Bulletin of Entomological Research (2015) **105**, 717–727
doi:10.1017/S0007485315000681

DNA metabarcoding of insects and allies: an evaluation of primers and pipelines

G.-J. Brandon-Mong[1,2], H.-M. Gan[3,4], K.-W. Sing[1,2], P.-S. Lee[1,2], P.-E. Lim[5] and J.-J. Wilson[1,2]*

[1]Museum of Zoology, Institute of Biological Sciences, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia: [2]Ecology and Biodiversity Program, Institute of Biological Sciences, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia: [3]School of Science, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway, 47500 Petaling Jaya, Selangor, Malaysia: [4]Monash University Malaysia Genomics Facility, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway 47500 Petaling Jaya, Selangor, Malaysia: [5]Institute of Ocean and Earth Sciences (IOES), University of Malaya, 50603 Kuala Lumpur, Malaysia

Abstract

Metabarcoding, the coupling of DNA-based species identification and high-throughput sequencing, offers enormous promise for arthropod biodiversity studies but factors such as cost, speed and ease-of-use of bioinformatic pipelines, crucial for making the leapt from demonstration studies to a real-world application, have not yet been adequately addressed. Here, four published and one newly designed primer sets were tested across a diverse set of 80 arthropod species, representing 11 orders, to establish optimal protocols for Illumina-based metabarcoding of tropical Malaise trap samples. Two primer sets which showed the highest amplification success with individual specimen polymerase chain reaction (PCR, 98%) were used for bulk PCR and Illumina MiSeq sequencing. The sequencing outputs were subjected to both manual and simple metagenomics quality control and filtering pipelines. We obtained acceptable detection rates after bulk PCR and high-throughput sequencing (80–90% of input species) but analyses were complicated by putative heteroplasmic sequences and contamination. The manual pipeline produced similar or better outputs to the simple metagenomics pipeline (1.4 compared with 0.5 expected:unexpected Operational Taxonomic Units). Our study suggests that metabarcoding is slowly becoming as cheap, fast and easy as conventional DNA barcoding, and that Malaise trap metabarcoding may soon fulfill its potential, providing a thermometer for biodiversity.

Keywords: Arthropoda, biodiversity, COI, high-throughput sequencing, Illumina MiSeq, Malaise trap

(Accepted 19 July 2015; First published online 7 September 2015)

Introduction

Much of our knowledge of biodiversity patterns and changes comes from the data based on mammals, birds and vascular plants (e.g., Gillison *et al.*, 2013). Yet these taxa represent only a fraction of biodiversity; the major component of terrestrial biodiversity comprises insects (Mora *et al.*, 2011). A recent meta-analysis of biodiversity studies revealed the dearth of information about most of the world's tropical biota (Gillison *et al.*, 2013), highlighting the fact that in order to decipher biodiversity patterns and change the major component can no longer be ignored. The absence of data on insects in biodiversity surveys, with the exception of small

*Author for correspondence
Fax: +603-7967-4178
E-mail: johnwilson@um.edu.my

https://doi.org/10.1017/S0007485315000681 Published online by Cambridge University Press

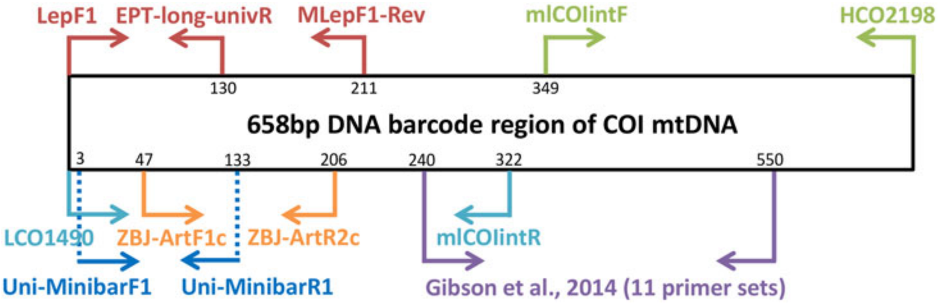

Fig. 1. Relative positions of primers on the COI barcode region.

groups of charismatic taxa such as butterflies, dragonflies and dung beetles (e.g., Korasaki *et al.*, 2013; Hart *et al.*, 2014; Zografou *et al.*, 2014), reflects the taxonomic challenges associated with the huge diversity of this group of relatively small-sized organisms (Floyd *et al.*, 2009). Obtaining insect samples is not an obstacle to collecting this data as many efficient sampling techniques have been developed (e.g., Russo *et al.*, 2011, and in particular Malaise traps) but the investment required to sort and classify these samples is prohibitive. Fortunately, modern technology is addressing this impediment. First, conventional (single specimen) DNA barcoding, the use of short cytochrome c oxidase I mtDNA (COI) sequences as species identification tags (Hebert *et al.*, 2003), has been applied to rapidly accelerate biodiversity surveys in hyperdiverse insect groups (e.g., ants of Madagascar; Smith *et al.*, 2005). Now, with next-generation-sequencing technologies allowing simultaneously sequencing of DNA fragments from multiple specimens in a bulk mixture of diverse taxa, termed metabarcoding (Yu *et al.*, 2012), the impediment is being alleviated further.

Metabarcoding is simply the pairing of DNA-based species recognition with high-throughput (next-generation) DNA sequencing (HTS) (Ji *et al.*, 2013). Consequently, metabarcoding, like conventional DNA barcoding, relies on 'universal' polymerase chain reaction (PCR) primers that can amplify a fragment of a standard DNA region from diverse taxa (Ji *et al.*, 2013). Due to the limitations in the size of DNA fragments sequenced by HTS platforms (see Shokralla *et al.*, 2014), metabarcoding has typically been restricted to targeting short fragments of the COI DNA barcode (e.g., Hajibabaei *et al.*, 2011; Zeale *et al.*, 2011). Prior to the coining of the word 'metabarcoding', the idea of 'mini-barcodes' had been investigated in the context of degraded DNA samples (Hajibabaei *et al.*, 2006). Hajibabaei *et al.* (2006) concluded that 135 bp fragments of COI can distinguish most species, but the location of the fragment within the full-length DNA barcode (~658 bp) is important. After further exploration of primer binding sites and species resolution offered by different fragments within the COI DNA barcode region, Meusnier *et al.* (2008) designed and advocated the use of Uni-MinibarF1 and Uni-MinibarR1 primers (amplifying a 130 bp fragment; see fig. 1) as a universal (eukaryote) primer set for the amplification of mini-barcodes (Meusnier *et al.*, 2008). Zeale *et al.* (2011) designed and tested primers (ZBJ-ArtF1c and ZBJ-ArtR2c; see fig. 1) which amplify a 160 bp fragment from the 5′ end of COI for application in the study of arthropod prey in bat guano. These primers have since been used extensively in metabarcoding-type studies of diets (e.g., Bohmann *et al.*, 2011; Razgour *et al.*, 2011; Vesterinen *et al.*, 2013; Burgar *et al.*, 2014; Hope *et al.*, 2014; Piñol *et al.*, 2014*a*). Other primers were designed by Leray *et al.* (2013) for the analysis of the (metazoan) diets of fish collected at coral reefs, targeting ~330 bp fragments of DNA barcode suitable for amplicon 454 pyrosequencing.

Besides diet studies, metabarcoding has been applied to environmental monitoring (Hajibabaei *et al.*, 2011). Hajibabaei *et al.* (2011) collected aquatic insect samples in southern Ontario, Canada, for a test of metabarcoding, targeting a 130 bp fragment of COI (LepF1 primer paired with a newly designed reverse primer – EPT-long-univR; see fig. 1). Metabarcoding of bulk Malaise trap samples took off with Yu *et al.* (2012) with a 'biodiversity soup' study. This study employed primers typically used for DNA barcoding of insects (Folmer *et al.*, 1994; also see Wilson, 2012) for amplicon 454 pyrosequencing producing sequenced fragments (~400 bp) which when assembled together cover the full-length DNA barcode (~658 bp). Liu *et al.* (2013) used the same samples (from Yu *et al.*, 2012) to develop a new bioinformatics pipeline 'SOAPBarcode' utilizing Illumina (HiSeq 2000) shotgun sequencing of the amplicons; in brief, 150 bp sections of the amplicon are sequenced and then assembled together to form the full-length DNA barcode. The use of metabarcoding as a source of data for conservation policy-making was validated by Ji *et al.* (2013) who compared metabarcoding datasets against standard biodiversity datasets in Malaysia (metabarcoded Malaise dataset versus birds, dung beetles, ants), China (light trap collected moths, both metabarcoded and morphologically identified) and England (metabarcoded whole pitfall-trap dataset versus ants, spiders, carabid beetles). Like Yu *et al.* (2012), Ji *et al.* (2013) used degenerate Folmer primers for amplicon 454 pyrosequencing. Yang *et al.* (2014) also followed the protocols of Yu *et al.* (2012) to test the metabarcoding approach on soil and leaf-litter samples for rapid environmental monitoring in terrestrial ecosystems.

The reliance on 'universal' primers and associated biases has been of concern to the early practitioners of metabarcoding. Hajibabaei *et al.* (2011) reported a taxonomic bias for their LepF1 and EPT-long-univR primer set (fig. 1) as well as biases due to varying abundances of species in bulk samples (Hajibabaei *et al.*, 2011). Yu *et al.* (2012) reported limitations of the classic 'Folmer' barcoding primers, particularly in regard to amplification of hymenopterans. *In vitro* PCR analyses by Clarke *et al.* (2014) with five primer sets targeting COI and 16S rDNA suggested a primer bias of COI markers towards lepidopteran and dipteran species with certain orders failing to amplify. In response to these simulated and empirical observations of primer biases in metabarcoding, Zhou *et al.* (2013) developed a new PCR-free Illumina pipeline for DNA-based

Table 1. Amplification success for five tested primer sets. Amplification success for bulk PCR using two primer sets was estimated by BLAST-matching Illumina reads to Sanger sequences (e-value <1e-100); the results from two Illumina runs are shown in parentheses.

Forward primer	LCO1490	Uni-MinibarF1	LCO1490	mlColintF	LepF1
Reverse primer	ZBJ-ArtR2c	Uni-MinibarR1	mlCOlintR	HCO2198	MLepF1_Rev
Amplicon length	207 bp	130 bp	319 bp	313 bp	218 bp
Araneae 2 spp.	0	1	0	2 (0.1)	1 (1.1)
Blattodea 2 spp.	1	2	0	2 (0.1)	2 (2.2)
Coleoptera 14 spp.	1	5	0	13 (5.13)	14 (6.10)
Diptera 19 spp.	4	15	0	19 (13.19)	18 (14.18)
Hemiptera 4 spp.	2	2	1	4 (2.4)	4 (3.3)
Hymenoptera 16 spp.	0	9	1	16 (6.12)	16 (4.10)
Lepidoptera 17 spp.	8	12	1	16 (13.17)	17 (15.17)
Mantodea 1 sp.	0	1	0	1 (1.1)	1 (1.1)
Odonata 1 sp.	0	1	0	1 (1.1)	1 (0.0)
Orthoptera 3 spp.	0	1	1	3 (2.3)	3 (2.2)
Collembola 1 sp.	0	1	0	1 (0.0)	1 (0.0)
Total 80 spp.	16	50	4	78 (43.72)	78 (48.64)

biodiversity assessment in bulk samples. Although the PCR-free Illumina pipeline (Zhou *et al.*, 2013) enabled the successful identification of 97% of (73) species in a pooled sample, the pipeline produced large amounts (99.47%) of redundant data i.e., sequences not (presently) useful for taxonomic identification purposes, despite a mitochondrial enrichment step. Tang *et al.* (2014) followed this work omitting mitochondrial enrichment with similar results. Another approach to limit primer biases in metabarcoding has been to use multiplex PCR (multiple primers) prior to amplicon sequencing (Gibson *et al.*, 2014). In order to maximize taxon detection, Gibson *et al.* (2014) used 11 unique PCR primer sets which all targeted the same 310 bp fragment of the standard COI DNA barcode (see fig. 1).

Despite the advancements in PCR-free and multiplex PCR pipelines, metabarcoding using universal primers for bulk PCR amplification still remains the most cost-effective and time-efficient protocol. PCR-free approaches generate a huge volume of redundant (un-utilized i.e., non-barcode) sequences. Only ~0.53% of raw sequences were mitochondrial sequences (Tang *et al.*, 2014), even after mitochondrial enrichment (Zhou *et al.*, 2013). In addition, shotgun, PCR-free approaches could miss the COI barcode target due to insufficient sequencing (Tang *et al.*, 2014), especially as bulk samples are pooled for cost-efficiency. The payoffs from multiplex PCR deserve a more systematic evaluation – 11 primer sets recovered 91% of the known species in a pooled sample (Gibson *et al.*, 2014), whereas a single primer set has been reported as successfully amplifying 91% of tested taxa (Leray *et al.*, 2013). Such comparisons are likely to be idiosyncratic, but undoubtedly depend on the 'universality' of the single primer sets being compared; and it is worth to note that Gibson *et al.* (2014) did not make a comparison with the standard DNA barcoding primers. In addition, for cost-effective metabarcoding, the Illumina sequencing platform may be preferred due to its low sequencing cost compared with Roche 454 platforms (Yu *et al.*, 2012; Liu *et al.*, 2013; Yang *et al.*, 2014). The Roche 454 platform has an estimated processing cost of US $240–415 per metabarcode sample, whereas the processing cost using an Illumina platform is estimated as half this value (Yu *et al.*, 2012; Liu *et al.*, 2013; Yang *et al.*, 2014). An Illumina Miseq v2 can produce up to 15 million reads per run while Roche 454 FLX Titanium can generate 1 million reads per run (Glenn, 2014; also see Liu *et al.*, 2013; Shokralla *et al.*, 2014). Furthermore, Roche is considered a 'Zombie platform' (Glenn, 2014). The objectives of this study were: (1) to test and compare universal primer sets on a diverse set of arthropod orders to establish the optimal primer set for metabarcoding a tropical Malaise trap sample, while also considering the short read-length requirements of the Illumina platform; and (2) to test and compare the ease-of-use and reliability of the outputs from two bioinformatic pipelines (metagenomic and manual) to establish the optimal quality control and filtering pipeline for 'real-world' application of arthropod metabarcoding.

Materials and methods

Sample collection, selection and DNA extraction

A Malaise trap was set at Rimba Ilmu Botanic Garden, University of Malaya, Kuala Lumpur, Malaysia between 7 and 13 June 2014. From the bulk sample collected, 80 morphologically distinct specimens were selected as a test dataset, with the aim of maximizing taxonomic diversity. The specimens were pinned and oven-dried for 24 h. Based on examination of morphological characters (Triplehorn & Johnson, 2005) the test dataset included species from the orders: Lepidoptera, Hymenoptera, Araneae, Blattodea, Coleoptera, Orthoptera, Odonata, Diptera, Hemiptera, Collembola and Mantodea (table 1). Genomic DNA was extracted from the whole bodies of smaller specimens and two legs of larger specimens using a NucleoSpin Tissue kit (Macherey-Nagel, Germany), following the manufacturer's instructions. A NanoDrop spectrophotometer (NanoDrop 2000c UV-Vis Spectrophotometer, Thermo Scientific) was used for DNA purity and concentration assessment.

Primer selection and testing: individual specimen

Four primer sets were retrieved from the metabarcoding literature: (i) ZBJ-ArtF1c/ZBJ-ArtR2c (Zeale *et al.*, 2011), (ii) Uni-MinibarF1/Uni-MinibarR1 (Meusnier *et al.*, 2008), (iii) mlCOIintF/HCO2198 (Leray *et al.*, 2013), (iv) LCO1490/mlCOIintR (Leray *et al.*, 2013) (table 1). In addition, a new reverse primer (MLepF1-Rev) was designed for use with the standard barcoding primer LepF1 (Hebert *et al.*, 2004). In our previous studies we have found high amplification success with the standard barcoding primer MLepF1 (Smith *et al.*, 2008*b*; also see Wilson, 2012) and noted its binding site around

https://doi.org/10.1017/S0007485315000681 Published online by Cambridge University Press

200 bp from LepF1 (fig. 1). Consequently, we used the program Primer3 Plus (Rozen & Skaletsky, 2000) and a set of diverse high-quality insect COI sequences from another study (Wong *et al.*, 2015) to select a 22 bp region slightly downstream of MLepF1 with appropriate structural and physical properties for primer binding. We included two degenerate bases (Ws) to create a reverse version of MLepF1 (named MLepF1-Rev). In preliminary testing with individual specimens, we found very low amplification success with ZBJ-ArtF1c/ZBJ-ArtR2c, therefore we proceeded with LC01490 (the standard COI barcoding primer) and ZBJ-ArtR2c as an alternative combination.

PCR amplification was performed in a total volume of 25 µl with 0.25 µl of each forward and reverse primer (10 µM), 12.5 µl of Taq98® Hot Start 2X Master Mix (Lucigen, USA), 10 µl ddH$_2$0 and 2 µl of genomic DNA. For each primer set, we followed the thermocycling programs recommended by previous studies: for LCO1490/ZBJ-ArtR2c a touch-down program (40 cycles) with annealing temperatures 61–53°C was followed (Zeale *et al.*, 2011); for Uni-MinibarF1/Uni-MinibarR1 a 'touch-up' program (40 cycles) with annealing temperatures of 46 and 53°C was followed (Meusnier *et al.*, 2008); for mlColintF/HCO2198 and LCO1490/mlCOlintR a touch-down program (41 cycles) with annealing temperatures between 62 and 46°C was followed (Leray *et al.*, 2013); for LepF1/MLepF1-Rev we followed 'COI Fast' (40 cycles) (Wilson, 2012). Success of PCR amplifications was checked on 2% agarose gels. A clear band of expected length (refer to fig. 1) indicated amplification success, whereas the absence of a band was recorded as PCR amplification failure.

Primer selection and testing: bulk PCR and Illumina sequencing

Based on the results from the individual specimen tests (see above), the two primer sets with the highest amplification success were selected and modified to include Illumina sequencing adapters with multiplex identifiers (following Bartram *et al.*, 2011; table 2). The 80 test DNA extracts were pooled (1–15 µl of DNA extract from each specimen depending on the measured DNA concentration) and used for bulk PCR using the two modified primer sets. Initially PCR amplification was performed in a total volume of 25 µl with 0.25 µl of each forward and reverse primer (10 µM), 12.5 µl of Taq98® Hot Start 2X Master Mix (Lucigen, USA), 10 µl ddH$_2$0 and 2 µl of pooled genomic DNA. For each primer set, we followed the thermocycling programs from above. Success of PCR amplifications was checked on 2% agarose gels. Amplicons were gel extracted and purified using a NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel, Germany), following the manufacturer's instructions. The libraries were quantified using KAPA library quantification kit (KAPA Biosystems, South Africa), normalized, pooled and sequenced on a MiSeq Desktop Sequencer (Illumina, USA) constituting approximately 0.5% of a MiSeq V2 500 cycle kit. Paired-end sequencing was performed at the Monash University Malaysia Genomics Facility.

Subsequently a second Illumina MiSeq run was conducted with amplicons produced by PCR in a total volume of 25 µl with 2.5 µl of each forward and reverse primer (10 µM), 2.0 µl of dNTPs, 0.25 µl of Accura™ High-Fidelity Polymerase (Lucigen, USA), 12.5 µl of Accura™ 2X HF buffer (Lucigen, USA), 3.25 µl of ddH$_2$O and 2 µl of mixed genomic DNA. The thermocycling profile was modified to minimize chimera formation (fewer cycles with longer extension times) during PCR: for [V3]mlColintF/[MID96]HCO2198 , 95°C for 2 min;

Table 2. Primers used in this study. The Illumina adapter sequences incorporating a multiplex identifier are shown in square brackets.

Primer	Sequence (5′–3′)	Paired with	References
LCO1490	GGTCAACAAATCATAAAGATATTGG	ZBJ-ArtR2c mlCOIintR	Folmer *et al.* (1994)
[MID96] HCO2198	[CAAGCAGAAGACGGCATACGAGATGGATGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT] TAAACTTCAGGGTGACCAAAAAATCA	[V3]mlCOIintF	Folmer *et al.* (1994)
ZBJ-ArtR2c	WACTAATCAATTWCCAAATCCTCC	LCO1490	Zeale *et al.* (2011)
Uni-MinibarF1	TCCACTAATCACAARGATATTGGTAC	Uni-MinibarR1	Meusnier *et al.* (2008)
Uni-MinibarR1	GAAAATCATAATGAAGGCATGAGC	Uni-MinibarF1	Meusnier *et al.* (2008)
[V3]mlCOIintF	[AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNN] GGWACWGGWTGAACWGTWTAYCCYCC	[MID96]HCO2198	Leray *et al.* (2013)
mlCOIintR	GGRGGRTASACSGTTCASCCSGTSCC	LCO1490	Leray *et al.* (2013)
[V3]LepF1	[AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNN] ATTCAACCAATCATAAAGATATTGG	[MID95] MLepF1-Rev	Hebert *et al.* (2004)
[MID95] MLepF1-Rev	[CAAGCAGAAGACGGCATACGAGATGAACGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT] CGTGGAAAWGCTATATCWGGTG	[V3]LepF1	This study

https://doi.org/10.1017/S0007485315000681 Published online by Cambridge University Press

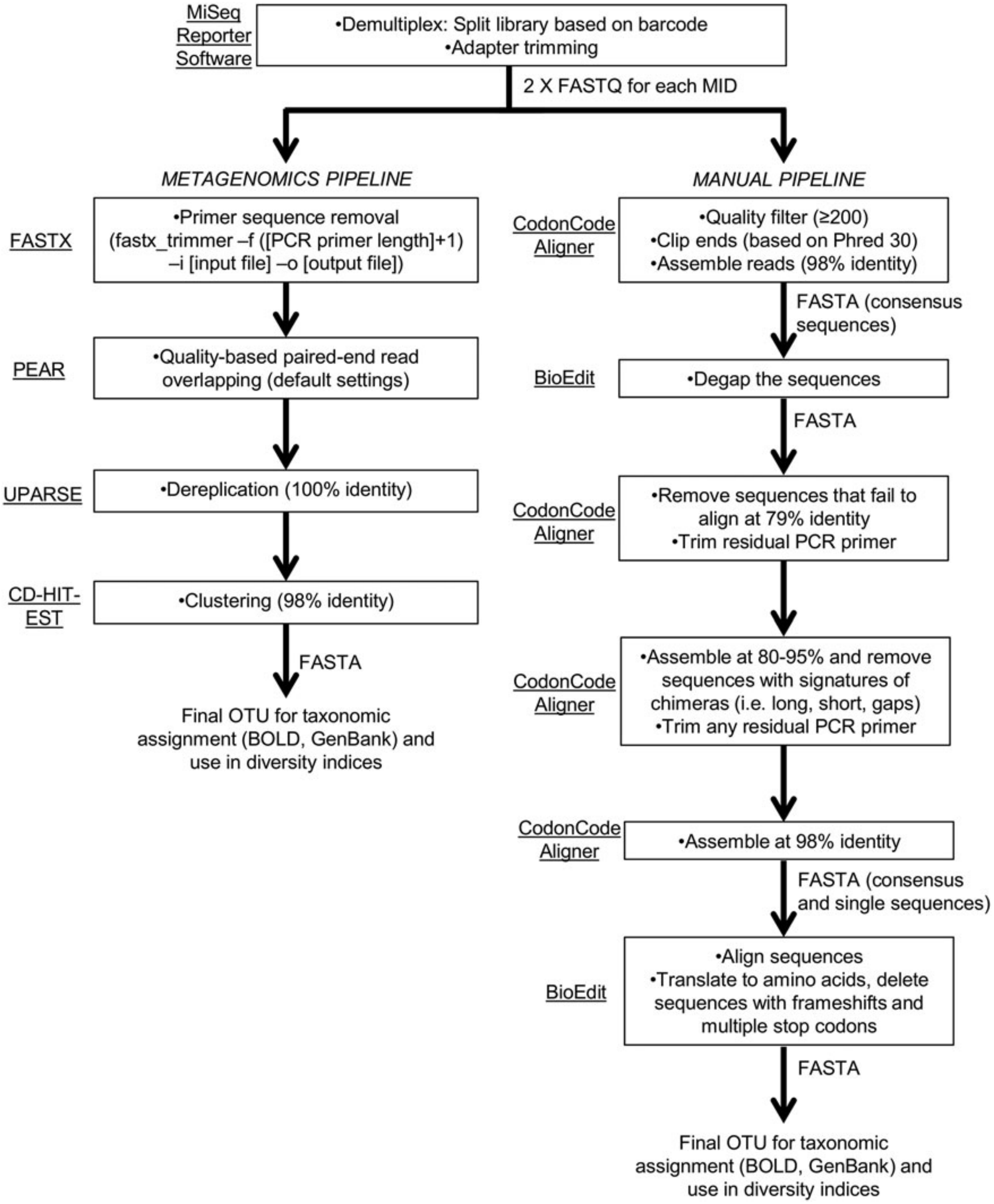
MiSeq Reporter Software
•Demultiplex: Split library based on barcode
•Adapter trimming
2 X FASTQ for each MID
METAGENOMICS PIPELINE
MANUAL PIPELINE
FASTX
•Primer sequence removal (fastx_trimmer –f ([PCR primer length]+1) –i [input file] –o [output file])
PEAR
•Quality-based paired-end read overlapping (default settings)
UPARSE
•Dereplication (100% identity)
CD-HIT-EST
•Clustering (98% identity)
FASTA
Final OTU for taxonomic assignment (BOLD, GenBank) and use in diversity indices
CodonCode Aligner
•Quality filter (≥200)
•Clip ends (based on Phred 30)
•Assemble reads (98% identity)
FASTA (consensus sequences)
BioEdit
•Degap the sequences
FASTA
CodonCode Aligner
•Remove sequences that fail to align at 79% identity
•Trim residual PCR primer
CodonCode Aligner
•Assemble at 80-95% and remove sequences with signatures of chimeras (i.e. long, short, gaps)
•Trim any residual PCR primer
CodonCode Aligner
•Assemble at 98% identity
FASTA (consensus and single sequences)
BioEdit
•Align sequences
•Translate to amino acids, delete sequences with frameshifts and multiple stop codons
FASTA
Final OTU for taxonomic assignment (BOLD, GenBank) and use in diversity indices

Fig. 2. Schematic of bioinformatic steps (metagenomic and manual pipelines).

25 cycles of 95°C for 15 s; 51°C for 30 s; 72°C for 3 min and a final extension of 72°C for 10 min; for [V3]LepF1/[MID95] MLepF1-Rev, 95°C for 2 min; 25 cycles of 95°C for 15 s; 45°C for 30 s; 72°C for 3 min and a final extension of 72°C for 10 min. Five independent PCR products for each primer set were pooled prior to gel extraction. The next steps followed as above with each sample comprising approximately 2.75% of the sequencing run.

Quality control and filtering pipelines

Sequencing reads were demultiplexed and adapter-trimmed onboard the MiSeq using the MiSeq Reporter software. This resulted in a 'raw' output of two paired-end FASTQ files for each primer set (fig. 2). We followed two pipelines (fig. 2) for quality control and filtering of the paired-end reads: (a) a simplified metagenomics pipeline (by HMG)

https://doi.org/10.1017/S0007485315000681 Published online by Cambridge University Press

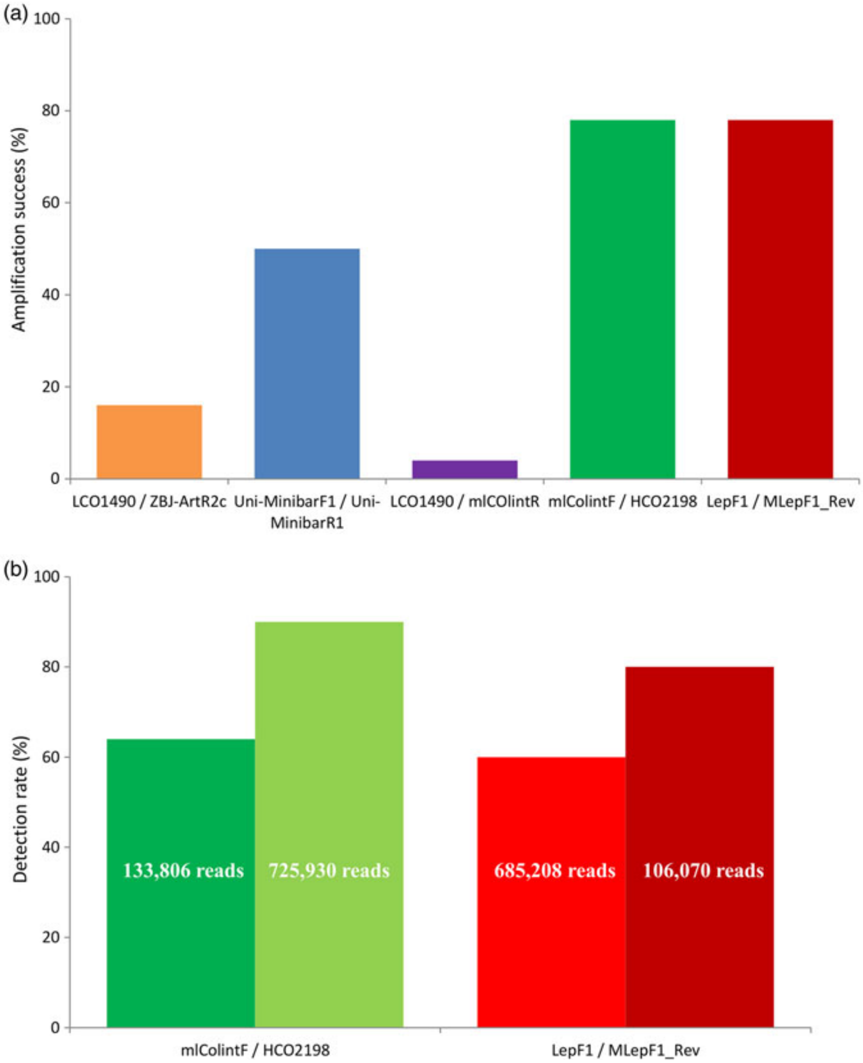

Fig. 3. (a) Amplification success rate for primer sets in conventional single specimen PCR; (b) detection rate of two primer sets used in bulk PCR and Illumina sequencing based on the percentage of Sanger sequences BLAST-matched to HTS reads with an e-value $<1e^{-100}$.

incorporating FASTX (Hannon Lab, 2014), PEAR (Zhang *et al.*, 2014), UPARSE (Edgar, 2013) and CD-HIT-EST (Fu *et al.*, 2012); (b) we screened and filtered the reads 'manually' using CodonCode Aligner (CodonCode Corp.) and BioEdit (Hall, 1999) (by GJBM & JJW).

The 80 individual DNA extractions were also used for PCR amplification with Folmer primers (following standard methods; Wilson, 2012) and the products, or alternatively PCR products generated during the individual specimen primer tests (see above), were sequenced by a local company (MyTACG Bioscience). These Sanger sequences are available on BOLD (Ratnasingham & Hebert, 2007) in the public project MBPT. The assembled and dereplicated Illumina reads were 'BLASTed' (Altschul *et al.*, 1990) against the Sanger sequences to give an estimate of the species surviving the bulk PCR and Illumina sequencing (i.e., the detection rate based on hits with an e-value of $<1e^{-100}$). Additionally, we built neighbor-joining (NJ) trees (in MEGA 6; Tamura *et al.*, 2013) combining the filtered metabarcode Operational Taxonomic Units (OTU) and Sanger sequences. The single representative of Odonata did not generate a Sanger sequence, but was traced among the OTU by BLAST searches against GenBank.

Results

Primer testing: individual specimen

The primer sets mlCOIintF/HCO2198 and LepF1/MLepF1-Rev showed the highest amplification success (both 98%), followed by Uni-MinibarF1/Uni-MinibarR1 (63%), LCO1490/ZBJ-ArtR2c (20%) and LCO1490/mlCOIintR (5%) (table 2, fig. 3). Consequently mlCOIintF/HCO2198 and LepF1/MLepF1-Rev were used for further evaluation.

Primer testing: bulk PCR and Illumina sequencing

FASTQ files related to this study are available in the NCBI short read archive under accession SRR1848965. The first sequencing run produced 106,070 paired-end reads for the LepF1/MLepF1-Rev primer set (28 Mb) and 133,806

https://doi.org/10.1017/S0007485315000681 Published online by Cambridge University Press

Table 3. Comparison of quality control and filtering pipelines applied to Illumina MiSeq metabarcodes.

		METAGENOMIC (PEAR/CD-HIT) PIPEPLINE OTU		MANUAL (CODONCODE/ BIOEDIT) PIPELINE OTU	
Primer set	Paired-end reads	Expected (Sanger sequence match)	Unexpected	Expected (Sanger sequence match)	Unexpected
LepF1/MLepF1-Rev	685,208	44	127	67	55
mlCOIintF/ HCO2198	725,930	64	137	62	43

paired-end reads for the mlCOlintF/HCO2198 primer set (35M b). The detection rate of input species was 60% for the LepF1/MLepF1-Rev primer set and 64% for the mlCOlintF/HCO2198 primer set (fig. 3). A second sequencing run at greater sequencing depth produced 685,208 paired-end reads for the LepF1/MlepF1-Rev primer set (167 Mb) and 725,930 paired-end reads for the mlCOlintF/HCO2198 primer set (136 Mb) (table 3). The detection rate of input species was 80% for the LepF1/MLepF1-Rev primer set and 90% for the mlCOlintF/HCO2198 primer set (fig. 3). There were notable differences in the abundance of reads produced for each species. LepF1/MLepF1-Rev produced double the number of expected lepidopteran (17 input species) reads (16% of reads) compared with mlCOlintF/HCO2198 (8% of reads) and significantly fewer expected hymenoteran (16 input species) reads (0.1% of reads compared with 0.8%). LepF1/MLepF1-Rev produced a large number of expected hemipteran (4 input species) reads (6% of reads) compared with mlCOlintF/HCO2198 (0.02% of reads) whereas the primers produced comparable number of expected dipteran (19 input species) reads (9% of reads) and expected coleopteran (14 input species) reads (1% of reads).

Quality control and filtering pipelines

After the simple metagenomics quality control and filtering pipeline, 171 OTU were retained for the LepF1/MLepF1-Rev primer set, 44 of which corresponded to the 80 input species (i.e., matched Sanger sequences so were 'expected' whereas the remaining 127 OTU did not match Sanger sequences so were 'unexpected') and 201 OTU were retained for the mlCOlintF/HCO2198 primer set, 64 of which corresponded to the 80 input species (table 3). After applying the manual quality control and filtering pipeline, 122 OTU were retained for the LepF1/MLepF1-Rev primer set, 67 of which corresponded to the 80 input species and 105 OTU were retained for the mlCOlintF/HCO2198 primer set, 62 of which corresponded to the 80 input species (table 3). The unexpected OTU included contaminants such as *Wolbachia* and bats for the LepF1/MLepF1-Rev primer set and fungi and mammals for the mlCOlintF/HCO2198 primer set, based on BLAST hits in GenBank (online Supplemental Figure).

Discussion

To move from demonstration technology to a practical, widely employed, biodiversity monitoring tool, Malaise trap metabarcoding, must be (i) easy to understand, (ii) easy to use, (iii) fast and cheap. Our metabarcoding approach using a single primer set targeting a short mini/metabarcode is very similar in essence to conventional DNA barcoding, which has already gained considerable acceptance among conservation practitioners and the general public (e.g., Bucklin *et al.*, 2011; Fišer Pečnikar & Buzan, 2014; Kress *et al.*, 2015). Our DNA extraction from fresh caught (Malaise trap) specimens and PCR with a single primer set (such as conducted in this study) can be completed in a basic molecular lab in a few hours, while an Illumina MiSeq v2 run takes ~39 h. HTS can be outsourced to commercial companies at reasonable (and dropping) cost (US$2.5 per Mb in Malaysia). Therefore, to obtain 100 Mb for a 1-week Malaise trap sample (~100–300 specimens) can cost around US$250, theoretically US$1–2.5 per specimen or less (DNA extraction and PCR would add approximately US$1 per specimen).

The commercial companies will also provide bioinformatics analysis of the submitted samples up to BLAST hit, however the company may not be familiar with the specific protocols or purpose of the study, so we would always recommend the end-user retains control of the quality control and filtering pipeline. In our view, it is unrealistic to expect the users of applied metabarcoding (e.g., conservation officers in government agencies or NGOs) to master a series of command line programs to analyze their metabarcodes. A specific step-by-step web interface (such as those available for phylogenetic analyses; Dereeper *et al.*, 2008) would be a significant step in the development of metabarcoding as a practical tool. Alternatively, easy-to-use GUI DNA sequence editing software, such as the widely used CodonCode Aligner, can be used to filter moderately sized metabarcode samples (e.g., weekly Malaise trap collections) and produces similar or better outputs to the 'conventional' pipelines adopted from bacterial metagenomics – 1.4 compared with 0.5 expected(Sanger matching):unexpected (without Sanger matches). Consequently, quality control and filtering of metabarcode datasets has the potential to be straightforward with considerable room for user input as opposed to the 'blackbox' of more complex pipelines (especially those requiring advanced sequence assembly e.g., Liu *et al.*, 2013). Several examples from traditional DNA barcoding studies illustrate the need for careful understanding and review of sequence data by the user (e.g., Wilson & Sing, 2013).

Despite the significant progress made in metabarcoding in recent years, several issues remain. Particularly important issues concern what is considered an acceptable detection rate (influenced both by sequencing depth and difficult to amplify taxonomic groups i.e., PCR bias) and species identification (incorporating species resolution, heteroplasmy and contamination). Further issues relating to species delimitation methods and the completeness of DNA barcode reference libraries for the identification of OTU are also critical (e.g., Wilson *et al.*, 2011).

The detection rate for 80 input species, which is slightly less than that found in a weekly Malaise trap sample in Malaysia (~100–300 species; Ji *et al.*, 2013), was 80 and 60% at 167 and 35 Mb of sequencing output, respectively, for the LepF1/MLepF1-Rev and 90 and 64% at 136 and 53 Mb of sequencing output, respectively, for the mlCOIintF/HCO2198 primer set.

This is less than detection at >97% in the PCR-free Illumina pipeline of Zhou *et al.* (2013) and Tang *et al.* (2014). However, the detection rate is comparable with that reported for bulk amplification with Folmer primers – 81% in the 'biodiversity soup' pipeline (Folmer primers and 454 sequencing) of Yu *et al.* (2012) and 84.9% using Illumina shotgun sequencing of the 'biodiversity soup' amplicons (Liu *et al.*, 2013). Considering the size of the sequencing output: >1.1 Gb for Liu *et al.* (2013), and 13.2–31.7 Gb for PCR-free pipelines (Zhou *et al.*, 2013; Tang *et al.*, 2014), this is an unfair comparison and represents the trade-off between cost and detection.

Previous studies have reported a low detection rate for species of Hymenoptera (Yu *et al.*, 2012; Zhou *et al.*, 2013) and this was also seen in bulk PCR and Illumina sequencing in the current study (25% of hymenopteran species were detected during the first sequencing run). Interestingly, the amplification success rate for hymenopterans using single-specimen PCR was 100% for the best primer sets, LepF1/MLepF1-Rev and mlCOIintF/HCO2198, showing that these primers can amplify hymentopteran COI but that there may be a bias during bulk PCR. It has been suggested that species with lower affinities with primer binding sites will yield lower level amplicons and fewer, if any, reads (Hajibabaei *et al.*, 2011); but primer affinity is hard to predict (see Lee *et al.*, 2015). To alleviate or at least minimize taxonomic bias in primer sets, lower PCR annealing temperatures (Ishii & Fukui, 2001; Sipos *et al.*, 2007) and deeper sequencing (Hajibabaei *et al.*, 2011) can be performed (88% of hymenopteran species were detected during the second deeper sequencing run), but may involve a trade-off in terms of non-specific binding and increased cost. Although a high number of hymenopteran species were detected by BLAST hit of the raw dereplicated reads, a large proportion of these reads were filtered out by the quality control pipelines, suggesting that although they were generated during sequencing, the hymenopteran reads were low abundance, low quality or characteristic of 'error' sequences. Hymenopteran reference sequences containing the poly-T region that is difficult to sequence cleanly and accurately (Zhou *et al.*, 2013) may be beneficial in guiding quality control and filtering pipelines and help avoid discard of genuine hymenopteran OTU. The bias towards amplification of lepidopteran and dipteran sequences (as reported by Clarke *et al.*, 2014) was seen in this study although appeared less severe for the primer set mlCOIintF/HCO2198. Deagle *et al.* (2014) argued that because COI has poorly conserved regions for primer design, the list of potential markers for metabarcoding has to be broadened. This is a controversial position as the opportunities for species identification based on the large, curated DNA barcode (COI) libraries (Ratnasingham & Hebert, 2007) would be forfeit (Yu *et al.*, 2012). Quality control and filtering techniques (especially for chimeric sequences) often rely on properties of protein-coding sequences (Yu *et al.*, 2012; Leray & Knowlton, 2015) and the advantages of mitochondrial protein-coding genes for species identification are well-established. Our study supports the previous work that has shown that reasonable detection rates and taxonomic coverage can be achieved with COI metabarcodes.

Although the majority of Sanger and OTU matches were within 2% p-distance, some Sanger sequences fell into clusters with closely grouping OTU on the NJ trees. There are a number of potential explanations for these sequence clusters including chimera formation, PCR or Illumina sequencing errors (Haas *et al.*, 2011; Quail *et al.*, 2012) and mitochondrial heteroplasmy (Shokralla *et al.*, 2014). Haas *et al.* (2011) reported that the number of PCR amplification cycles has a dominant effect on chimera formation. By increasing the PCR extension time, reducing the concentration of template DNA and the number of amplification cycles to the fewest number (approximately 20 cycles) still able to yield sufficient amplicons for sequencing, chimera formation can be alleviated or at least be minimized (Lahr & Katz, 2009; Haas *et al.*, 2011; Stevens *et al.*, 2013). Rapid changes in temperature might produce incomplete products which subsequently anneal to other DNA templates, creating chimeras, thus slowing the PCR ramp speed to 1°C s^{-1} has been recommended as another modification to inhibit chimera formation (Stevens *et al.*, 2013). Several potential chimeric sequences were observed and removed during our manual filtering steps, for example, when a group of reads showed close matches across a significant portion of the read and major divergences across another portion, or the presence of large gaps and frameshifts. Common methods for chimera filtering rely on analyzing the distribution and abundance of closely matching reads (Boyer *et al.*, 2014) but we have found this approach will significantly reduce the detection rate when low abundance but 'real' (100% match to Sanger sequences) sequences are inadvertently filtered (also reported by Yu *et al.*, 2012). Other common approaches for chimera removal rely on reference alignments (Edgar *et al.*, 2011), but this is problematic for datasets consisting of many novel sequences, such as tropical Malaise samples. Other chimera detection methods based on signatures of recombination within a dataset (Martin *et al.*, 2010) may be suitable additions to metabarcoding pipelines.

The observed divergences between Sanger sequences and OTU may partially be explained by heteroplasmy (coexistence of multiple mitochondrial haplotypes in an individual) (Magnacca & Brown, 2010; Shokralla *et al.*, 2014). COI heteroplasmy has been documented in many insects species across several orders; Lepidoptera (12% of species examined; Shokralla *et al.*, 2014), Orthoptera (2–24% of individuals observed; Moulton *et al.*, 2010), Hymenoptera (13% of Hawaiian *Hylaeus*; Magnacca & Brown, 2010) and Diptera (17% of individuals of *Drosophila melanogaster*; Townsend & Rand, 2004). After our manual filtering pipeline we encountered putative heteroplasmic sequences in 58% of species, representing all the insect orders included, but especially among Diptera. Decreasing the OTU clustering threshold (e.g., to 95% as used by Yu *et al.*, 2010) may mask the presence of heteroplasmic sequences, but could also merge 'valid' species showing low COI divergences. Another potential complication is nuclear-mitochondrial pseudogenes (numts). Numts have been highlighted as a potential source of ambiguity for DNA barcoding (Song *et al.*, 2008), however, numts are generally easily spotted among amplified COI sequences by patterns in the amino acid translation (numts are noncoding), and probable numts would be removed when sequences were aligned and translated into amino acids.

Previous metabarcoding studies have all reported high levels of unexpected OTU (sequences) (13–35%; Hajibabaei *et al.*, 2011; Liu *et al.*, 2013). Our study was no exception with 22% (mlCOIintF/HCO2198) to 39% (LepF1/MLepF1-Rev) of OTU surviving the metagenomics pipeline being probable contaminants or error sequences. However, this was reduced to <10% using the manual pipeline. Contamination may be caused by environmental DNA in the field or laboratory, if mis-priming (hybridization of sequencing primers during sequencing of libraries) occurs, if residual tissues (gut contents including mammal tissue such as detected in our study, eggs, minute specimens, etc.) or endosymbiotic bacteria (e.g.,

https://doi.org/10.1017/S0007485315000681 Published online by Cambridge University Press

Wolbachia) are present in pooled samples, or be carried over from previous sequencing runs using the same MID (Hajibabaei *et al.*, 2011; Kircher *et al.*, 2011; Liu *et al.*, 2013; Zhou *et al.*, 2013; Nelson *et al.*, 2014; Shokralla *et al.*, 2014). In a real sample (when no corresponding Sanger sequences are available), it will be difficult to detect contamination and for particularly sensitive work, including that with legal ramifications and the detection of invasive species (Boykin *et al.*, 2012), Yu *et al.* (2012) suggested that specialized protocols such as those followed in ancient DNA laboratories will be necessary. Another suggestion has been to split samples and conduct independent amplification and sequencing on both halves followed by statistical comparison with detected error reads (Lange, 2015). This of course would massively increase cost. Options for the elimination and detection of contamination without the need (and expense) of using ancient DNA protocols or additional sequencing runs include not reusing MID tags, sequencing extraction blank controls, and including technical replicates on the same HTS run.

Conclusion

Metabarcoding is a young field. Despite the successes of early empirical studies, a number of commentaries, simulations and thought-experiments have been published highlighting perceived shortcomings (e.g., Coissac *et al.*, 2012; Cristescu, 2014; Deagle *et al.*, 2014; Piñol *et al.*, 2014*b*), in particular, many focus around marker choice. This is reminiscent of the wave of publications following Hebert *et al.* (2003) (e.g., Rubinoff *et al.*, 2006) which decreased abruptly after the build-up of empirical data clearly demonstrated the technical plausibility and utility of COI as a standard marker to recognize species boundaries (e.g., Smith *et al.*, 2008*a*). The single primer sets used here, both targeting the COI barcode region, showed an acceptable detection rate for arthropod biodiversity analysis but there were complications due to putative heteroplasmic sequences and contamination. Considering the higher detection rate, and lower components of unexpected sequences after filtering, the mlCOIintF/HCO2198 primer set seems to best target around which to develop future protocols. To date, most metabarcoding-type studies have focused only on the efficiency of pipelines (detection rate), but factors such as cost, time and ease-of-use of the bioinformatics pipeline, that are crucial for making the leapt from demonstration studies to a real-world application have not been realistically addressed. Our study suggests that DNA metabarcoding is slowly becoming as easy, fast and cheap as conventional DNA barcoding, and that Malaise trap metabarcoding may soon fulfill its potential, providing a thermometer for biodiversity.

Supplementary material

The supplementary material for this article can be found at http://www.journals.cambridge.org/BER

Acknowledgements

This study was supported by a High Impact Research Grant (Chancellery) from the University of Malaya to J. J. W. and P. E. L.; K. W. S. and P. S. L. are supported by postgraduate scholarships from the University of Malaya. Yee Yoke Pei (Ecology and Biodiversity Program, UM) and Narong Jaturas (Department of Parasitology, UM) assisted in sample collection and laboratory work. M. Sugumaran (coordinator of Rimba Ilmu Botanic Garden) provided permission to deploy the Malaise trap. H. M. G. is grateful to the Monash University Malaysia Tropical Medicine and Biology Platform for financial and infrastructure support.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410.

Bartram, A.K., Lynch, M.D., Stearns, J.C., Moreno-Hagelsieb, G. & Neufeld, J.D. (2011) Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. *Applied and Environmental Microbiology* **77**, 3846–3852.

Bohmann, K., Monadjem, A., Lehmkuhl-Noer, C., Rasmussen, M., Zeale, M.R., Clare, E., Jones, G., Willerslev, E. & Gilbert, M.T. (2011) Molecular diet analysis of two African free-tailed bats (molossidae) using high throughput sequencing. *PLoS ONE* **6**, e21441.

Boyer, F., Mercier, C., Bonin, A., Taberlet, P. & Coissac, E. (2014) OBITools: a Unix-inspired software package for DNA metabarcoding. Available online at http://metabarcoding.org/obitools/doc/index.html (accessed 28 February 2015).

Boykin, L.M., Armstrong, K.F., Kubatko, L. & De Barro, P. (2012). Species delimitation and global biosecurity. *Evolutionary Bioinformatics Online* **8**, 1–37.

Bucklin, A., Steinke, D. & Blanco-Bercial, L. (2011) DNA barcoding of marine metazoa. *Annual Review of Marine Science* **3**, 471–508.

Burgar, J.M., Murray, D.C., Craig, M.D., Haile, J., Houston, J., Stokes, V. & Bunce, M. (2014) Who's for dinner? High-throughput sequencing reveals bat dietary differentiation in a biodiversity hotspot where prey taxonomy is largely undescribed. *Molecular Ecology* **23**, 3605–3617.

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.

Coissac, E., Riaz, T. & Puillandre, N. (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology* **21**, 1834–1847.

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.

Deagle, B.E., Jarman, S.N., Coissac, E., Pompanon, F. & Taberlet, P. (2014) DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology Letters* **10**, 20140562.

Dereeper, A., Guignon, V., Blanc, G., Audic, S. & Buffet, S. (2008). Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Research* **36**, W465–W469.

Edgar, R.C. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* **10**, 996–998.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C. & Knight, R. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194–2200.

Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152.

Fišer-Pečnikar, Ž. & Buzan, E.V. (2014) 20 years since the introduction of DNA barcoding: from theory to application. *Journal of Applied Genetics* **55**, 43–52.

https://doi.org/10.1017/S0007485315000681 Published online by Cambridge University Press

Floyd, R.M., Wilson, J.J. & Hebert, P.D.N. (2009) DNA barcodes and insect biodiversity. pp. 417–431 *in* Foottit, R.G. & Aler, P.H. (*Eds*) *Insect Biodiversity: Science and Society*. Oxford, Blackwell Publishing Ltd.

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.

Gibson, J., Shokralla, S., Porter, T.M., King, I., van-Konynenburg, S., Janzen, D.H., Hallwachs, W. & Hajibabaei, M. (2014) Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 8007–8012.

Gillison, A.N., Bignell, D.E., Brewer, K.R.W., Fernandes, E.C.M. & Jones, D.T. (2013) Plant functional types and traits as biodiversity indicators for tropical forests: two biogeographically separated case studies including birds, mammals and termites. *Biodiversity and Conservation* **22**, 1909–1930.

Glenn, T.C. (2014) 2014 NGS Field Guide: Overview (The Molecular Ecologist). Available online at http://www.molecularecologist.com/next-gen-fieldguide-2014/ (accessed 28 February 2015).

Hall, T.A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* **41**, 95–98.

Haas, B.J., Gevers, D., Earl, A.M., Feldgarden, M., Ward, D.V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S.K., Sodergren, E., Methé, B., DeSantis, T.Z., Petrosino, J.F., Knight, R. & Birren, B.W. (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research* **21**, 494–504.

Hart, L.A., Bowker, M.B., Tarboton, W. & Downs, C.T. (2014) Species composition, distribution and habitat types of Odonata in the iSimangaliso Wetland Park, KwaZulu-Natal, South Africa and the associated conservation implications. *PLoS ONE* **9**, e92588.

Hajibabaei, M., Smith, M.A., Janzen, D.H., Rodriguez, J.J. & Whitefield, J.B. (2006) A minimalist barcode can identify a specimen whose DNA is degraded. *Molecular Ecology* **6**, 959–964.

Hajibabaei, M., Shokralla, S., Zhou, X., 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.

Hannon Lab. (2014) FASTX-Toolkit: FASTQ/A short-reads preprocessing tools. Available online at http://hannonlab.cshl.edu/fastx_toolkit/index.html (accessed 5 May 2015).

Hebert, P.D.N., Cywinska, A., Ball, S.L. & deWaard, J.R. (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B, Biological Sciences* **270**, 313–321.

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.

Hope, P.R., Bohmann, K., Gilbert, M.T., Zepeda-Mendoza, M.L., Razgour, O. & Jones, G. (2014) Second generation sequencing and morphological faecal analysis reveal unexpected foraging behaviour by *Myotis nattereri* (Chiroptera, Vespertilionidae) in winter. *Frontiers in Zoology* **11**, 39.

Ishii, K. & Fukui, M. (2001) Optimization of annealing temperature to reduce bias caused by a primer mismatch in multitemplate PCR. *Applied and Environmental Microbiology* **67**, 3753–3755.

Ji, Y., Ashton, L., Pedley, S.M., Edwards, D.P., Tang, Y., Nakamura, A., Kitching, R., Dolman, P.M., Woodcock, P., Edwards, F.A., Larsen, T.H., Hsu, W.W., Benedick, S., Hamer, K.C., Wilcove, D.S., Bruce, C., Wang, X., Levi, T., Lott, M., Emerson, B.C. & Yu, D.W. (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters* **16**, 1245–1257.

Kress, W.J., García-Robledo, C., Uriarte, M. & Erickson, D.L. (2015) DNA barcodes for ecology, evolution, and conservation. *Trends in Ecology and Evolution* **30**, 25–35.

Kircher, M., Heyn, P. & Kelso, J. (2011) Addressing challenges in the production and analysis of Illumina sequencing data. *BMC Genomics* **12**, 382.

Korasaki, V., Lopes, J., Gardner-Brown, G. & Louzada, J. (2013) Using dung beetles to evaluate the effects of urbanization on Atlantic Forest biodiversity. *Insect Science* **20**, 393–406.

Lahr, D.J. & Katz, L.A. (2009) Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *BioTechniques* **47**, 857–866.

Lange, A. (2015) Statistical analysis of amplicon data of the same sample to identify artefacts. Available online at http://cran.r-project.org/web/packages/AmpliconDuo/AmpliconDuo.pdf (accessed 5 May 2015).

Lee, P.S., Sing, K.W. & Wilson, J.J. (2015) Reading mammal diversity from flies: the persistence period of amplifiable mammal mtDNA in blowfly guts (*Chrysomya megacephala*) and a new DNA mini-barcode target. *PLoS ONE* **10**, e0123871.

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.

Leray, M., Yang, J.Y., Meyer, C.P., Mills, S.C., Agudelo, N., Ranwez, V., Boehm, J.T. & Machida, R.J. (2013) A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Frontiers in Zoology* **10**, 34.

Liu, S., Li, Y., Lu, J., Su, X., Tang, M., Zhang, R., Zhou, L., Zhou, C., Yang, Q., Ji, Y., Yu, D.W. & Zhou, X. (2013) SOAPBarcode: revealing arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons. *Methods in Ecology and Evolution* **4**, 1142–1150.

Magnacca, K.N. & Brown, M.J. (2010) Mitochondrial heteroplasmy and DNA barcoding in Hawaiian *Hylaeus* (*Nesoprosopis*) bees (Hymenoptera: Colletidae). *BMC Evolutionary Biology* **10**, 174.

Martin, D.P., Lemey, P., Lott, M., Moulton, V., Posada, D. & Lefeuvre, P. (2010) RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* **26**, 2462–2463.

Meusnier, I., Singer, G.A., Landry, J.F., Hickey, D.A., Hebert, P.D.N. & Hajibabaei, M. (2008) A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics* **9**, 214.

Mora, C., Tittensor, D.P., Adl, S., Simpson, A.G. & Worm, B. (2011) How many species are there on Earth and in the ocean? *PLOS Biology* **9**, e1001127.

Moulton, M.J., Song, H. & Whiting, M.F. (2010) Assessing the effects of primer specificity on eliminating numt coamplification in DNA barcoding: a case study from Orthoptera (Arthropoda: Insecta). *Molecular Ecology Resources* **10**, 615–627.

https://doi.org/10.1017/S0007485315000681 Published online by Cambridge University Press

Nelson, M.C., Morrison, H.G., Benjamino, J., Grim, S.L. & Graf, J. (2014) Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. *PLoS ONE* **9**, e94249.

Piñol, J., San-Andrés, V., Clare, E.L., Mir, G. & Symondson, W. O. (2014*a*) A pragmatic approach to the analysis of diets of generalist predators: the use of next-generation sequencing with no blocking probes. *Molecular Ecology Resources* **14**, 18–26.

Piñol, J., Mir, G., Gomez-Polo, P. & Agustí, N. (2014*b*) 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.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P. & Gu, Y. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**, 341.

Ratnasingham, S. & Hebert, P.D.N. (2007) BOLD: the Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes* **7**, 355–364.

Razgour, O., Clare, E.L., Zeale, M.R., Hanmer, J., Schnell, I.B., Rasmussen, M., Gilbert, T.P. & Jones, G. (2011) High-throughput sequencing offers insight into mechanisms of resource partitioning in cryptic bat species. *Ecology and Evolution* **1**, 556–570.

Rozen, S. & Skaletsky, H.J. (2000) Primer3 on the www for general users and for biologist programmers. pp. 365–386 *in* Krawetz, S. & Misener, S. (*Eds*) *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. New Jersey, Humana Press.

Rubinoff, D., Cameron, S. & Will, K. (2006) A genomic perspective on the shortcomings of mitochondrial DNA for 'barcoding' identification. *Journal of Heredity* **97**, 581–594.

Russo, L., Stehouwer, R., Heberling, J.M. & Shea, K. (2011) The composite insect trap: an innovative combination trap for biologically diverse sampling. *PLoS ONE* **6**, e21079.

Shokralla, S., Gibson, J.F., Nikbakht, H., Janzen, D.H., Hallwachs, W. & Hajibabaei, M. (2014) Next-generation barcoding: using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Molecular Ecology Resources* **14**, 892–901.

Sipos, R., Székely, A.J., Palatinszky, M., Révész, S., Márialigeti, K. & Nikolausz, M. (2007) Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiology Ecology* **60**, 341–350.

Smith, M.A., Fisher, B.L. & Hebert, P.D.N. (2005) DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **360**, 1825–1834.

Smith, M.A., Poyarkov, N.A. Jr. & Hebert, P.D.N. (2008*a*) CO1 DNA barcoding amphibians: take the chance, meet the challenge. *Molecular Ecology Resources* **8**, 235–246.

Smith, M.A. & Rodriguez, J.J., Whitfield, J.B., Deans, A.R., Janzen, D.H., Hallwachs, W. & Hebert, P.D.N. (2008*b*) Extreme diversity of tropical parasitoid wasps exposed by iterative integration of natural history, DNA barcoding, morphology, and collections. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 12359–12364.

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.

Stevens, J.L., Jackson, R.L. & Olson, J.B. (2013) Slowing PCR ramp speed reduces chimera formation from environmental samples. *Journal of Microbiological Methods* **93**, 203–205.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution* **30**, 2725–2729.

Tang, M., Tan, M., Meng, G., Yang, S., Su, X., Liu, S., Song, W., Li, Y., Wu, Q., Zhang, A. & Zhou, X. (2014) Multiplex sequencing of pooled mitochondrial genomes-a crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Research* **42**, e166.

Townsend, J.P. & Rand, D.M. (2004) Mitochondrial genome size variation in New World and Old World populations of *Drosophila melanogaster*. *Heredity* **93**, 98–103.

Triplehorn, C.A. & Johnson, N.F. (2005) *Borror and DeLong's Introduction to the Study of Insects*. 7th edn. California, Thomson Brooks/Cole.

Vesterinen, E.J., Lilley, T., Laine, V.N. & Wahlberg, N. (2013) Next generation sequencing of fecal DNA reveals the dietary diversity of the widespread insectivorous predator Daubenton's Bat (*Myotis daubentonii*) in Southwestern Finland. *PLoS ONE* **8**, e82168.

Wilson, J.J. (2012) DNA barcodes for insects. pp. 17–45 *in* Kress, W.J. & Erikson, D.L. (*Eds*) *DNA Barcodes: Methods and Protocols*. New York, Humana Press.

Wilson, J.J. & Sing, K.W. (2013) DNA barcoding can successfully identify *Penaeus monodon*, associate life cycle stages, and generate hypotheses of unrecognized diversity. *Sains Malaysiana* **42**, 1827–1829.

Wilson, J.J., Rougerie, R., Schonfeld, J., Janzen, D.H., Hallwachs, W., Hajibabaei, M., Kitching, I.J., Haxaire, J. & Hebert, P.D.N. (2011) When species matches are unavailable are DNA barcodes correctly assigned to higher taxa? An assessment using sphingid moths. *BMC Ecology* **11**, 18.

Wong, M.M., Lim, C.L. & Wilson, J.J. (2015) DNA barcoding implicates 23 species and four orders as potential pollinators of Chinese knotweed (*Persicaria chinensis*) in peninsular Malaysia. *Bulletin of Entomological Research* **105**, 515–520.

Yang, C.X., Wang, X.Y., Miller, J.A., Marleen-de-Blécourt, Ji, Y. Q., Yang, C.Y., Harrison, R.D. & Yu, D.W. (2014) Using metabarcoding to ask if easily collected soil and leaf-litter samples can be used as a general biodiversity indicator. *Ecological Indicators* **46**, 379–389.

Yu, D.W., Ji, Y.Q., Emerson, B.C., Wang, X.Y., Ye, C.X., Yang, C.Y. & Ding, Z.L. (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution* **4**, 613–623.

Zeale, M.R.K., Butlin, R.K., Barker, G.L.A., Lees, D.C. & Jones, G. (2011) Taxon-specific PCR for DNA barcoding arthropod prey in bat faeces. *Molecular Ecology Resources* **11**, 236–244.

Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. (2014) PEAR: a fast and accurate Illumina paired-end reAD mergeR. *Bioinformatics* **30**, 614–620.

Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X., Zhou, L., Tang, M., Fu, R., Li, J. & Huang, Q. (2013) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *GigaScience* **2**, 4.

Zografou, K., Kati, V., Grill, A., Wilson, R.J., Tzirkalli, E., Pamperis, L.N., Halley, J.M. (2014) Signals of climate change in butterfly communities in a Mediterranean protected area. *PLoS ONE* **9**, e87245.

https://doi.org/10.1017/S0007485315000681 Published online by Cambridge University Press