4852>
- Tedersoo, L., Tooming-Klunderud, A., & Anslan, S. (2018). PacBio metabarcoding of Fungi and other eukaryotes: Errors, biases, and perspectives. *New Phytologist*, 217, 1370–1385. <https://doi.org/10.1111/nph.14776>
- Tessler, M., Neumann, J. S., Afshinnekoo, E., Pineda, M., Hersch, R., Velho, L. F. M., ... Brugler, M. R. (2017). Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Scientific Reports*, 7, 6589. <https://doi.org/10.1038/s41598-017-06665-3>
- Thomas, A. C., Deagle, B. E., Eveson, J. P., Harsch, C. H., & Trites, A. W. (2015). Quantitative DNA metabarcoding: Improved estimates of species proportional biomass using correction factors derived from control material. *Molecular Ecology Resources*, 16, 714–726. <https://doi.org/10.1111/1755-0998.12490>
- Vasselon, V., Bouchez, A., Rimet, F., Jacquet, S., Trobajo, R., Corniquel, M., ... Domaizon, I. (2017). Avoiding quantification bias in metabarcoding: Application of a cell biovolume correction factor in diatom molecular biomonitoring. *Methods in Ecology and Evolution*, 9, 1060–1069. <https://doi.org/10.1111/2041-210X.12960>
- Veltri, K. L., Espiritu, M., & Singh, G. (1990). Distinct genomic copy number in mitochondria of different mammalian organs. *Journal of Cellular Physiology*, 143, 160–164. <https://doi.org/10.1002/jcp.1041430122>
- Vivien, R., Lejzerowicz, F., & Pawlowski, J. (2016). Next-generation sequencing of aquatic oligochaetes: Comparison of experimental communities. *PLoS ONE*, 11, e0148644. <https://doi.org/10.1371/journal.pone.0148644>
- Vogel, G. (2017). Where have all the insects gone? *Science*, 356, 569–579. <https://doi.org/10.1126/science.356.6338.576>
- Waldron, A., Miller, D. C., Redding, D., Mooers, A., Kuhn, T. S., Nibbelink, N., ... Gittleman, J. L. (2017). Reductions in global biodiversity loss predicted from conservation spending. *Nature*, 551, 364–367. <https://doi.org/10.1038/nature24295>
- Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer-Verlag.
- Yamamoto, S., Masuda, R., Sato, Y., Sado, T., Araki, H., Kondoh, M., ... Miya, M. (2017). Environmental DNA metabarcoding reveals local fish communities in a species-rich coastal sea. *Scientific Reports*, 7, 40368. <https://doi.org/10.1038/srep40368>
- Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., & Ding, Z. (2012). Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, 3, 613–623. <https://doi.org/10.1111/j.2041-210X.2012.00198.x>
- Zimmerman, J., Abarca, N., Enk, N., Skibbe, O., Kusber, W.-H., & Jahn, R. (2014). Taxonomic reference libraries for environmental barcoding: A best practice example from diatom research. *PLoS ONE*, 9, e114758. <https://doi.org/10.1371/journal.pone.0114758>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Braukmann TWA, Ivanova NV, Prosser SWJ, et al. Metabarcoding a diverse arthropod mock community. *Mol Ecol Resour*. 2019;19:711–727. <https://doi.org/10.1111/1755-0998.13008>