Received: 5 September 2017 | Revised: 20 December 2017 | Accepted: 20 December 2017

DOI: 10.1111/1755-0998.12751

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

Sorting specimen-rich invertebrate samples with cost-effective NGS barcodes: Validating a reverse workflow for specimen processing

Wendy Y. Wang[1] | Amrita Srivathsan[2] | Maosheng Foo[1] | Seiki K. Yamane[3] | Rudolf Meier[1,2]

[1]Lee Kong Chian Natural History Museum, Faculty of Science, National University of Singapore, Singapore

[2]Evolutionary Biology Laboratory, Department of Biological Sciences, National University of Singapore, Singapore

[3]Kagoshima University Museum, Kagoshima, Japan

Correspondence
Rudolf Meier, Lee Kong Chian Natural History Museum, Faculty of Science, National University of Singapore, Singapore.
Email: meier@nus.edu.sg

Funding information
MOE, Grant/Award Number: R-154-000-A22-112; NUS, Grant/Award Number: R-154-000-648-646, R-154-000-648-733

Abstract

Biologists frequently sort specimen-rich samples to species. This process is daunting when based on morphology, and disadvantageous if performed using molecular methods that destroy vouchers (e.g., metabarcoding). An alternative is barcoding every specimen in a bulk sample and then presorting the specimens using DNA barcodes, thus mitigating downstream morphological work on presorted units. Such a "reverse workflow" is too expensive using Sanger sequencing, but we here demonstrate that is feasible with an next-generation sequencing (NGS) barcoding pipeline that allows for cost-effective high-throughput generation of short specimen-specific barcodes (313 bp of COI; laboratory cost <$0.50 per specimen) through next-generation sequencing of tagged amplicons. We applied our approach to a large sample of tropical ants, obtaining barcodes for 3,290 of 4,032 specimens (82%). NGS barcodes and their corresponding specimens were then sorted into molecular operational taxonomic units (mOTUs) based on objective clustering and Automated Barcode Gap Discovery (ABGD). High diversity of 88–90 mOTUs (4% clustering) was found and morphologically validated based on preserved vouchers. The mOTUs were overwhelmingly in agreement with morphospecies (match ratio 0.95 at 4% clustering). Because of lack of coverage in existing barcode databases, only 18 could be accurately identified to named species, but our study yielded new barcodes for 48 species, including 28 that are potentially new to science. With its low cost and technical simplicity, the NGS barcoding pipeline can be implemented by a large range of laboratories. It accelerates invertebrate species discovery, facilitates downstream taxonomic work, helps with building comprehensive barcode databases and yields precise abundance information.

KEYWORDS
community ecology, DNA barcoding, insects, invertebrates, systematics

1 | INTRODUCTION

Invertebrate biologists often process specimen- and species-rich bulk samples for purposes such as biomonitoring, biodiversity assessment and taxonomic revision (Miller, Hausmann, Hallwachs, & Janzen, 2016; Morinière et al., 2016). Traditionally, presorting of such bulk samples is based on morphology, but few invertebrate taxonomists have the time for species-level sorting, and many taxa lack

identification resources such as species-level keys (Aagaard et al., 2017; Gotelli, 2004). A potential alternative for species-level sorting is grouping specimens based on DNA barcodes, followed by validating molecular operational taxonomic units (mOTUs) using morphology. However, processing specimen-rich samples using such a strict "reverse workflow"—where all specimens are barcoded first before morphological validation—is very expensive with Sanger sequencing. For example, according to the pricing on the Canadian Centre for DNA Barcoding (CCDB) website (Centre for Biodiversity Genomics©: http://ccdb.ca/pricing/), the cost of Sanger barcodes is ca. USD 18 (CAD 23) per tissue sample; that is barcoding the 4,032 ant specimens processed in this study would cost more than USD 72,000.

High cost is presumably the main reason why large-scale studies do not adopt a reverse workflow for sorting invertebrate samples. Instead, such studies start with morphology-based presorting (usually carried out by parataxonomists), followed by testing of morphospecies by barcoding a few representative specimens per morphospecies. A typical example is Tänzler, Sagata, Surbakti, Balke, and Riedel's (2012) study of *Trigonopterus* weevils which involved barcoding 1,002 of the ~6,500 edaphic weevil specimens that were collected over 4 years and which were presorted into morphospecies prior to barcoding. Similarly, Renaud, Savage, and Adamowicz (2012) presorted 1,303 muscid flies based on morphology before barcoding a subsample. Even the largest barcoding study to date (>1 million barcodes) (Hebert et al., 2016) employed presorting, although the Canadian Centre for DNA Barcoding is the best-funded institution in this area of research.

Such a hybrid approach has several drawbacks. If presorting is carried out by parataxonomists, the validity and accuracy of morphospecies are uneven and unpredictable (Krell, 2004). This is undesirable because errors in presorting can have serious cascading effects (Bortolus, 2008) and interfere with subsequent testing of the morphospecies units with DNA barcodes. In addition, selectively mixing morphological and molecular techniques also makes it difficult to objectively assess the levels of conflict between molecular and morphological data, because not all specimens are studied using both types of data. This lack of independence is avoided when a strict reverse workflow—where all specimens are barcoded—is adopted. After barcoding, specimens are grouped into mOTUs, before being validated by taxonomic experts based on morphology.

The advent of affordable high-throughput next-generation sequencing (NGS) technologies has generated additional options for assessing bulk invertebrate samples. Numerous studies have demonstrated that metabarcoding and metagenomics can rapidly generate large amounts of sequence data for complex, mixed samples which can be used for biomonitoring, building phylogenies and studying community ecology (e.g., Crampton-Platt et al., 2015; Gómez-Rodríguez, Crampton-Platt, Timmermans, Baselga, & Vogler, 2015; Hajibabaei, Spall, Shokralla, & van Konynenburg, 2012; Ji et al., 2013; Morinière et al., 2016; Yu et al., 2012). However, these techniques also have drawbacks. First, DNA sequences cannot be traced back to voucher specimens because vouchers are either completely or partially destroyed during DNA extraction. Second, the bioinformatics of metabarcoding data is not straightforward because most sequences cannot be identified to species based on online databases. It is thus difficult to assess whether similar reads are sequence variants or evidence for closely related species in the sample (e.g., Gompert et al., 2014). This is a significant problem because ca. 30% of all species are rare (Lim et al., 2016), and a recent study suggests that random sampling of DNA during metabarcoding affects the discovery probability for these rare species (Leray & Knowlton, 2017). Third, it is unclear how one can obtain reliable abundance and biomass information from metabarcoding data although such data are often needed for ecological analyses (Gibb et al., 2017). Lastly, metabarcoding studies of poorly known faunas tend to generate large numbers of barcodes that cannot be identified to species (Lim et al., 2016; Srivathsan, Ang, Vogler, & Meier, 2016; Srivathsan, Sha, Vogler, & Meier, 2015), because the available barcoding databases typically have poor species coverage, especially for invertebrates (Kwong, Srivathsan, & Meier, 2012). This interferes with biological interpretation of the data because unidentified sequences remain singular observations, while sequences that can be identified to species yield scientific names that allow for accessing the relevant scientific literature.

In the light of these issues, we here propose and test a NGS barcoding-based "reverse workflow," which can be used for processing specimen-rich invertebrate samples cost-effectively. With its relative simplicity and reliance on established common laboratory techniques, this workflow can be used widely. If adopted by a sufficiently large number of laboratories across continents, it would accelerate the building of comprehensive and curated taxonomic barcode databases for invertebrates, which are also essential for in-depth interpretation of metabarcoding data. The basic techniques for obtaining NGS barcodes were described in Meier, Wong, Srivathsan, and Foo (2016), but our study is the first to upscale and apply this workflow to a sample that is sufficiently large to derive reliable success rates and cost estimates. In addition, we explicitly assess whether the mOTUs obtained with short NGS barcodes (313 bp) are congruent with morphospecies and optimize the procedures for an arthropod taxon commonly used in studies on terrestrial biodiversity, that is ants (Formicidae).

Ants comprise a major component of most terrestrial ecosystems, and by virtue of their abundance, often influence or perform critical ecosystem functions. Despite their abundance in insect surveys, there are no sizable studies on ant diversity that use a strict DNA barcoding approach involving the generation of barcodes for all specimens. Instead, most studies adopt hybrid approaches that start with presorting based on morphology (e.g., Smith, Fisher, & Hebert, 2005: 268 barcodes from 280 specimens; Delsinne et al., 2012: 187 barcodes for 10,260 ants representing ca. 70 morphospecies). We here demonstrate that NGS barcodes can be used for obtaining reliable species estimates for ant samples based on a reverse workflow. Successful implementation of the NGS pipeline could mitigate many challenges associated with biodiversity surveys of this ecologically important family. This includes the matching of reproductive individuals to conspecific sterile workers.

2 | MATERIALS AND METHODS

2.1 | Sampling

Ant samples were collected as part of a larger insect survey based on Malaise trapping that was conducted over six consecutive months in 2015 (April–September) and included four forested sites (>1 km apart) across the campus of the National University of Singapore. Each site was named after the nearest adjoining man-made structure: (i) Prince George's Park Residences (PGP) (N 1.2924°, E 103.7787°); (ii) I-Cube Building (ICube) (N 1.2935°, E 103.7763°); (iii) University Hall (Uhall) (N 1.2971°, E 103.7766°); and (iv) University Town (Utown) (N 1.3062°, E 103.7746°). One Malaise trap was set up in each site, and trap samples were collected on a weekly basis. Samples from each collection event were allocated a unique registration number to facilitate specimen tracking. Samples were presorted to insect orders or families based on morphology. Ants (Hymenoptera: Formicidae) collected over the entire 6 months were selected for this study. As Malaise traps mainly capture flying insects, samples consisted predominantly of winged reproductives (ant queens and males).

2.2 | DNA barcoding and next-generation sequencing

Direct polymerase chain reactions (dPCR) were used to amplify DNA barcodes for each specimen. Amplification procedures were adapted from Meier et al. (2016) and Wong et al. (2014), and optimized for Formicidae; the protocol was upscaled to a 96-well microplate format. In the optimized protocol, a piece of tissue approximately 0.5–1.5 mm^2 in mass was dissected from each specimen and used as template for dPCR. Tissue was dissected from a single leg, the specific portions of which depended on body and corresponding leg size: (i) large (>9 mm body length) specimens—a small portion of the femur; (ii) mid-sized (3–9 mm) specimens—half or whole femur; (iii) small (1.5–3 mm) specimens—femur and tibia (tarsus removed); and (iv) very small specimens (<1.5 mm body length)—whole individual. To prevent ethanol contamination, forceps used to retrieve specimens from ethanol storage tubes were kept separate from those used to insert template tissue into the dPCR mastermix. Forceps were also cleaned with sterile cut C-fold towels in between dissections. Dissected specimens were subsequently stored in individual ethanol vials; whole individuals used as template were also retrieved with sterile pipette tips from their respective reaction tubes after dPCR and stored in ethanol vials. Each specimen vial was allocated a label with a unique identifier code (ZRC_BDPXXX), which also serves as the museum accession number. This identifier was incorporated downstream into each individual barcode sequence header in *.fasta* format, enabling barcodes to be traced to their respective specimens-of-origin. Specimen vials were stored in the same order as their PCR positions on the 96-well plate; this arrangement was recorded in a spreadsheet.

A 313-bp fragment of cytochrome oxidase I [COI; m1COIintF: 5′-GGWACWGGWTGAACWGTWTAYCCYCC-3′ (Leray et al., 2013) and modified jgHCO2198: 5′-TANACYTCNGGRTGNCCRAARAAYCA-3′ (Geller, Meyer, Parker, & Hawk, 2013)/5′-TAAACYTCAGGRTGCCCRAARAAYCA-3′ (Meier et al., 2016)] was amplified using labelled forward and reverse primers. Each label was 9-bp long, differing from other labels by ≥3 bp; labels were generated using the online freeware "Barcode Generator" (http://comailab.genomecenter.ucdavis.edu/index.php/Barcode_generator). In this study, we used an assortment of different combinations from 250 available pairs of labelled primers; each specimen barcode was amplified with a uniquely labelled primer combination. Each position on the 96-well microplate was linked to a unique tagged primer combination. The number of pairs of labelled primers required depends on the number of libraries going into each MiSeq run. For example with three libraries in one run, about 70 primer pairs are needed. PCR mixtures of 20 μl reaction volume each were prepared (2 μl of 10× BioReady rTaq 10× Reaction Buffer, 1.5 μl of 2.5 mM dNTP mix, 0.2 μl of BioReady rTaq DNA polymerase, 2 μl each of 5 μM forward and reverse primers and 2 μl of 1 mg/ml Bovine Serum Albumin, RNase/DNase-free sterile water), and cycling conditions were as follows: initial denaturation at 94°C for 5 min, 35 cycles of denaturation at 94°C for 1 min, annealing at 47°C for 2 min and extension at 72°C for 1 min, thereafter a final extension at 72°C for 5 min.

Amplified PCR products were combined and cleaned in aliquots of up to 100 μl using SureClean (Bioline Inc., London, UK); the cleaned amplicon products were then re-eluted in RNase/DNase-free water. NGS libraries were prepared for the PCR pools using the TruSeq Nano DNA Library Preparation kit and sequenced on an Illumina MiSeq 2 × 300 bp platform. Paired-end (PE) read data (*.fastq*) were assembled using PEAR version 0.9.6 (Zhang, Kobert, Flouri, & Stamatakis, 2014).

2.3 | Demultiplexing and quality filters

We followed the barcoding pipeline and scripts in Meier et al. (2016) for downstream processing of PEAR-assembled PE reads. Demultiplexing of sequences, whereby PE reads are assigned to their respective specimen-of-origin, was carried out with a Python script. For demultiplexing, no mismatch was allowed for the tag region while two mismatches were allowed for the primer sequence. Besides demultiplexing reads, this script also generates the following output in tabular format: (i) total number of reads for each specimen/sample (total read coverage); (ii) total count for the largest set of identical reads (barcodes) including their length variants (total barcode count); (iii) ratio of the read counts for the second-largest set of identical reads to the number of reads in the largest set; and (iv) nucleotide sequences for the largest read set and second-largest read set.

Demultiplexed data were screened with a series of quality filters. For a specimen to be considered successfully barcoded, total read coverage must be >50, followed by total barcode count >10, and finally, coverage for the second-largest set of identical reads should not exceed 20% of the total barcode count, that is ratio ≤0.2. Successful barcodes were checked for contamination against the

GenBank (NCBI) nucleotide database (Benson et al., 2013), using the Unix® command-line version of NCBI Basic Local Alignment Search Tool (BLAST) version 2.6.0+ (Altschul, Gish, Miller, Myers, & Lipman, 1990). BLAST searches were performed under default parameters in *Megablast* (word size: 28). Barcodes with top hits to anything other than ants (Formicidae) were removed.

2.4 | mOTU estimation

The final set of barcodes was aligned using MAFFT version 7 (Katoh & Standley, 2013); alignments were checked on MEGA 6 (Tamura, Stecher, Peterson, Filipski, & Kumar, 2013). We used two methods of grouping the aligned sequences into mOTUs: objective clustering (Meier, Shiyang, Vaidya, & Ng, 2006) and Automatic Barcode Gap Discovery (ABGD) analysis (Puillandre, Lambert, Brouillet, & Achaz, 2012). Objective clustering as implemented in SpeciesIdentifier (TaxonDNA 1.6.2; Meier et al., 2006) was used to determine mOTUs at different percentage thresholds (0%–10%). In objective clustering, sequences are grouped according to uncorrected *p*-distances (Meier, Zhang, & Ali, 2008; Meier et al., 2006; Srivathsan & Meier, 2012)—members of a set of putative conspecific sequences have at least one match to a sequence in the set that falls within the specified threshold distance. Multiple thresholds were applied (0%–10%) to test stability of the results. Cluster splitting and/or merging events amongst individual sequences were visualized using a custom-designed software—OBJ_CLUST version 0.1.2 (A. Srivathsan, unpublished; an implementation of objective clustering as described by Meier et al., 2006).

Automated Barcode Gap Discovery analyses were conducted using the command-line version on Unix. We used simple distance and ran ABGD on the following parameters: p_{min} = .005, p_{max} = .1 and no. of steps = 20. The mOTU assignments over 20 recursions for prior maximum intraspecific divergence (*p*) values between .005 and .1 were recorded.

2.5 | Morphological verification of mOTUs

To check whether mOTU delimitation was in agreement with morphology both within and between mOTUs, barcoded specimens were identified to species using relevant taxonomic keys and reference collections (Meier, 2017: see Table S2). All specimens of a mOTU (at the chosen percentage threshold) were examined if the mOTU had ≤20 individuals. For mOTUs with >20 members, 20 specimens were subsampled per mOTU for morphological verification according to a framework aimed at optimizing representation across time and space. If there were specimens collected across multiple sites and months, specimens from every site and month sampled were drawn for examination. Sorting of specimens per mOTU according to this framework was easily achieved, because sampling locality and date information are indicated in every barcode sequence header, and linked to the corresponding specimen code. Morphological differences within a molecular cluster suggest recent species divergence, whereas morphologically identical mOTUs suggest cryptic species and/or morphological stasis. Specimens that could not be identified to described species were designated tentative morphospecies identities.

Congruence of morphological species units with mOTUs was quantified with the match ratio adapted from Ahrens et al. (2016); the match ratio is calculated as follows: $2 \times N_{match}/(N_{mOTU} + N_{morph})$, where N_{match} is the number of putative species that are the same between the two methods of delimitation, and N_{mOTU} and N_{morph} refer to the total number of mOTUs at a given percentage distance threshold and total number of morphological species units, respectively.

A BLAST check of representative 4% cluster sequences against GenBank (NCBI) was performed to assess the adequacy of the database for obtaining species-level identifications. Barcodes that could not be matched to a species with high confidence (≥96% identity at 100% query cover) were checked against databases of the Barcode Of Life Data System (BOLD; Ratnasingham & Hebert, 2007). The mOTU identities based on morphological examination of specimens were also compared with matches from GenBank and BOLD.

Finally, mounted specimens of each 4% mOTU were imaged with a Dun Inc.™ Passport II macrophotography imaging system, using a Canon MP-E 65 mm lens. Images were edited and scale bars added using Adobe Photoshop CS6; processed images were uploaded onto the Biodiversity of Singapore database (https://singapore.biodiversity.online/). All barcoded specimens were deposited at the Lee Kong Chian Natural History Museum, as part of its Zoological Reference Collection (ZRC). Successful barcodes were submitted and are accessible through GenBank (see Table S1, Supplemental Information for Accession numbers).

3 | RESULTS

A total of 4,032 ants underwent barcoding. For direct PCR, up to 200 specimens were processed per person in about 8 hr of a work day; one person could barcode all 4,032 ants in 320-hr spread over 20 full work days. Manpower costs were low because the work requires minimal training and can be accomplished by interns and students. The sequence data for the ant amplicons were obtained in two MiSeq runs that were shared with multiple other projects (see below for read numbers and coverage).

3.1 | Consumable costs

Mean consumable cost for dPCR was USD 0.16/specimen (overall USD 645.12). In total, 4,964,876 sequencing reads were generated for the 4,032 specimens; that is mean coverage was very high 1,231× per specimen. This requires 1/3 of an Illumina MiSeq run (2 × 300 bp PE) (based on an estimated output of 15,000,000 paired reads of 300 bp) and a single library (given that each specimen has been barcoded with a unique combination of indexed primers); overall sequencing cost would therefore be USD 571.67 + 190 = 761.67 (rates acc. to http://research.ncsu.edu/gsl/pricing/). Note that lowering mean coverage can significantly reduce this cost. For example, in our study, halving mean coverage (=sequencing cost) would have only resulted in the

loss of barcodes for ca. 100 specimens (due to violation of the >50 total read count criterion: Figure 1). Following quality filtering and omission of contaminant sequences, specimen success rates were on average 81% per run, and 3,290 ants were successfully barcoded. Overall, it cost USD 1,406.79/3,290 = 0.43 to barcode each specimen. Note that the main objective was not maximizing success rates, rather we emphasize rapid processing given that <100 species were expected from the sample. All successful barcodes are available in *.fasta* format as Supplementary Data (S2).

3.2 | mOTU estimation

The interspecies threshold for COI divergence across most arthropods ranges from 2% to 4% (Hebert, Cywinska, & Ball, 2003; Meier et al., 2008). Sequence clustering using objective clustering (Meier et al., 2006) yielded 94, 92 and 90 mOTUs at 2%, 3% and 4% thresholds, respectively (Table 1; 124 haplotypes; Supplementary Data S1); that is the numbers of clusters were very stable across the threshold range of 2%–4% and few specimens were involved in mOTU reassignment between the thresholds. Initial partitioning using ABGD (Puillandre et al., 2012) produced 88 mOTUs at prior intraspecific divergences of 2%–4% (p = .02–.04; Figure S1). Recursive partitioning identified 1–3 additional mOTUs (91, 90 and 89 mOTUs at 2%, 3% and 4%, respectively) within the same threshold range. Objective clustering and initial partitioning on ABGD at a 4% intraspecific threshold each generated mOTUs that were mostly identical between the two methods. The only exceptions were two cluster splits produced by objective clustering.

Downstream sorting was based on the 90 mOTUs obtained using objective clustering at 4% threshold; mOTUs that were lumped together using ABGD at the same percentage threshold were marked for later morphological investigation. One mOTU that was a singleton—n11—was lost and hence omitted from later congruence analyses. Sixty-five 4% mOTUs had 1–20 members each, while only 25 mOTUs comprised >20 specimens each (8 mOTUs: 21–40 specimens; four mOTUs; 41–60; four mOTU: 61–80 specimens; nine mOTUs >100 specimens) and were subsampled for morphological verification. All barcoded specimens were subsequently sorted and separated into their respective 4% mOTU clusters; this process was expedited by the fact that each specimen had been allocated a label with a unique identifier and positional-storage information at the point of PCR. The process of physical sorting into mOTUs and mounting of representative specimens per mOTU was carried out by one laboratory person and took approximately 1 week. The entire pipeline, combined with 10 days of molecular processing by two parties, 3 days of MiSeq sequencing and downstream demultiplexing of tagged amplicons, could be completed in less than 1 month.

TABLE 1 Relationship between threshold values (uncorrected *p*-distances) and number of mOTUs, singletons and doubletons

Clustering threshold (%)	0	1	2	3	4	5–7	8	9	10
Total no. of mOTUs	124	97	94	92	90	88	87	83	80
Singletons	41	29	27	25	24	22	22	21	17
Doubletons	16	11	11	12	12	13	13	11	12

3.3 | Congruence between mOTUs and morphology

Morphological validation of congruence between putative species units based on DNA and morphology was performed at two different levels: (i) species delimitation; (ii) and species identification. Validation of delimitation requires investigation of physical clusters based on DNA against those based on external morphology. This

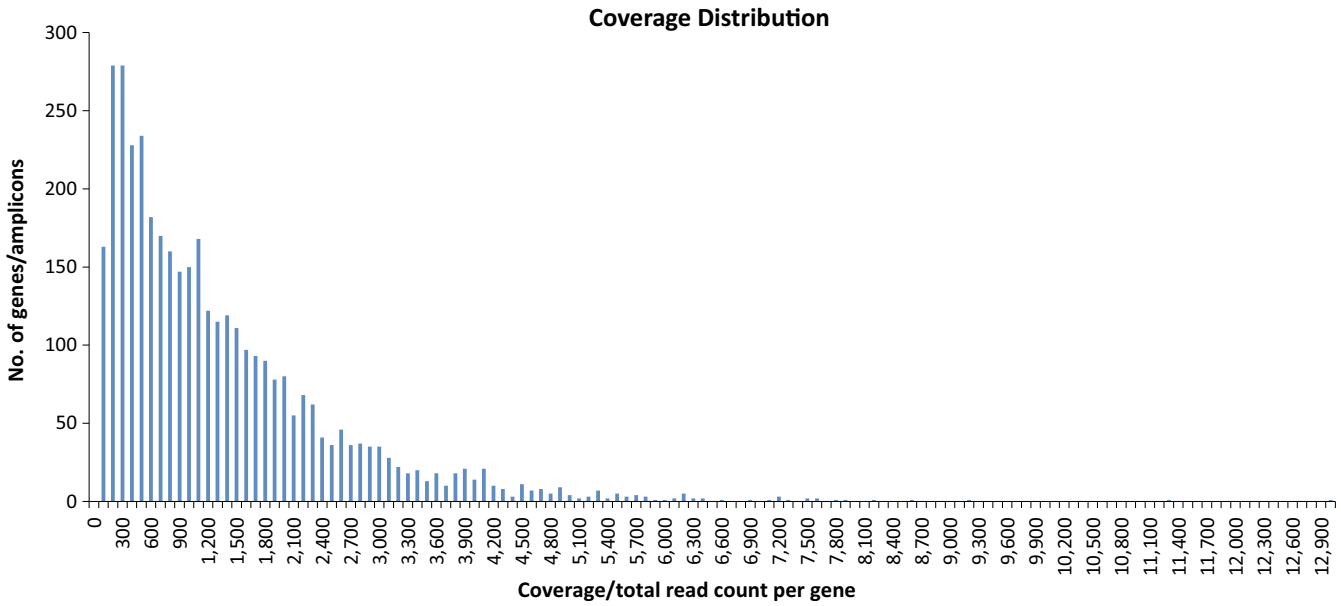

FIGURE 1 Distribution of mean coverage per gene successfully sequenced (total read count >50) via Illumina MiSeq

also means checking large mOTUs for morphological differences, which may be evidence for recently diverged species. This is necessary because COI is not a speciation gene (Kwong, Srivathsan, Vaidya, & Meier, 2012), and COI barcodes are known to fare poorly in detecting recently diverged species (Meier et al., 2008; van Velzen, Weitschek, Felici, & Bakker, 2012). In contrast to species delimitation, species identification involves assigning available species names based on identified sequences in online nucleotide databases or based on identifying the specimens in the clusters with morphological keys or reference collections (Meier, 2017).

Expert examination and verification of mounted specimens from 89 mOTUs (87 based on ABGD) required almost 2 weeks. Representative specimens from all 4% mOTUs could be identified to at least genus level based on morphology (see Table S2 for taxonomic resources and references used for identification). Morphological species units were congruent with most mOTUs examined, yielding 86 morphospecies and a 0.95 match ratio with mOTUs generated from either objective clustering or initial partitioning on ABGD. Match ratios did not improve at smaller intraspecific divergence thresholds, that is objective clustering—0.92, 0.88, ABGD—0.95 (consistent mOTU count from initial partitioning) at 2% and 3% thresholds, respectively.

There were two cases of conflict with specimens belonging to different 4% mOTUs from objective clustering, and ABGD being identical based on morphology (Table S2: n8 and n34—*Tetraponera allaborans*; n38 and n40—*Tetramorium pacificum*). There is a third case where conflict is only observed between objective clustering and morphology (Table S2: n26 and n42—*Paratrechina longicornis*) and a forth case where ABGD's initial partitioning results in lumping: two singleton mOTUs based on objective clustering (4% threshold) were also separate species based on morphology (Table S2: n05 *Tapinoma* cf.sp.13.of.SKY and n23 *Tapinoma* sp.1.of.WW).

3.4 | Species identification using barcode databases

Local BLAST searches of 4% mOTU haplotypes against Genbank or BOLD databases yielded only 42 top hits (out of 90 mOTUs) that could be confidently identified at least to genus, that is 100% query cover and ≥96% identity (Table S2). Twenty-one of these mOTUs could be identified to named species based on both database searches and morphology; 18 of 21 database matches were in agreement with morphology-based species identities (Table 2). Three database matches, however, conflicted with morphological identities: (i) mOTU n7—*Camponotus rufoglaucus* Jerdon, 1851, was matched with a database sequence derived from a specimen identified as *Polyrhachis paracamponota* at 97%; (ii) mOTU n75—*Polyrhachis* cf. *proxima* was matched with a GenBank sequence identified as *Polyrhachis beccarii*; and (iii) mOTU n78—*Tapinoma indicum* was identified on BOLD as *Tapinoma melanocephalum* at 100% (Table 2). Another eight mOTUs could only be identified to genus using both databases (i.e., reference sequences did not contain species names) and morphology; all database matches for these eight mOTUs concurred with genus-level identification based on morphology (Table 3). Finally, 13 mOTUs were matched to genera in agreement with morphology and could be identified to named species based on either morphology (seven mOTUs) or matches to existing identified database sequences (six mOTUs) (Table 4).

4 | DISCUSSION

4.1 | Tackling invertebrate diversity and building barcode databases with NGS barcoding

We here demonstrate that the NGS barcoding-based pipeline of Meier et al. (2016) can be expanded and simplified to standards suitable for invertebrate laboratories with access to minimal equipment, and staff and students with limited background in molecular biology and bioinformatics. The straightforward laboratory techniques and bioinformatics procedures can be learned within a few days, and in our study, the laboratory cost per specimen was USD 0.43. This cost can be further reduced through the use of higher throughput sequencing. Based on our results, NGS barcoding of ~5,000 specimens would only require 4% of a sequencing lane (Illumina® HiSeq2500; 2 × 250 bp; rapid run) and the use of this NGS technology would reduce the barcode cost per specimen from USD 0.43 to USD 0.17 (see http://research.ncsu.edu/gsl/pricing/). A further cost reduction to ca. USD 0.10 can be achieved by reusing plastic consumables (tips and 96-well plates contribute ca. USD 0.06 to PCR cost: Meier et al., 2016) and reducing the PCR reaction volume from 20 to 10 μl. The low cost of NGS barcoding means that even laboratories with very limited access to resources can use this pipeline for processing thousands of specimens into mOTUs. Developing such cost-effective and scalable techniques is particularly important given that much unknown invertebrate diversity resides in countries with limited funding for scientific research.

Our reverse workflow—consisting of NGS barcoding with subsequent morphological work—is not only technically simple, but also particularly suitable for efficient building of barcode databases. The latter are currently woefully incomplete for invertebrates (Ball, Hebert, Burian, & Webb, 2005; Kwong, Srivathsan, & Meier, 2012; Kwong, Srivathsan, & Vaidya et al., 2012; Renaud et al., 2012), which interferes with evaluation of data generated from metabarcoding or metagenomics (Lim et al., 2016; Srivathsan et al., 2015, 2016). Most sequences in such studies cannot be identified to species/genus, that is the scientific literature for these unidentified species cannot be consulted for additional insights. Our study illustrates the extent of this problem—fewer than half of all mOTUs, that is 42, generated in this study could be identified to genus/species using online databases (Table S2), and only 19 of these could be confidently identified to named species in agreement with morphology (Table 2). However, we also show how our NGS-based pipeline can help address this problem. In our study, barcodes are generated first and the species are then identified based on vouchers examined post hoc by experts using morphology. Note that this task is facilitated by the molecular presorting which assigned the >4,000 ants into 90 mOTUs; that is taxonomic experts do not have to spend time on presorting. Through the strict application of the reverse

TABLE 2 Species identification of 4% mOTUs based on Genbank BLAST and BOLDSYSTEMS. Only 4% mOTUs with ≥96% matches in identity are shown

	mOTU code	Database, BLAST % identity	GenBank/BOLD sp. ID	Morphology-based sp. ID	Conflict or agreement (C/A)	
					Genus	Species
1	n06	GenBank,98	*Acropyga acutiventris*	*Acropyga acutiventris*	A	A
2	n10	GenBank,100	*Crematogaster* sp.10.of.SKY (*C. treubi)*	*Crematogaster treubi*	A	A
3	n26	GenBank,100	*Paratrechina longicornis*	*Paratrechina longicornis*	A	A
4	n28	BOLD,100	*Pyramica* (syn. *Strumigenys*) *eggersi*	*Strumigenys eggersi*	A	A
5	n29	BOLD,98.71	*Technomyrmex kraepelini*	*Technomyrmex kraepelini*	A	A
6	n31	GenBank,99	*Tetramorium bicarinatum*	*Tetramorium bicarinatum*	A	A
7	n33	GenBank,100	*Anoplolepis gracilipes*	*Anoplolepis gracilipes*	A	A
8	n36	GenBank,100	*Strumigenys emmae*	*Strumigenys emmae*	A	A
9	n40	GenBank,100	*Tetramorium pacificum*	*Tetramorium pacificum*	A	A
10	n42	GenBank,100	*Paratrechina longicornis*	*Paratrechina longicornis*	A	A
11	n43	BOLD,100	*Tetramorium tonganum*	*Tetramorium tonganum*	A	A
12	n48	GenBank,100	*Polyrhachis proxima*	*Polyrhachis proxima*	A	A
13	n50	GenBank,100	*Monomorium floricola*	*Monomorium floricola*	A	A
14	n58	GenBank,99	*Cardiocondyla obscurior*	*Cardiocondyla obscurior*	A	A
15	n59	BOLD,98.38	*Strumigenys nepalensis*	*Strumigenys nepalensis*	A	A
16	n60	GenBank,99	*Oecophylla smaragdina*	*Oecophylla smaragdina*	A	A
17	n64	BOLD,97.09	*Tapinoma melanocephalum*	*Tapinoma melanocephalum*	A	A
18	n89	GenBank,97	*Polyrhachis armata*	*Polyrhachis armata*	A	A
Conflict cases						
19	n07	GenBank,97	*Polyrhachis paracamponota*	*Camponotus rufoglaucus*	C	C
20	n75	GenBank,100	*Polyrhachis beccarii*	*Polyrhachis* cf. *proxima*	A	C
21	n78	BOLD,100	*Tapinoma melanocephalum*	*Tapinoma indicum*	A	C

Conflict or agreement between database-matched and morphology-based identities is indicated in the last column.

TABLE 3 Genus-level identification of 4% mOTUs

	mOTU code	Database, BLAST % identity	GenBank/BOLD sp. ID	Morphology-based sp. ID	Conflict or agreement (C/A) Genus
1	n04	BOLD,99.68	*Philidris* MY04	*Philidris* sp.4.of.SKY	A
2	n22	BOLD,97.41	*Solenopsis* sp.	*Solenopsis* sp.15.of.SKY	A
3	n35	GenBank,100	*Nylanderia* sp.TIMO001	*Nylanderia* sp.1.*of.WW*	A
4	n45	GenBank,97	*Nylanderia* sp. BOR008	*Nylanderia cf.* sp.5.*of.SKY*	A
5	n57	GenBank,97	*Nylanderia* sp.BOR012	*Nylanderia* sp.2.of.WW	A
6	n65	GenBank,100	*Ponera* sp. MU01	*Ponera* sp.2.of.WW	A
7	n70	GenBank,100	*Camponotus* sp. CSIRO.66	*Camponotus* (*Tanaemyrmex*) sp.153.of.SKY	A
8	n71	BOLD,100	*Monomorium* sp.	*Monomorium* sp.2.of.WW	A

Only 4% mOTUs with ≥96% identities are shown.

flow, we are here able to contribute barcodes for an additional 48 species not previously available online, all of which could be identified at least to genus-level based on morphology. Of these, 28 are potentially new to science (Table S2). Amongst the remaining mOTUs (Tables 4–6), we detected three probable cases of misidentification in the sequence databases (Table 2) and provided species identities for seven database matches that previously had only genus-level identifications (Table 4). Preserved and identified barcoded vouchers were furthermore imaged and made available on a digital reference collection (https://singapore.biodiversity.online/taxon/A-Arth-Hexa-Hymn-Form). Such collections are important for taxonomic validation (Ang et al., 2013), but they can also be used to identify cases of molecular cross-contamination.

Such cross-contamination is almost inevitable in a study involving 4,000+ specimens, but it occurs on a manageable scale. In this study, 16 specimens were involved in molecular cross-contamination, a

TABLE 4 List of 4% mOTUs identified to species based on either online database matches or morphology (≥96% identity)

	mOTU code	Database, BLAST % identity	GenBank/BOLD sp. ID	Morphology-based sp. ID	Named species ID (D/M)
1	n02	BOLD,100	*Paraparatrechina illusio*	*Paraparatrechina* cf.sp.11.of.SKY	D
2	n09	GenBank,100	*Pheidole cf. sauteri*	*Pheidole parva*	M
3	n19	GenBank,99	*Camponotus* sp.CSIRO.55.1	*Camponotus carin var. tenuisquamis* (sp.48.of. SKY)	M
4	n27	GenBank,99	*Camponotus nr. reticulatus*	*Camponotus bedoti*	M
5	n44	BOLD,100	*Prionopelta* sp.	*Prionopelta kraepelini*	M
6	n56	GenBank,98	*Anochetus pubescens*	*Anochetus graeffei* sp. complex	D
7	n63	BOLD,98.71	*Pheidole clypeocornis*	*Pheidole cf. hortensis*	D
8	n67	GenBank,99	*Camponotus* sp. AEAN	*Camponotus* (*Myrmamblys*) *reticulatus sericella*	M
9	n74	GenBank,98	*Camponotus* sp ZC-2012	*Camponotus* (*Tanaemyrmex*) *albosparsus*	M
10	n77	BOLD,98.06	*Leptogenys kraepelini*	*Leptogenys* sp. (*kraepelini* group)	D
11	n82	GenBank,99	*Camponotus* YN-2011	*Camponotus* (*Tanaemyrmex*) *irritans*	M
12	n88	BOLD,100	*Brachyponera obscurans*	*Brachyponera cf. luteipes*	D
13	n90	BOLD,100	*Technomyrmex albipes*	*Technomyrmex* nr.vitiensis	D

D, database; M, morphology.
Database matches and morphology-based genus identities were in agreement for all mOTUs.

very small proportion of the 4,000+ specimens (0.004%). All cases of cross-contamination involved species from different subfamilies that could easily be distinguished even by nonexperts. Indeed, cross-contamination in a molecular laboratory tends to be random with regard to taxonomic affinity and thus rarely involves closely related species that are morphologically very similar. Many cases can also be resolved because one morphotype is much more numerous than the other.

Even without having names for all mOTUs, meaningful ecological analyses of poorly known faunas can be carried out based on NGS-barcoded specimens while experts work on species identifications and descriptions in the background. Such analyses are feasible because we here show that the pipeline allows for specimens to be reliably sorted into species-level clusters without help from taxonomic experts. This means that—in contrast to other bulk sample processing techniques that do not preserve specimen vouchers—NGS barcoding data can be used for ecological analyses that require precise data on species abundances or biomass (Gibb et al., 2017). NGS barcoding data can also be used for studying species richness and community patterns over time and space because each barcode comes with precise information on collection locality and date.

4.2 | Taxonomic rigour and accuracy are compatible with reverse workflow

While low-cost barcoding is desirable, it would be of little use if it were to yield units that have little correspondence to species. We thus rigorously assessed congruence between mOTUs and morphology. All mOTUs were morphologically validated by checking the specimens belonging to the same mOTU for morphological uniformity. In our study, we found 95% correspondence between 4% mOTUs and morphological species—this is close to the correspondence levels found in studies based on full-length barcodes by Renaud et al. (2012) (98% congruent for 1,114 barcodes, 160 morphospecies), and Zenker et al. (2016) (94% congruent for 1,075 barcodes, 286 morphospecies). It should be noted, however, that these studies only tested the molecular integrity of morphologically presorted species units. We therefore expected lower congruence in our study, but this was not observed. Our study thus expounds that accurate species sorting can be achieved for large numbers of specimens even with short barcodes (~313 bp). Some studies have demonstrated similar discriminatory powers of "minibarcodes" (e.g., Doña et al., 2015; Hajibabaei et al., 2006; Meusnier et al., 2008), but they were comparatively small scale while large scale in silico studies implied that minibarcodes may have lower power for discriminating dense species samples (Hajibabaei et al., 2006). We note, however, that our study is a limited test; results may differ when more intraspecific variability across the species' ranges is sampled (Bergsten et al., 2012).

4.3 | The case for expert verification of mOTUs

Proponents of accelerated biodiversity research may argue for further simplification of the workflow by eliminating the morphological step of our recommended pipeline. However, we would counter that expert-validated barcode databases are key for establishing meaningful links between DNA data and voucher specimen information, which is desirable for broader applications such as metabarcoding, biomonitoring and diet analyses. It is important for species information in such databases to be quality-checked in order to avoid cascading errors (Bortolus, 2008). It has been highlighted repeatedly that existing databases are not exempt from misidentified sequences, and identification errors are common. In this study, for instance, mOTU n7 was identified as *P. paracamponota* on BOLD (Table 2),

even though the corresponding image was identified by an expert as *C. rufoglaucus* (pers.comm., Sk. Yamane). Furthermore, mOTU n75 was identified as *P. beccarii* on GenBank (Table 2), but specimens from the cluster were morphologically different from actual *P. beccarii* specimens, including the species' holotype (pers.comm., Sk. Yamane). In addition, morphological validation is needed to clarify species boundaries, given that barcodes can lump recently diverged species and split older lineages of the same species that have deep genetic splits between allopatric populations (Meier et al., 2008).

4.4 | Resolution of cluster conflicts

While mOTUs produced with the reverse NGS workflow are largely in agreement with morphology, a few conflicts were flagged. In two cases, both objective clustering and ABGD yielded identical mOTUs that were incongruent with morphology. In two other cases, objective clustering and ABGD gave different results: one case of conflict involved two objective clustering mOTUs that were morphologically inferred to be one species; i.e., this may be a case of cryptic species. The same two mOTUs were identified as one cluster using ABGD. But there was also one case of conflict involving ABGD mOTUs. ABGD lumped two mOTUs that were disparate species based on morphology and objective clustering.

4.5 | Cryptic species

The first two cases of morphology–DNA conflict involve the species *Tetraponera allaborans* and *Tetramorium pacificum*. We observed morphological differences between specimens from the different mOTUs, but these differences could be interpreted as intraspecific variability based on existing taxonomic keys. Ward (2001) surmised that *T. allaborans* could possibly consist of up to five species, but proposed that they should be treated as conspecific until additional genetic evidence became available. We indeed find two barcode clusters which imply that *T. allaborans* consists of at least two species. Similarly for *T. pacificum*, pronounced morphological variation between members from different populations suggests the existence of multiple species (Schlick-Steiner, Steiner, & Zettel, 2006), even though both mOTUs keyed to *T. pacificum* based on Bolton's (1977) seminal taxonomic key.

Unlike the conflicts involving *T. allaborans* and *T. pacificum*, *Paratrechina longicornis* was a case of morphologically identical specimens being sorted into two separate mOTUs differing by 5% (average uncorrected *p*-distances). Bearing in mind that *P. longicornis* is a widespread cosmopolitan species (Fox et al., 2007; McGlynn, 1999) of disputed native range (LaPolla & Fisher, 2014; Wetterer, 2008), high genetic variability in Asia may imply multiple introductions or one introduction involving multiple haplotypes (Bergsten et al., 2012; Wild, 2009). This species appears morphologically uniform despite high underlying genetic variability in mitochondrial markers. Our result is only suggestive and more empirical evidence is needed for drawing definite conclusions, because relying excessively on genetic methods for delimiting species may yield misleading results (Wild, 2009).

Apparent high genetic variability in one morphospecies can also be due to the presence of nuclear mitochondrial pseudogenes (NUMTs), which can lead to an overestimation of true species diversity based on mitochondrial markers alone. NUMTs are known to frequently occur in insects, including ants (Aguilar-Velasco et al., 2016; Cristiano, Cardoso, & Fernandes-Salomão, 2014), although their prevalence varies between taxa (Song, Buhay, Whiting, & Crandall, 2008). Fortunately, the preservation and morphological examination of specimen vouchers by our pipeline facilitate re-investigation. In addition, NUMTs are likely to be detected by NGS-based pipelines because multiple, distinct reads are likely to be found for such specimens.

4.6 | Practical recommendations

Central towards the usefulness of our NGS pipeline is the flexibility of the barcode amplification step. PCR may be performed using either DNA extracts or raw organismal tissue (i.e., direct PCR) as template. Relatively cheap and rapid DNA extraction can be carried out using QuickExtract™ DNA extraction solution (Kranzfelder, Ekrem, & Stur, 2016), at minimal cost. Alternatively, direct PCR can be used as long as users adapt the pipeline to their own invertebrate group (see Wong et al., 2014) by optimizing reagent amounts, template tissue volume and cycling conditions. A physiological understanding of the target taxon to be barcoded is also useful for boosting dPCR success rates. For example, ant legs, which contributed most template tissue for our dPCR, contain glands that produce PCR-inhibiting compounds (Billen, 2009; Schrader, Schielke, Ellerbroek, & Johne, 2012). We therefore modified the PCR recipe and cycling conditions by adding bovine serum albumin (BSA; anti-inhibitor adjuvant) and increased the primer annealing time to 2 min. These two adjustments proved effective and drove the PCR success rates past 80% which was much higher than reported for ants in Wong et al. (2014). Possible modifications for other taxa also include lowering annealing temperature and utilization of body parts other than legs, for example antennal tips.

Another advantage of the NGS-based pipeline is its versatility and malleability to fit any form of biodiversity study involving bulk samples; application and utility of the workflow are not restricted to any particular sampling method or invertebrate taxon (Kutty et al., 2017). In cases where each sample is dominated by many specimens that are likely to belong to the same species, for example ants from the same colony, subsampling can be performed. However, we only recommend this strategy if there is no doubt that specimens which are similar in appearance definitely belong to the same species.

As each barcode is labelled with a sequence header that contains a unique specimen identifier (i.e., ZRC_BDPXXX), it is also possible to annotate collection data en masse to every barcode, enabling meaningful downstream ecological interpretation and/or environmental correlation analyses. This includes time and date information, which are pertinent in biodiversity monitoring studies that involve repeated surveys. We suggest that databasing of each specimen record be carried out; the collection information can subsequently be easily retrieved and associated with the respective unique

identifiers and their associated sequences. Downstream processing of unique sequence identifiers annotated with collection data then allows for tracking of changes in (molecular) species abundances and composition over time or other gradients.

5 | CONCLUSION

We here demonstrate the feasibility of a reverse workflow based on NGS barcoding for the rapid generation of accurate species data from a diverse invertebrate sample. The pipeline is suitable for building robust barcode databases and can be adopted across a broad range of ecological applications. This study was based on a single taxon, but the pipeline is suitable for many taxa (Kutty et al., 2017). NGS barcoding presents a promising and feasible panacea for invertebrate biologists in various disciplines who are otherwise daunted by overwhelmingly large numbers of specimens and species. We advocate the adoption of reverse workflows for processing bulk samples, given its proven technical simplicity and utility and its effective streamlining of presorting and expert morphological verification.

ACKNOWLEDGEMENTS

We would like to acknowledge support from the following grants: MOE grant for biodiversity discovery (R-154-000-A22-112) and NUS grant for SEABIG (R-154-000-648-646 and R-154-000-648-733) that supported computational resources and salary for AS. Lastly, we thank NUS OFM for granting permission to set up Malaise traps on the university campus.

AUTHOR CONTRIBUTIONS

W.W. performed the laboratory work and data analyses, and wrote most of the manuscript. A.S. designed the computational scripts and facilitated bioinformatics processes in the pipeline. M.S.F. and W.W. both organized and conducted the collection and initial processing of specimens. S.K.Y. and W.W. performed morphological validation and identification of molecular species units. R.M. edited and revised the entire manuscript.

DATA ACCESSIBILITY

The cluster-fusion dendrogram used to visualize objective clustering results, and all sequence data (barcodes) in *.fasta* format (sequence headers annotated with species IDs) are available as Supplementary Information. Sequence data have also been deposited in GenBank, accession numbers provided in Supplementary Information. Specimen images are available on the "Biodiversity of Singapore" online image reference database: https://singapore.biodiversity.online/.

ORCID

Rudolf Meier http://orcid.org/0000-0002-4452-2885

REFERENCES

Aagaard, K., Berggren, K., Hebert, P. D., Sones, J., McClenaghan, B., & Ekrem, T. (2017). Investigating suburban micromoth diversity using DNA barcoding of malaise trap samples. *Urban Ecosystems*, *20*(2), 353–361.

Aguilar-Velasco, R. G., Poteaux, C., Meza-Lázaro, R., Lachaud, J. P., Dubovikoff, D., & Zaldívar-Riverón, A. (2016). Uncovering species boundaries in the Neotropical ant complex *Ectatomma ruidum* (Ectatomminae) under the presence of nuclear mitochondrial paralogues. *Zoological Journal of the Linnean Society*, *178*(2), 226–240. https://doi.org/10.1111/zoj.12407

Ahrens, D., Fujisawa, T., Krammer, H. J., Eberle, J., Fabrizi, S., & Vogler, A. P. (2016). Rarity and incomplete sampling in DNA-based species delimitation. *Systematic Biology*, *65*(3), 478–494. https://doi.org/10.1093/sysbio/syw002

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Ang, Y., Puniamoorthy, J., Pont, A. C., Bartak, M., Blanckenhorn, W. U., Eberhard, W. G., & Meier, R. (2013). A plea for digital reference collections and other science-based digitization initiatives in taxonomy: Sepsidnet as exemplar. *Systematic Entomology*, *38*(3), 637–644. https://doi.org/10.1111/syen.12015

Ball, S. L., Hebert, P. D., Burian, S. K., & Webb, J. M. (2005). Biological identifications of mayflies (Ephemeroptera) using DNA barcodes. *Journal of the North American Benthological Society*, *24*(3), 508–524. https://doi.org/10.1899/04-142.1

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, *41* (D1), D36–D42.

Bergsten, J., Bilton, D. T., Fujisawa, T., Elliott, M., Monaghan, M. T., Balke, M., … Vogler, A. P. (2012). The effect of geographical scale of sampling on DNA barcoding. *Systematic Biology*, *61*(5), 851–869. https://doi.org/10.1093/sysbio/sys037

Billen, J. (2009). Occurrence and structural organization of the exocrine glands in the legs of ants. *Arthropod Structure & Development*, *38*(1), 2–15. https://doi.org/10.1016/j.asd.2008.08.002

Bolton, B. (1977). The ant tribe Tetramoriini (Hymenoptera: Formicidae). The genus *Tetramorium* Mayr in the Oriental and Indo-Australian regions, and in Australia. *Bulletin of the British Museum (Natural History)(Entomology)*, *36*(2), 67–151.

Bortolus, A. (2008). Error cascades in the biological sciences: The unwanted consequences of using bad taxonomy in ecology. *AMBIO: A Journal of the Human Environment*, *37*(2), 114–118. https://doi.org/10.1579/0044-7447(2008)37[114:ECITBS]2.0.CO;2

Crampton-Platt, A., Timmermans, M. J., Gimmel, M. L., Kutty, S. N., Cockerill, T. D., Vun Khen, C., & Vogler, A. P. (2015). Soup to tree: The phylogeny of beetles inferred by mitochondrial metagenomics of a Bornean rainforest sample. *Molecular Biology and Evolution*, *32*(9), 2302–2316. https://doi.org/10.1093/molbev/msv111

Cristiano, M. P., Cardoso, D. C., & Fernandes-Salomão, T. M. (2014). Could pseudogenes be widespread in ants? Evidence of *numts* in the leafcutter ant *Acromyrmex striatus* (Roger, 1863) (Formicidae: Attini). *Comptes Rendus Biologies*, *337*(2), 78–85. https://doi.org/10.1016/j.crvi.2013.11.007

Delsinne, T., Sonet, G., Nagy, Z. T., Wauters, N., Jacquemin, J., & Leponce, M. (2012). High species turnover of the ant genus Solenopsis (Hymenoptera: Formicidae) along an altitudinal gradient in the Ecuadorian Andes, indicated by a combined DNA sequencing and morphological approach. *Invertebrate Systematics*, *26*(6), 457–469. https://doi.org/10.1071/IS12030

Doña, J., Diaz-Real, J., Mironov, S., Bazaga, P., Serrano, D., & Jovani, R. (2015). DNA barcoding and minibarcoding as a powerful tool for feather mite studies. *Molecular Ecology Resources*, *15*(5), 1216–1225.

Fox, E. G. P., Solis, D. R., Jesus, C. M. D., Bueno, O. C., Yabuki, A. T., & Rossi, M. L. (2007). On the immature stages of the crazy ant *Paratrechina longicornis* (Latreille 1802)(Hymenoptera: Formicidae). *Zootaxa*, *1503*, 1–11.

Geller, J., Meyer, C., Parker, M., & Hawk, H. (2013). Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. *Molecular Ecology Resources*, *13*(5), 851–861. https://doi.org/10.1111/1755-0998.12138

Gibb, H., Dunn, R. R., Sanders, N. J., Grossman, B. F., Photakis, M., Abril, S., … Arnan, X. (2017). A global database of ant species abundances. *Ecology*, *98*(3), 883–884. https://doi.org/10.1002/ecy.1682

Gómez-Rodríguez, C., Crampton-Platt, A., Timmermans, M. J., Baselga, A., & Vogler, A. P. (2015). Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. *Methods in Ecology and Evolution*, *6*(8), 883–894. https://doi.org/10.1111/2041-210X.12376

Gompert, Z., Lucas, L. K., Buerkle, C. A., Forister, M. L., Fordyce, J. A., & Nice, C. C. (2014). Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. *Molecular Ecology*, *23*(18), 4555–4573. https://doi.org/10.1111/mec.12811

Gotelli, N. J. (2004). A taxonomic wish–list for community ecology. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *359*(1444), 585–597. https://doi.org/10.1098/rstb.2003.1443

Hajibabaei, M., Smith, M., Janzen, D. H., Rodriguez, J. J., Whitfield, J. B., & Hebert, P. D. (2006). A minimalist barcode can identify a specimen whose DNA is degraded. *Molecular Ecology Notes*, *6*(4), 959–964. https://doi.org/10.1111/j.1471-8286.2006.01470.x

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*(1), 28. https://doi.org/10.1186/1472-6785-12-28

Hebert, P. D., Cywinska, A., & Ball, S. L. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B: Biological Sciences*, *270*(1512), 313–321. https://doi.org/10.1098/rspb.2002.2218

Hebert, P. D., Ratnasingham, S., Zakharov, E. V., Telfer, A. C., Levesque-Beaudin, V., Milton, M. A., … Jannetta, P. (2016). Counting animal species with DNA barcodes: Canadian insects. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *371*(1702), 20150333. https://doi.org/10.1098/rstb.2015.0333

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., … Larsen, T. H. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, *16*(10), 1245–1257. https://doi.org/10.1111/ele.12162

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kranzfelder, P., Ekrem, T., & Stur, E. (2016). Trace DNA from insect skins: A comparison of five extraction protocols and direct PCR on chironomid pupal exuviae. *Molecular Ecology Resources*, *16*(1), 353–363. https://doi.org/10.1111/1755-0998.12446

Krell, F. T. (2004). Parataxonomy vs. taxonomy in biodiversity studies–pitfalls and applicability of 'morphospecies' sorting. *Biodiversity and Conservation*, *13*(4), 795–812. https://doi.org/10.1023/B:BIOC.0000011727.53780.63

Kutty, S. N., Wang, W., Ang, Y., Tay, Y. C., Ho, J. K. L., & Meier, R. (2017). Next-generation identification tools For Nee Soon Swamp Forest. *Gardens' Bulletin Singapore*, accepted.

Kwong, S., Srivathsan, A., & Meier, R. (2012). An update on DNA barcoding: Low species coverage and numerous unidentified sequences. *Cladistics*, *28*(6), 639–644. https://doi.org/10.1111/j.1096-0031.2012.00408.x

Kwong, S., Srivathsan, A., Vaidya, G., & Meier, R. (2012). Is the COI barcoding gene involved in speciation through intergenomic conflict? *Molecular Phylogenetics and Evolution*, *62*(3), 1009–1012. https://doi.org/10.1016/j.ympev.2011.11.034

LaPolla, J. S., & Fisher, B. L. (2014). Then there were five: A reexamination of the ant genus *Paratrechina* (Hymenoptera, Formicidae). *ZooKeys*, *422*, 35–48. https://doi.org/10.3897/zookeys.422.7779

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

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., … 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*(1), 34. https://doi.org/10.1186/1742-9994-10-34

Lim, N. K., Tay, Y. C., Srivathsan, A., Tan, J. W., Kwik, J. T., Baloğlu, B., … Yeo, D. C. (2016). Next-generation freshwater bioassessment: eDNA metabarcoding with a conserved metazoan primer reveals species-rich and reservoir-specific communities. *Royal Society Open Science*, *3*(11), 160635. https://doi.org/10.1098/rsos.160635

McGlynn, T. P. (1999). The worldwide transfer of ants: Geographical distribution and ecological invasions. *Journal of Biogeography*, *26*(3), 535–548. https://doi.org/10.1046/j.1365-2699.1999.00310.x

Meier, R. (2017). Citation of taxonomic publications: The why, when, what and what not. *Systematic Entomology*, *42*(2), 301–304. https://doi.org/10.1111/syen.12215

Meier, R., Shiyang, K., Vaidya, G., & Ng, P. K. (2006). DNA barcoding and taxonomy in Diptera: A tale of high intraspecific variability and low identification success. *Systematic Biology*, *55*(5), 715–728. https://doi.org/10.1080/10635150600969864

Meier, R., Wong, W., Srivathsan, A., & Foo, M. (2016). $1 DNA barcodes for reconstructing complex phenomes and finding rare species in specimen-rich samples. *Cladistics*, *32*(1), 100–110. https://doi.org/10.1111/cla.12115

Meier, R., Zhang, G., & Ali, F. (2008). The use of mean instead of smallest interspecific distances exaggerates the size of the "barcoding gap" and leads to misidentification. *Systematic Biology*, *57*(5), 809–813. https://doi.org/10.1080/10635150802406343

Meusnier, I., Singer, G. A., Landry, J. F., Hickey, D. A., Hebert, P. D., & Hajibabaei, M. (2008). A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics*, *9*(1), 214. https://doi.org/10.1186/1471-2164-9-214

Miller, S. E., Hausmann, A., Hallwachs, W., & Janzen, D. H. (2016). Advancing taxonomy and bioinventories with DNA barcodes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *371*(1702), 20150339. https://doi.org/10.1098/rstb.2015.0339

Morinière, J., deAraujo, 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*(5), e0155497. https://doi.org/10.1371/journal.pone.0155497

Puillandre, N., Lambert, A., Brouillet, S., & Achaz, G. (2012). ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology*, *21*(8), 1864–1877. https://doi.org/10.1111/j.1365-294x.2011.05239.x

Ratnasingham, S., & Hebert, P. D. (2007). BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes*, *7*(3), 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Renaud, A. K., Savage, J., & Adamowicz, S. J. (2012). DNA barcoding of Northern Nearctic Muscidae (Diptera) reveals high correspondence between morphological and molecular species limits. *BMC Ecology*, *12*(1), 24. https://doi.org/10.1186/1472-6785-12-24

Schlick-Steiner, B. C., Steiner, F. M., & Zettel, H. (2006). *Tetramorium pacificum* MAYR, 1870, *T. scabrum* MAYR, 1879 sp. rev., *T. manobo* (CALILUNG, 2000) (Hymenoptera: Formicidae)–three good species. *Myrmecologische Nachrichten*, *8*, 181–191.

Schrader, C., Schielke, A., Ellerbroek, L., & Johne, R. (2012). PCR inhibitors–occurrence, properties and removal. *Journal of Applied Microbiology*, *113*(5), 1014–1026. https://doi.org/10.1111/j.1365-2672.2012.05384.x

Smith, M. A., Fisher, B. L., & Hebert, P. D. (2005). DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: The ants of Madagascar. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1462), 1825–1834. https://doi.org/10.1098/rstb.2005.1714

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*, *105*(36), 13486–13491. https://doi.org/10.1073/pnas.0803076105

Srivathsan, A., Ang, A., Vogler, A. P., & Meier, R. (2016). Fecal metagenomics for the simultaneous assessment of diet, parasites, and population genetics of an understudied primate. *Frontiers in Zoology*, *13*(1), 17. https://doi.org/10.1186/s12983-016-0150-4

Srivathsan, A., & Meier, R. (2012). On the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA-barcoding literature. *Cladistics*, *28*(2), 190–194. https://doi.org/10.1111/j.1096-0031.2011.00370.x

Srivathsan, A., Sha, J., Vogler, A. P., & Meier, R. (2015). Comparing the effectiveness of metagenomics and metabarcoding for diet analysis of a leaf-feeding monkey (*Pygathrix nemaeus*). *Molecular Ecology Resources*, *15*(2), 250–261. https://doi.org/10.1111/1755-0998.12302

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, *30*(12), 2725–2729. https://doi.org/10.1093/molbev/mst197

Tänzler, R., Sagata, K., Surbakti, S., Balke, M., & Riedel, A. (2012). DNA barcoding for community ecology-how to tackle a hyperdiverse, mostly undescribed Melanesian fauna. *PLoS ONE*, *7*(1), e28832. https://doi.org/10.1371/journal.pone.0028832

van Velzen, R., Weitschek, E., Felici, G., & Bakker, F. T. (2012). DNA barcoding of recently diverged species: Relative performance of matching methods. *PLoS ONE*, *7*(1), e30490. https://doi.org/10.1371/journal.pone.0030490

Ward, P. S. (2001). Taxonomy, phylogeny and biogeography of the ant genus *Tetraponera* (Hymenoptera: Formicidae) in the Oriental and Australian regions. *Invertebrate Systematics*, *15*(5), 589–665. https://doi.org/10.1071/IT01001

Wetterer, J. K. (2008). Worldwide spread of the longhorn crazy ant, *Paratrechina longicornis* (Hymenoptera: Formicidae). *Myrmecological News*, *11*, 137–149.

Wild, A. L. (2009). Evolution of the Neotropical ant genus *Linepithema*. *Systematic Entomology*, *34*(1), 49–62. https://doi.org/10.1111/j.1365-3113.2008.00435.x

Wong, W. H., Tay, Y. C., Puniamoorthy, J., Balke, M., Cranston, P. S., & Meier, R. (2014). 'Direct PCR' optimization yields a rapid, cost-effective, nondestructive and efficient method for obtaining DNA barcodes without DNA extraction. *Molecular Ecology Resources*, *14*(6), 1271–1280. https://doi.org/10.1111/1755-0998.12275

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*(4), 613–623. https://doi.org/10.1111/j.2041-210X.2012.00198.x

Zenker, M. M., Rougerie, R., Teston, J. A., Laguerre, M., Pie, M. R., & Freitas, A. V. (2016). Fast census of moth diversity in the neotropics: A comparison of field-assigned morphospecies and DNA barcoding in tiger moths. *PLoS ONE*, *11*(2), e0148423. https://doi.org/10.1371/journal.pone.0148423

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, *30*(5), 614–620. https://doi.org/10.1093/bioinformatics/btt593

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Wang WY, Srivathsan A, Foo M, Yamane SK, Meier R. Sorting specimen-rich invertebrate samples with cost-effective NGS barcodes: Validating a reverse workflow for specimen processing. *Mol Ecol Resour*. 2018;00:1–12. https://doi.org/10.1111/1755-0998.12751