Received: 19 July 2021 | Revised: 7 February 2022 | Accepted: 26 April 2022

DOI: 10.1111/fwb.13926

Freshwater Biology WILEY

ORIGINAL ARTICLE

DNA metabarcoding unravels unknown diversity and distribution patterns of tropical freshwater invertebrates

Alexandra Zieritz[1,2] | Ping Shin Lee[2,3] | Wilhelm Wei Han Eng[4] | Shu Yong Lim[4] | Kong Wah Sing[2] | Wei Ning Chan[2] | Jey Sern Loo[2] | Farah Najwa Mahadzir[2] | Ting Hui Ng[5,6,7] | Darren C.J. Yeo[5,6] | Lydia Xinjie Gan[5] | Jing Ye Gan[2] | Christopher Gibbins[2] | Muhammad Zarul Hanifah Md Zoqratt[4] | John-James Wilson[8]

[1]School of Geography, University of Nottingham, Nottingham, UK

[2]School of Environmental and Geographical Sciences, University of Nottingham Malaysia, Semenyih, Malaysia

[3]College of Life Sciences, Anhui Normal University, Wuhu, China

[4]School of Science, Monash University Malaysia Genomics Facility, Bandar Sunway, Malaysia

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

[6]Department of Biological Sciences, Faculty of Science, National University of Singapore, Singapore City, Singapore

[7]Institute for Tropical Biology and Conservation, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia

[8]Vertebrate Zoology at World Museum, National Museums Liverpool, Liverpool, UK

Correspondence
Alexandra Zieritz, School of Geography, University of Nottingham, Sir Clive Granger Building, University Park, NG7 2RD Nottingham, UK.
Email: alexandra.zieritz@nottingham.ac.uk

Funding information
University of Nottingham

Abstract

1. Tropical freshwater invertebrate species are becoming extinct without being described, and effective conservation is hampered by a lack of taxonomic and distribution data. DNA metabarcoding is a promising tool for rapid biodiversity assessments that has never been applied to tropical freshwater invertebrates across large spatial and taxonomic scales.
2. Here we use DNA metabarcoding to comprehensively assess the benthic freshwater invertebrate fauna of the Perak River basin, Malaysia. Specific objectives were to: (1) assess performance of two DNA metabarcoding protocols; (2) identify gaps in reference databases; (3) generate new data on species diversity and distribution; and (4) draw conclusions regarding the potential value of DNA metabarcoding in tropical freshwater conservation.
3. Organisms were collected by hand and net at 34 sites and divided into small (retained in 0.5-mm but passing through 1-mm mesh) and large (retained in 1-mm mesh) fractions, and a 313-bp cytochrome c oxidase subunit I fragment amplified and sequenced using general Metazoa primers.
4. Bioinformatic analysis resulted in 468 operational taxonomic units (~species) from 12 phyla. Only 29% of species could be assigned binominal names through matches to public sequence libraries, indicating varying levels of library completeness across Orders. Extraction of small-fraction DNA with a *soil* kit resulted in a significantly higher species count than with a general kit, but this was not even across taxa.
5. Metabarcoding (amplification) success rate, estimated via comparison to morphological identifications of the large-fraction specimens, was high in most taxa analysed but low, for example, in ampullariid and viviparid gastropods. Conversely, a large proportion of species-site records for Decapoda and Bivalvia came from metabarcoding only. Species richness averaged 29 ± 16 species per site, dominated by Diptera, Annelida, and Odonata, and was particularly high in

tributaries of the mountainous Titiwangsa Range. At least eight species are new records for Malaysia, including the non-natives *Ferrissia fragilis* (Gastropoda) and *Dugesia notogaea* (Platyhelminthes).

6. Our study showed that DNA metabarcoding is generally more effective in detecting tropical freshwater invertebrate species than traditional morphological approaches, and can efficiently improve knowledge of distribution patterns and ranges of native and non-native species. However, current gaps in reference databases, particularly for bioindicator taxa, such as the Plecoptera, Ephemeroptera, and Coleoptera, need to be addressed urgently.

KEYWORDS
conservation biology, DNA barcoding, invertebrates, Southeast Asia, tropical biodiversity

1 | INTRODUCTION

Biodiversity of freshwater ecosystems is declining at a far greater rate than terrestrial or marine biodiversity (Reid et al., 2019). This is in part due to biodiversity being concentrated in fresh waters, which harbour 10% of global animal biodiversity despite occupying only 2% of the Earth's surface (Reid et al., 2019). Declines stem from anthropogenic pressure on freshwater ecosystems related to water abstraction and regulation, pollution, land-use change, overexploitation of biological resources, introduction of non-native species and climate change (Dudgeon, 2019). Rates of freshwater species loss are particularly acute in tropical biodiversity hotspots, including those in Southeast Asia, where levels of endemism and anthropogenic pressures are exceptionally high (Dudgeon et al., 2006; Mittermeier et al., 2011). Most species extinctions in tropical fresh waters are likely to be undocumented, involving species that are yet to be formally described or even discovered, or are not prevented because we lack the data to ascertain their conservation status (Dudgeon, 2003). At the same time, conservation measures to effectively protect threatened tropical freshwater species and their ecosystems, including the identification of key biodiversity areas (Holland et al., 2012), are hampered by a lack of data on their distribution, biology, and ecology.

In comparison to fish, amphibians and other vertebrates, tropical freshwater invertebrates are particularly poorly studied (Dudgeon, 2019; Liew et al., 2020). This is despite the fact that these animals support millions of livelihoods as a food source, and provide crucial ecosystem services, including water purification (Chowdhury et al., 2016; Covich et al., 2004; Macadam & Stockan, 2015). For most of these taxa, we lack even the most basic data, such as numbers and identities of species, as well as identification tools (Morse et al., 2007). New freshwater invertebrate species are described from Southeast Asia every year (Jeratthitikul et al., 2021; Mendoza & Yeo, 2014; Zieritz, Jainih, et al., 2021a), but morphological descriptive work is time-intensive and will not be completed before many of these species will have become extinct (Morse et al., 2007). In recent decades, however, molecular tools have emerged as an alternative for detecting and identifying species, particularly for morphologically variable or cryptic, poorly studied, or yet undescribed taxa. DNA barcoding allows for identification of organisms by matching a short DNA sequence from a given specimen against a reference database, such as the Barcode of Life project (BOLD, http://www.boldsystems.org; Ratnasingham & Hebert, 2013) or GenBank (https://www.ncbi.nlm.nih.gov/genbank/). Whilst taxonomic gaps in these reference databases are ubiquitous (Kvist, 2013; Porter & Hajibabaei, 2018), even in the absence of a database match, data from such unidentified or potentially undescribed species can be retained in analyses as molecular operational taxonomic units (OTUs) or barcode index numbers, leading to a more accurate and complete analysis of species diversity and distribution (Wilson et al., 2016). Since the advent of DNA metabarcoding (Taberlet et al., 2012), multiple specimens representing many different species from a single bulk sample can be processed simultaneously, rendering this a promising tool for rapidly gathering data on freshwater invertebrate diversity and distribution. Several studies have shown that DNA metabarcoding can perform equally well as or better than traditional morphology-based surveys of freshwater invertebrates (Beermann et al., 2018; Elbrecht et al., 2017; Emilson et al., 2017; Jackson et al., 2014; Kutty et al., 2018; Lim et al., 2016).

Despite its great potential for rapid biodiversity assessments in poorly studied tropical systems, DNA metabarcoding of freshwater invertebrates has predominantly been applied in well studied, temperate systems (Andújar et al., 2018; Beermann et al., 2018; Compson et al., 2019; Gardham et al., 2014; Theissinger et al., 2018). The few studies that have applied DNA metabarcoding on tropical freshwater invertebrates to date are restricted in spatial and/or taxonomic scale (e.g. focus on general metazoan diversity in two reservoirs in Singapore (Lim et al., 2016), Chironomidae in a swamp forest in Singapore (Baloğlu et al., 2018) and Southeast Asian dytiscid beetles (Balke et al., 2013)). DNA metabarcoding has not yet been used to improve our knowledge of invertebrate diversity and distribution across tropical river basins, which are known to harbour the bulk of global freshwater biodiversity (Dudgeon, 2000; Dudgeon et al., 2006), let alone identify sites of special conservation importance. One particular challenge in conducting DNA-based surveys in remote tropical regions is a potentially more rapid degradation of

DNA due to imperfect storage conditions during field campaigns, for example, because access to ice is not always available. It remains to be quantified how effectively available DNA metabarcoding protocols, including different DNA extraction methods, perform in these circumstances and systems across taxonomic groups.

The present study represents the first to apply DNA metabarcoding to comprehensively assess the benthic freshwater invertebrate fauna across a major tropical river basin. Specific objectives were to: (1) assess performance of two variations of a DNA metabarcoding protocol (with regard to DNA extraction from specimens <1 mm) and identify shortcomings, including quantification of false negatives (through comparison of morphological identification), and discuss potential mitigation measures; (2) identify the most significant gaps in reference databases for freshwater invertebrates of this region; (3) gather new information on tropical freshwater invertebrate species diversity and distribution in the study region; and (4) ultimately draw conclusions regarding the potential and limitations of DNA metabarcoding in tropical freshwater conservation.

2 | METHODS

2.1 | Study sites

The study was conducted at 34 sites in the Perak River basin in the Sundaland biodiversity hotspot (Mittermeier et al., 2011). Sites were selected haphazardly, but with the aim of including and sampling across a diversity of habitats and stream order (from small tributaries to the main stem of the river; Figure 1). The Perak River is the second longest river in Peninsular Malaysia, with a length of approximately 400 km and a basin area of 14,900 km^2 (Figure 1). It runs from the Thai border in the north through a mosaic of protected primary rainforest (Royal Belum National Park), secondary forest, urban areas, and agricultural land (including rice and oil palm plantations). The river has four consecutive, medium to large hydroelectric dams, i.e. from up- to downstream: Temenggor (reservoir area 153 km^2), Bersia (5.7 km^2), Kenering (40.5 km^2), and Chenderoh (25 km^2; Figure 1). The river exhibits very high biodiversity, including at least 107 fish species from 33 families (Hashim et al., 2012). Nevertheless, local extirpation due to damming and other anthropogenic pressures has been observed for some species, including the Critically Endangered *Probarbus jullieni* (Hashim et al., 2012). Freshwater mussel and testudine diversity is also high but declining (Sharma & Tisen, 2000; Zieritz et al., 2016).

2.2 | Field methods

Field sampling was conducted during the wet season (northeast monsoon), from November 2018 to January 2019. At each site, typically covering 50 m of river length, benthic organisms were sampled using a hand-held D-frame net (0.5-mm mesh) and by hand from all present microhabitats that were accessible by wading. Riffles and runs were sampled by kick-sampling; pools and submerged macrophytes were sampled by repeatedly jabbing the net into substrate and quickly sweeping upward to the water's surface; large rocks and boulders were sampled on the surface and underneath by hand, whilst placing the net downstream to collect dislodged specimens. Samples were washed with copious amounts of river water through a 0.5-mm wire mesh sieve and cleaned of large detritus (vegetation, stones, etc.) by hand before being preserved in absolute ethanol (in 50-ml Falcon tubes) and kept on ice whenever possible until return to the laboratory (after a maximum of 5 days), where samples were immediately stored at −20°C until processing. At each site, four to six 50-ml Falcon tubes worth of samples were taken depending on microhabitat heterogeneity. Rarely, specimens were too large to be fitted in Falcon-tubes; in these cases, we used sterilised, leakproof glass jars. Sub-samples were taken for large gastropods (e.g. *Pomacea* spp.), reducing the replicate number of each morphospecies to a minimum of three specimens from a maximum size range. All sampling equipment was sterilised by soaking in 25% bleach solution in between sampling sites.

2.3 | Laboratory methods

We loosely followed Andújar et al. (2018) to prepare bulk samples for DNA extraction. All specimens collected from a respective site were processed together. Each sample was filtered through a sieving tower of 1 mm on top of a 0.5-mm wire mesh to separate large fractions (>1 mm) from small fractions (0.5–1 mm). All four to six large fraction samples from a given site were pooled, photographed for quantification of amplification and molecular identification success rates and then immediately processed for DNA extraction (see below).

All large-fraction specimens from each site were pooled into one Falcon tube either whole (for specimens below *c*. 5 mm length, such as most dipteran larvae) or as leg or tissue snips of *c*. 4–5 mm (for specimens above *c*. 5 mm length, such as snails, prawns, caddisflies) and dried at 56°C for 3 h in the oven. DNA was extracted from pooled samples (up to 25 mg in total and <5 mg per specimen [snip]) of each site using NucleoSpin Tissue-Kit (Macherey-Nagel; 100 µL elution) following the manufacturer's protocol with overnight lysis. The remaining parts of the large specimens from which snips had been taken were stored in ethanol at −20°C for morphological reference.

For small-fraction samples, we tested the performance of two different DNA extraction kits. Approximately 50 mg was taken from each of the four to six homogenised small-fraction samples per site, dried in one Falcon tube at 56°C for 3 days and then crushed using sterilised (through autoclaving or bleach) mortar and pestle. Small-fraction DNA of 13 and 21 randomly selected sites was then extracted with either: (1) NucleoSpin Tissue-Kit (Macherey-Nagel, Germany; 100 µl elution); or (2) NucleoSpin Soil-Kit (Macherey-Nagel; 100 µl elution), respectively, following the manufacturer's instructions for higher DNA purity and yield (Table 1).

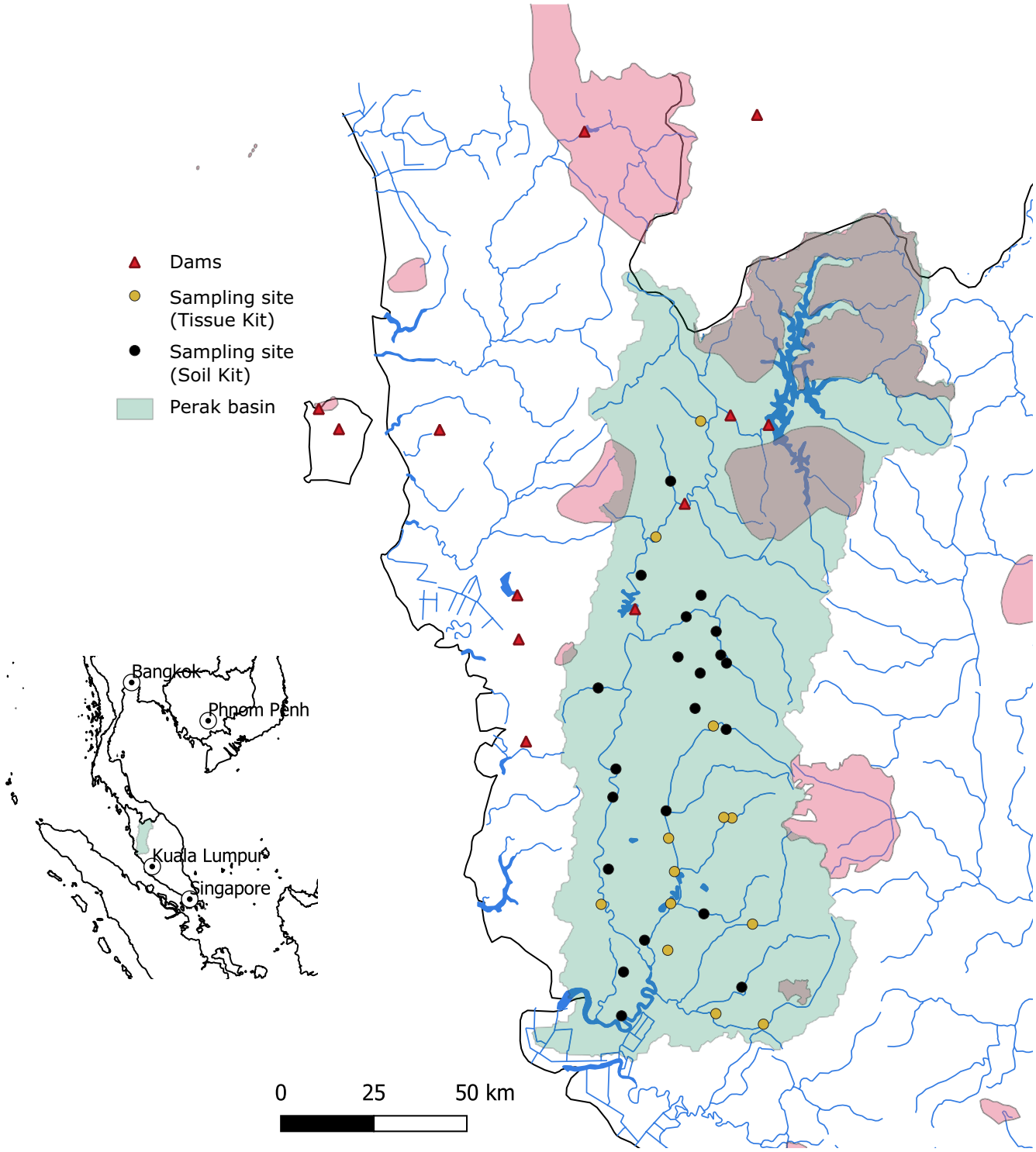

FIGURE 1 Location of the 34 sampling sites in the Perak River basin, Malaysia, for which DNA of specimens <1 mm size was extracted with NucleoSpin tissue-kit (yellow circles) and NucleoSpin soil-kit (black circles), respectively. Dam data from Mulligan et al. (2020); protected areas (red areas) from UNEP-WCMC and IUCN (2014–2020)

We amplified a 313-bp fragment (316-bp for Rotifera) of cytochrome c oxidase subunit I (COI) using primer pair mlCOIintF and jgHCO2198 (Leray et al., 2013) including overhanging adapter sequences for subsequent nested polymerase chain reaction (PCR) in analogy to the Illumina protocol—TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG and GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG onto the 5′ end of the forward and reverse primers respectively. PCR was performed in a total volume of 25 µl with 12.5 µl of Taq98™ Hot Start 2× Master Mix (Lucigen; for large-fraction samples) or 2× My Taq Red Mix (for small-fraction samples; Bioline),

TABLE 1 Numbers of reads and operational taxonomic units (OTUs) obtained in mBRAVE (see main text for details of bioinformatic analysis) after Illumina-sequencing of polymerase chain reaction products amplified with primer pair mlCOIintF and jgHCO2198 from DNA extracts of bulk benthic freshwater macroinvertebrate samples collected by net- and hand-sampling from 34 sites of the Perak River basin (coordinates provided)

LAT	LON	DNA extraction kit used for small-fraction (<1 mm) specimens	Reads	mBRAVE OTUs	Final number of species
3.958	101.288	Nucleo-Spin Tissue Kit	68,036	120	14
3.983	101.174	Nucleo-Spin Tissue Kit	175,790	196	21
4.136	101.057	Nucleo-Spin Tissue Kit	241,254	407	63
4.200	101.262	Nucleo-Spin Tissue Kit	115,424	68	16
4.224	101.145	Nucleo-Spin Tissue Kit	96,184	64	14
4.249	101.065	Nucleo-Spin Tissue Kit	104,176	96	15
4.326	101.074	Nucleo-Spin Tissue Kit	101,762	120	20
4.407	101.059	Nucleo-Spin Tissue Kit	56,542	62	17
4.456	101.213	Nucleo-Spin Tissue Kit	96,130	79	10
4.457	101.193	Nucleo-Spin Tissue Kit	106,350	69	18
4.678	101.167	Nucleo-Spin Tissue Kit	174,954	132	11
5.270	101.064	Nucleo-Spin Tissue Kit	205,912	211	15
5.415	101.137	Nucleo-Spin Tissue Kit	196,646	168	32
3.978	100.946	Nucleo-Spin Soil Kit	141,458	248	25
4.047	101.236	Nucleo-Spin Soil Kit	210,556	1,114	64
4.084	100.951	Nucleo-Spin Soil Kit	180,828	456	34
4.161	101.001	Nucleo-Spin Soil Kit	136,680	739	28
4.247	100.897	Nucleo-Spin Soil Kit	263,242	408	22
4.332	100.914	Nucleo-Spin Soil Kit	182,860	250	37
4.473	101.053	Nucleo-Spin Soil Kit	263,234	1,135	23
4.506	100.925	Nucleo-Spin Soil Kit	264,664	1,121	43
4.574	100.932	Nucleo-Spin Soil Kit	236,318	189	24
4.670	101.198	Nucleo-Spin Soil Kit	150,424	229	25
4.721	101.123	Nucleo-Spin Soil Kit	165,330	623	34
4.770	100.889	Nucleo-Spin Soil Kit	238,128	1,793	43
4.806	101.136	Nucleo-Spin Soil Kit	61,574	57	13
4.830	101.199	Nucleo-Spin Soil Kit	188,748	647	59
4.845	101.082	Nucleo-Spin Soil Kit	128,880	387	33
4.850	101.186	Nucleo-Spin Soil Kit	61,174	175	15
4.907	101.174	Nucleo-Spin Soil Kit	167,408	467	44
4.942	101.102	Nucleo-Spin Soil Kit	192,532	1,109	37
4.994	101.138	Nucleo-Spin Soil Kit	224,536	1,045	57
5.043	100.993	Nucleo-Spin Soil Kit	165,266	124	13
5.135	101.029	Nucleo-Spin Soil Kit	239,484	1,061	46

Note: Final number of species refers to freshwater invertebrate species retained after additional clustering in BOLD, automated barcode gap discovery (ABGD) analysis and discarding of non-freshwater invertebrate taxa.

0.75 µl of 10 µM of each primer, 10 µl of ddH_2O, and 1 µl of genomic DNA. Thermal cycling conditions followed Leray et al. (2013), i.e. 95°C for 3 min; 16 cycles of 95°C for 10 s, 62°C (−1°C per cycle) for 30 s, 72°C for 60 s; 25 cycles of 95°C for 10 s, 46°C for 30 s, 72°C for 60 s; 72°C for 5 min; and hold at 4°C. Negative controls were run with each PCR batch. PCR products were visualised through agarose gel electrophoresis on a 1.5% agarose gel stained with 1 × SYBR® Safe DNA gel stain (Invitrogen). If amplification was not successful (i.e. did not produce a strong band on the gel), PCR was repeated using diluted DNA extract (1:10, 1:20 or 1:30) and/or increasing DNA template to up to 3 µl. For each sample, two independent reactions were performed for both the small-fraction and large-fraction extracts. This resulted in four PCR amplicons per site, which were pooled for Illumina sequencing.

2.4 | Library preparation and Illumina sequencing

The PCR amplicons from bulk samples were cleaned using Agencourt AMPure XP beads (Beckman Coulter) using a 0.8× ratio of bead to amplification products. A second round of amplification was performed to incorporate the Illumina i5 and i7 adapters and 8-bp indexes. The reactions were performed in 10-μl volumes with 5 μl of KAPA Hifi HotStart ReadyMix (KAPA Biosystems), 1 μl of Nextera XT Index 1 primer, 1 μl of Nextera XT Index 2 primer, and 3 μl of the PCR amplicons. The thermal profile was 95°C for 3 min, then 8 cycles of 95°C for 30 s, 55°C for 30 s, 72°C for 30 s, followed by 1 cycle of 72°C for 1 min, and 4°C hold. A second round of AMPure bead clean-up was performed using a 0.7× ratio of bead to amplification products to ensure removal of unwanted small fragments, including PCR products that failed to ligate to the adapters.

The final libraries were quantified using Qubit 2.0 fluorometer (Life Technologies). The size distribution of the libraries was assessed using TapeStation 2200 (Agilent Technologies). All libraries were pooled in equimolar concentrations and loaded on an Illumina MiSeq 2 × 250 bp flow cell at 11 pM.

2.5 | Bioinformatic analysis

Read libraries were uploaded to and analysed in mBRAVE (mbrave.net; Ratnasingham, 2019) with the following parameters: trimming primers front and end (26 bp, respectively); filtering sequences with (1) mean quality value (QV) <10; (2) >4% of bp with QV < 20; (3) >1% of bp with QV < 10; merging paired ends with a minimum overlap of 20 bp; and clustering reads at an OTU threshold of 2% (to produce mBRAVE-OTUs).

Single reps of each mBRAVE-OTU were downloaded from mBRAVE, uploaded to BOLD (http://www.boldsystems.org/) under project *FIM - Freshwater Invertebrates Malaysia metabarcoding* (Zieritz, Lee, et al., 2021b), and clustered using the *Cluster Sequence* option (which employs the refined single linkage algorithm for a staged clustering process; Ratnasingham & Hebert, 2013) and default parameters to produce BOLD-OTUs. To minimise the occurrence of false positives, for each site, all BOLD-OTUs with <0.03% read frequency were then discarded (following Port et al., 2016). In a final quality-control step, the most common sequence (in terms of reads) from all retained BOLD-OTUs were aligned by ClustalW in Mega-X. BOLD-OTUs with (most common) sequences <310 bp were discarded from the dataset as were those with sequencing errors or misalignments as identified by translating DNA sequences to protein sequences.

The most common sequence of all remaining BOLD-OTUs was then blasted against both the BOLD (Full DB) and NCBI Genbank database, and assigned a name to the lowest taxonomic level possible, applying taxon-specific thresholds of genetic similarity as determined by previous literature (Appendix S1). Multiple species names were given to a BOLD-OTU if more than one name was matched to the respective sequence with the exception of names occurring in <15% of respective matches in reference databases, which were considered as mis-identifications.

BOLD-OTUs that were not identified as invertebrates or that could not be reliably assigned to at least phylum-level were discarded from the dataset. To reveal misidentifications of BOLD-OTUs due to misidentified reference sequences, a neighbour-joining tree of all taxa (maximum composite likelihood; 1,000 bootstraps) was constructed in MEGA-X (Kumar et al., 2018).

Finally, considering that the threshold of species divergence can differ considerably between taxonomic groups (see Appendix S1), we identified remaining cases where a single species was potentially represented by two or more BOLD-OTUs: Firstly, for each of the 25 taxonomic groups (listed in Table 2), pairwise distance matrices (K2P) of BOLD-OTUs were computed in Mega-X. Secondly, the same datasets were analysed through the Automated Barcode Gap Discovery (ABGD) https://bioinfo.mnhn.fr/abi/public/abgd/abgdweb.html portal using K2P-distances and default settings. BOLD-OTUs were merged (as a single species) if: (1) they were retrieved as the same OTU by ABGD; and (2) K2P between respective BOLD-OTUs were not considerably above mean and not at all above max intraspecific K2P reported for the respective taxonomic group in previous literature.

Terrestrial taxa (including Hymenoptera and Lepidoptera) commonly associated with adjacent river banks and riparian vegetation that fall into the water, as well as Chordata were excluded from further analysis. This produced a dataset of FINAL-OTUs.

We tested for the effect of small-fraction DNA extraction method on mBRAVE-OTUs and FINAL-OTUs overall and per higher taxon using Welch two-sample t-tests (for samples with unequal variances).

2.6 | Quantification of amplification and morphological identification success rates

To allow for quantification of amplification success rates (ASRs) and assess the reliability of molecular taxonomic identifications, large-fraction specimens (i.e. those retained in 1 mm-sieves) of nine orders within the phyla Mollusca (Bivalvia and Gastropoda) and Arthropoda (Malacostraca: Decapoda and Insecta: Blattodea, Coleoptera, Hemiptera, Odonata, Plecoptera and Trichoptera) were identified by morphology to the lowest taxonomic level possible based on expert knowledge and using available literature (Yule & Yong, 2004; Zieritz & Lopes-Lima, 2018). Taxa were selected for this analysis on the basis of likelihood of a large proportion of specimens and species being retained in 1 mm-sieves and our ability to identify to species- or, at least, genus-level based on our available taxonomic expertise and existing identification literature. ASRs and morphological success rates (MSRs) for large-fraction specimens were calculated for each order as well as the families within the Gastropoda, Bivalvia, Odonata, and Trichoptera, as $\mathrm{ASR} = \frac{(P_{\mathrm{both}} + P_{\mathrm{mol}})}{(P_{\mathrm{both}} + P_{\mathrm{mol}} + P_{\mathrm{mor}})}$ and

TABLE 2 List of higher taxa recovered by DNA metabarcoding from 34 sites of the Perak River basin, including their predominant size-group (meiofauna vs. macrofaunal), number of species recovered, proportion of final operational taxonomic units (OTUs) that could be matched to sequences in reference databases Genbank and BOLD at species-level, and proportion of final-OTUs that could be identified to species- and at least genus-level by comparison to reference databases, and results of two-sample *t*-test, testing whether DNA extraction of small-fractions (<1 mm specimen size) with Nucleo-spin soil kit resulted in a higher number of final-OTUs than Nucleo-spin tissue kit

Taxon	Meio/macro	Number of species (BOLD OTUs)	Match to reference sequence (%)	ID to species (%)	ID to at least genus (%)	Tissue vs. soil kit			
						t	*df*	*p*	Corr. factor
Phylum Annelida	Macro	62 (91)	34	29	47	−1.7531	31.61	**0.0446**	1.9
Phylum Arthropoda									
Subphylum Chelicerata									
Class Arachnida									
Superorder Acariformes	Meio	19 (21)	5	0	5	−1.8321	23.54	**0.0398**	5.3
Subphylum Hexapoda									
Class Entognatha									
Subclass Collembola	Meio	10 (19)	60	10	10	−1.6241	28.64	0.0577	3.4
Class Insecta									
Order Blattodea	Macro	1 (4)	100	100	100	0.8635	22.70	0.8015	0.6
Order Coleoptera	Macro	17 (17)	18	0	12	−0.7557	27.73	0.2281	1.5
Order Diptera	Macro	145 (164)	51	28	63	−1.8728	26.44	**0.0361**	2.0
Order Ephemeroptera	Macro	28 (41)	14	14	43	−1.9944	29.68	**0.0277**	1.8
Order Hemiptera	Macro	18 (29)	28	17	39	−2.002	29.26	**0.0273**	3.0
Order Odonata	Macro	42 (71)	83	83	90	−0.2694	31.65	0.3947	1.1
Order Plecoptera	Macro	12 (31)	0	0	8	−1.8276	28.43	**0.0391**	2.8
Order Trichoptera	Macro	23 (23)	39	22	52	−0.4938	31.97	0.3124	1.2
Subphylum Crustacea									
Class Branchiopoda									
Order Cladocera	Meio	6 (7)	0	0	0	−2.0342	20	**0.0277**	N/A
Class Hexanauplia									
Subclass Copepoda	Meio	8 (9)	38	38	50	−2.3591	20	**0.0143**	N/A
Class Malacostraca									
Order Decapoda	Macro	10 (84)	70	70	80	−0.3351	24.51	0.3702	1.1
Class Ostracoda	Meio	7 (10)	14	14	14	−1.7974	30.18	**0.0411**	3.4
Phylum Bryozoa	Macro	6 (8)	17	17	33	−1.1819	31.12	0.1231	2.5
Phylum Cnidaria	Macro	2 (4)	0	0	0	−0.1811	27.19	0.4288	1.2
Phylum Gastrotricha	Meio	2 (2)	0	0	0	−1.451	20	0.0812	N/A
Phylum Mollusca	Macro	21 (24)	71	71	81	1.1096	17.77	0.859	0.7

(Continues)

TABLE 2 (Continued)

Taxon	Meio/macro	Number of species (BOLD OTUs)	Match to reference sequence (%)	ID to species (%)	ID to at least genus (%)	Tissue vs. soil kit			
						t	*df*	*p*	Corr. factor
Phylum Nematoda	Meio	6 (6)	17	17	17	−2.034	20	**0.0277**	N/A
Phylum Nemertea	Meio	1 (1)	100	0	100	−1	20	0.1646	N/A
Phylum Platyhelminthes	Macro	2 (3)	50	50	50	0.3504	16.44	0.6348	0.6
Phylum Porifera	Macro	1 (1)	0	0	0	−1	20	0.1646	N/A
Phylum Rotifera	Meio	13 (14)	0	0	8	−2087	20	**0.0250**	N/A
Phylum Tardigrada	Meio	6 (6)	0	0	0	−2.3355	20	**0.0150**	N/A
Overall meiofauna		78 (95)	17	8	13	−3.6104	21.32	**0.0008**	9.0
Overall macrofauna		390 (595)	45	33	57	−2.1419	23.95	**0.0213**	1.5
OVERALL		468 (690)	40	29	49	−2.454	25.21	**0.0107**	2.6

Note: Statistically significant *p*-values (0.05 level) in bold.

MSR = $\frac{(P_{both} + P_{mor})}{(P_{both} + P_{mol} + P_{mor})}$, where P_{both} is the proportion of species-site records that were determined by both DNA metabarcoding as well as by morphological identification, P_{mol} is the proportion of taxa-site records that were determined only by DNA metabarcoding, and P_{mor} is the proportion of taxa-site records that were determined only by morphological identification. A taxon was considered to be present in both the morphological and DNA metabarcoding datasets even if it was identified above the species-level either by morphology and/or by DNA metabarcoding. Morphological identifications at species levels were further used to identify potential mis-identifications by DNA metabarcoding.

2.7 | Linking morphologically identified *Melanoides jugicostis* specimens to bulk DNA metabarcoding sequences

To link DNA metabarcoding sequences to the *Melanoides jugicostis* specimens identified by morphology, representing the first record of this species for Malaysia (see below), we additionally extracted genomic DNA from the foot tissue of one specimen from the Kinta River (4.407°N, 101.059°E; Accession Number ZRC.MOL.024119, Zoological Collection of the Lee Kong Chian Natural History Museum) using E.Z.N.A. Mollusc DNA Kit (Omega Bio-tek) following the manufacturer's instructions. PCR was run using the same primers as above in a total volume of 25 µl with 12.5 µl of exTEN 2× PCR Master Mix (Axil Scientific), 1.5 µl of 10 µM of each primer, 6 µl of ddH_2O, 2.5 µl of 1 mg/ml bovine serum albumin, and 1.5 µl of genomic DNA. Cycling parameters were as follows: 94°C for 4 min; 35 cycles of 94°C for 30 s, 48°C for 90 s, and 72°C for 90 s; and a final extension step at 72°C for 10 min. Successful amplification was verified by gel electrophoresis. The PCR product was purified and both DNA strands sequenced at Axil Scientific, Singapore. Forward and reverse sequences were assembled, and the reading frame checked using MEGA-X (Kumar et al., 2018). The final 313-bp sequence was uploaded onto BOLD (SEANM001-22) and checked against our final DNA metabarcoding dataset.

3 | RESULTS

3.1 | Operational taxonomic unit determination

Read number ranged from 56,542 to 264,664 reads per site, and number of mBRAVE-OTUs ranged from 57 to 1,793 per site (Table 1). Clustering in BOLD resulted in a total of 8,428 OTUs (BOLD-OTUs), which were further reduced to 1,382 after discarding all BOLD-OTUs with <0.03% read frequency per site. Of these, 692 BOLD-OTUs were discarded from the dataset, as they were identified as non-invertebrates (n = 415) or terrestrial invertebrates (n = 36), could not be reliably assigned to at least phylum-level (n = 221), or exhibited a sequence length of <310 bp and/or stop codons (incorrect protein translation; n = 20). This resulted in a total of 690

freshwater invertebrate BOLD-OTUs. Final ABGD/K2P analysis reduced the dataset to 468 FINAL-OTUs, hereafter referred to as *species*, from 25 higher taxonomic groups and 12 different phyla (Table 1). The proportional reduction of the number of BOLD-OTUs to FINAL-OTUs was particularly high for Malacostraca (88% reduction), Blattodea (75% reduction, resulting in only 1 FINAL-OTU) and Plecoptera (61%), whilst each BOLD-OTU was translated into a separate FINAL-OTU within the insect orders Coleoptera and Trichoptera, and the phyla Gastrotricha, Nematoda, Nemertea, Porifera, and Tardigrada (Table 1).

3.2 | Comparison of DNA extraction kits

Of the 468 freshwater invertebrate species, 390 fell into 15 higher taxonomic groups that are considered predominantly or fully macrofaunal (hereafter macrofaunal taxa), and 78 species fell into 10 higher taxonomic groups that are predominantly or fully meiofaunal (hereafter meiofaunal taxa; Table 2). Sites for which small-fraction DNA was extracted by NucleoSpin Soil-Kit exhibited significantly higher numbers of mBRAVE-OTUs (*t*-test: $t = -4.7817$, $df = 22.652$, $p < 0.0001$) and FINAL-OTUs than those extracted with NucleoSpin Tissue-Kit (Table 2).

The effect of extraction method of small-fraction DNA on FINAL-OTU number was statistically significant for seven out of 10 meiofaunal taxa, and five out of 15 macrofaunal taxa, i.e. Annelida, Diptera, Ephemeroptera, Hemiptera, and Plecoptera (Table 2, Figures 2 and 3). In total, 60% of macrofaunal and 87% of meiofaunal species were recovered only from sites for which small-fraction DNA was extracted with a Soil-Kit. For seven meiofaunal taxa and Porifera, 100% of reads were retained from sites where small-fraction DNA was extracted with a Soil-Kit (Table 2). On average, 1.5× more macrofaunal, 9× more meiofaunal, and 2.6× more total species were recovered when using NucleoSpin Soil-Kit compared to NucleoSpin Tissue-Kit for small-fraction DNA extraction (Table 2).

3.3 | Differences and reliability of molecular identification among taxa

Of the 468 FINAL-OTUs recovered, 40% could be matched to sequences in reference databases at the species-level, and 29% and 49% could be assigned binominal species names and genus-only names, respectively (Table 2, Appendix S2). Within macrofaunal taxa, a high proportion of FINAL-OTUs could be assigned species and genus names for Odonata (83% and 90%, respectively), Mollusca (71% and 81%), and Decapoda (70% and 80%); moderate levels were achieved for Diptera (28% and 63%) and Trichoptera (22% and 52%); and the proportion of successfully named species was particularly low for Plecoptera (0% and 8%), Coleoptera (0% and 12%), Hemiptera (17% and 39%), Ephemeroptera (14% and 43%), and Annelida (29% and 47%). For meiofauna, only 17% of FINAL-OTUs could be matched to reference sequences at the species-level, and only 8% and 13% could be assigned binominal species and genus-only names, respectively.

Of the 136 FINAL-OTUs that could be assigned binominal species names, 20% were assigned more than one name based on common matches with reference databases (not including rare misidentifications; Appendix S2). The proportion of such species with multiple species names was particularly high for Decapoda (43%), Hemiptera (33%) and Odonata (26%). In addition, two species names were represented by more than one FINAL-OTU, i.e. the annelid *Limnodrilus hoffmeisteri* (four FINAL-OTUs) and the gastropod *Tarebia granifera* (two FINAL-OTUs). One gastropod species, i.e. *M. jugicostis*, was initially identified to species level by morphology. The partial COI sequence generated from one of these morphologically identified specimens (BOLD ID SEANM00-22) was 100% identical to our metabarcoding sequences FIM13977-19 and FIM14326-19, thus confirming the presence of this species at the Sungai Batang Patang near Tapah and the Sungai Kinta near Batu Gajah.

3.4 | Amplification and morphological success rates

Amplification success rates of large-fraction samples averaged 0.92 across all nine taxa analysed in this respect, ranging from 0.74 in Gastropoda to 1.0 in Bivalvia, Hemiptera and Blattodea (Figure 4). At the family-level, ASR was particularly low for (1) gastropod families Ampullariidae (represented only by *Pomacea canaliculata*) and Viviparidae (represented by *Filopadulina javanica* and *Filopadulina sumatrensis*) at 0.33, respectively, (2) trichopteran family Hydropsychidae at 0.62 (represented by *Amphipsyche meridiana*, *Cheumatopsyche charites*, *Polymorphanisus astictus/ocularis*, *Potamyia flavata*, *Hydropsyche* sp., and *Macrostemum* sp.), and (3) odonate family Macromiidae (represented by *Macromia berlandi*, *Macromia callisto*, and a third, unidentified *Macromia* species) at 0.78. By contrast, MSRs averaged only 0.47 across all taxa, and was particularly low for Blattodea (0.11), Plecoptera (0.32), Bivalvia (0.33), Decapoda (0.36), and Coleoptera (0.63), indicating that the proportion of taxa-site records determined only by DNA metabarcoding but without morphologically determined specimens in large-fraction samples was particularly high for these groups.

3.5 | Freshwater invertebrate species richness of the Perak River basin

Based on the DNA metabarcoding dataset, on average, 29 ± 16 species (26 ± 14 macrofaunal and 3 ± 4 meiofaunal species) were recorded per site. In total, 58% of macrofaunal and 92% of meiofaunal species were recorded only from a single site. Within macrofaunal taxa, the proportion of single-site species was particularly high for Coleoptera (82%), Trichoptera (77%), Hemiptera (71%), and Diptera (70%), but relatively low for Decapoda (0%), Odonata (36%), Ephemeroptera (39%), Plecoptera (42%), and Mollusca (50%; Figure 2). Species richness was

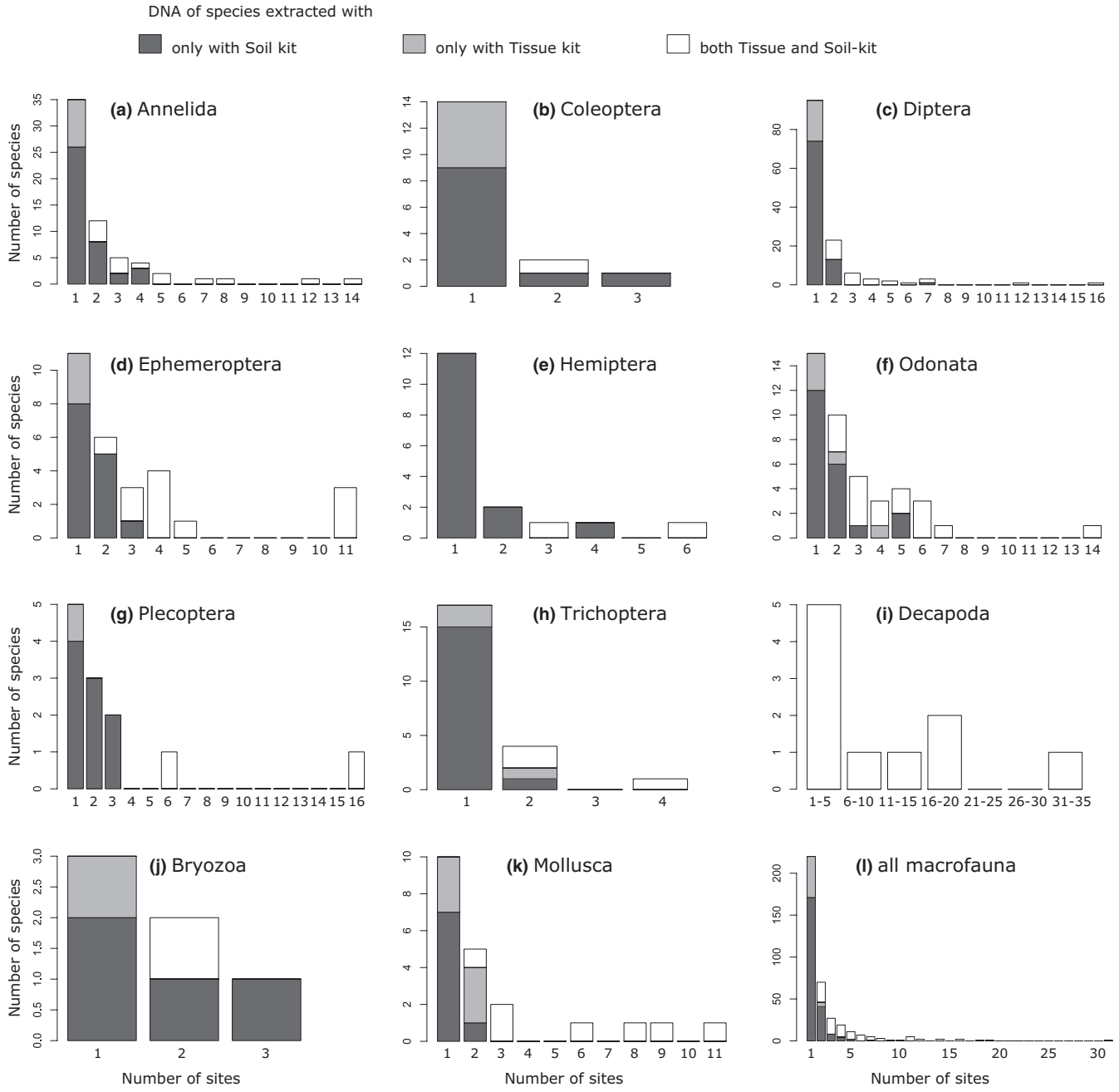

FIGURE 2 Number of species within (predominantly or exclusively) macrofaunal taxa versus number of sites where those species were recorded by DNA metabarcoding from a total of 34 sites across the Perak River basin, Malaysia

particularly high in the eastern tributaries situated in the mountainous Titiwangsa Range (Figure 5). Overall, species richness was dominated by Diptera (32% of total Final-OTUs), predominantly from the family Chironomidae (accounting for 71% of dipteran sequences matched to species-level; Appendix S2), followed by Annelida (13%) and Odonata (9%; Table 2). For the majority of species, the global IUCN status has not been assessed yet (52%), whilst 46% are listed as Least Concern and 2% as Data Deficient.

At least six species (four gastropods and two decapods) collected from the Perak River basin are considered non-native to Malaysia, whilst the non-native/native status of another 11 species could not be determined with certainty by us (Appendix S2). At least eight and potentially up to 29 species records are new for Malaysia (Appendix S2), including the non-native gastropod *Ferrissia fragilis* (native to North America; confirmed at three sites in the Perak River basin), the presumed native gastropod *M. jugicostis* (two sites), the non-native platyhelminth *Dugesia notogaea* (native to northern Australia; one site), and the presumed native odonate *Onychogomphus* cf. *risi* (six sites), the ephemeropteran *Thalerosphyrus vietnamensis*, the hemipteran *Ventidius sushmae*, and the chironomids *Polypedilum tamasemusi* and *Corynoneura yoshimurai* (present at one site each).

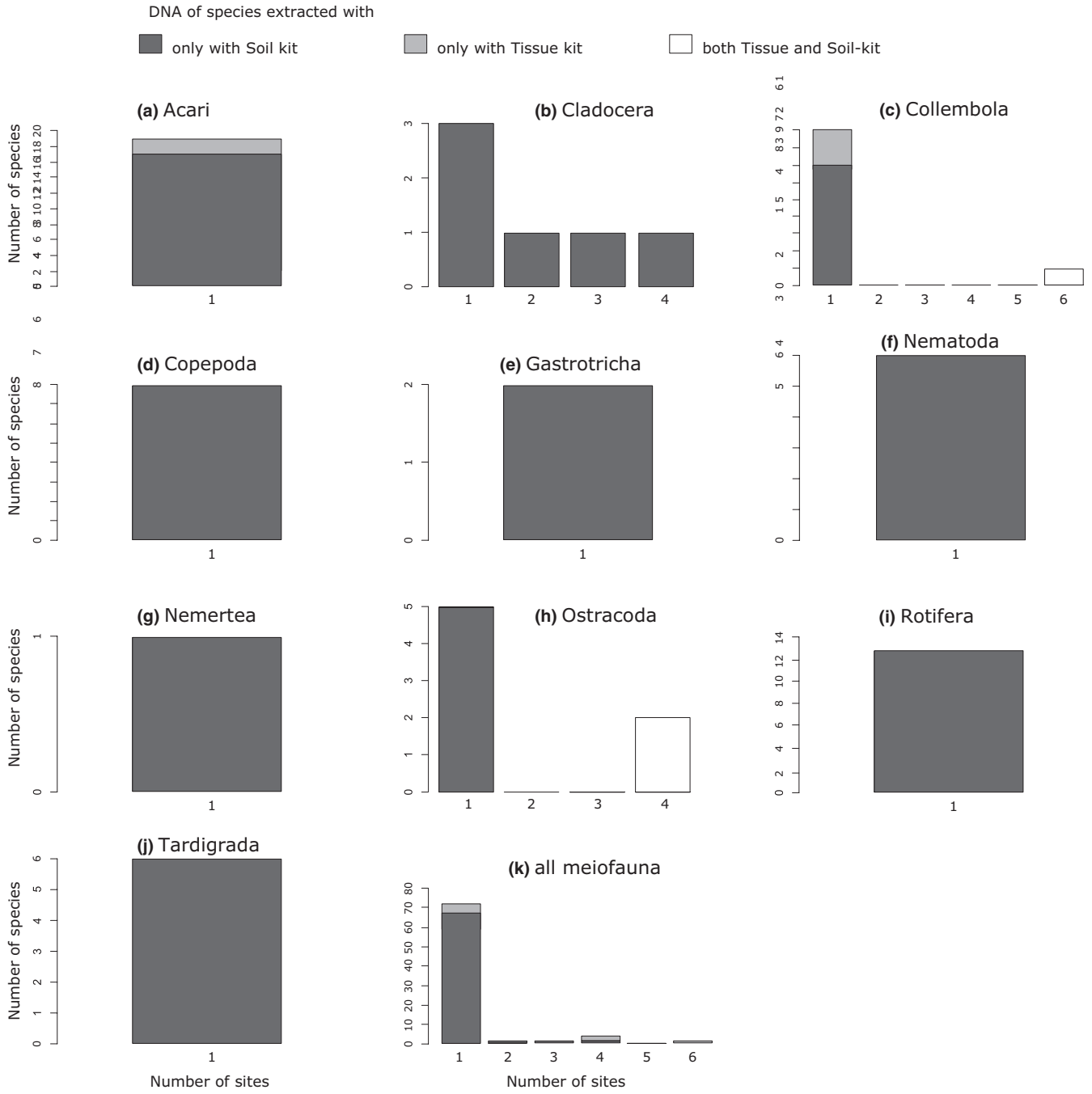

FIGURE 3 Number of species within meiofaunal taxa versus number of sites where those species were recorded by DNA metabarcoding from a total of 34 sites across the Perak River basin, Malaysia

4 | DISCUSSION

4.1 | Performance of DNA metabarcoding protocol across taxonomic groups and recommendations for future studies on tropical freshwater invertebrates

Our study confirmed previous work by, for example, Aylagas et al. (2014), Brandon-Mong et al. (2015), and Couton et al. (2019), in showing that the primer pair mlCOIintF/jgHCO2198 applied (Leray et al., 2013) successfully amplifies a COI fragment across a wide range of metazoan groups (i.e. 12 phyla in our study). However, our results also indicate that DNA extraction method can strongly affect PCR amplification success in certain taxonomic groups, and that the DNA storage and extraction protocols applied in this study may require adjustments to provide a more reliable and complete description of the diversity of particular taxonomic groups.

On average, samples for which DNA of small-fraction samples was extracted using DNA Soil-Kit retrieved 2.6 times more FINAL-OTUs than those extracted with Tissue-Kit. This effect was particularly strong and statistically significant for important bioindicator

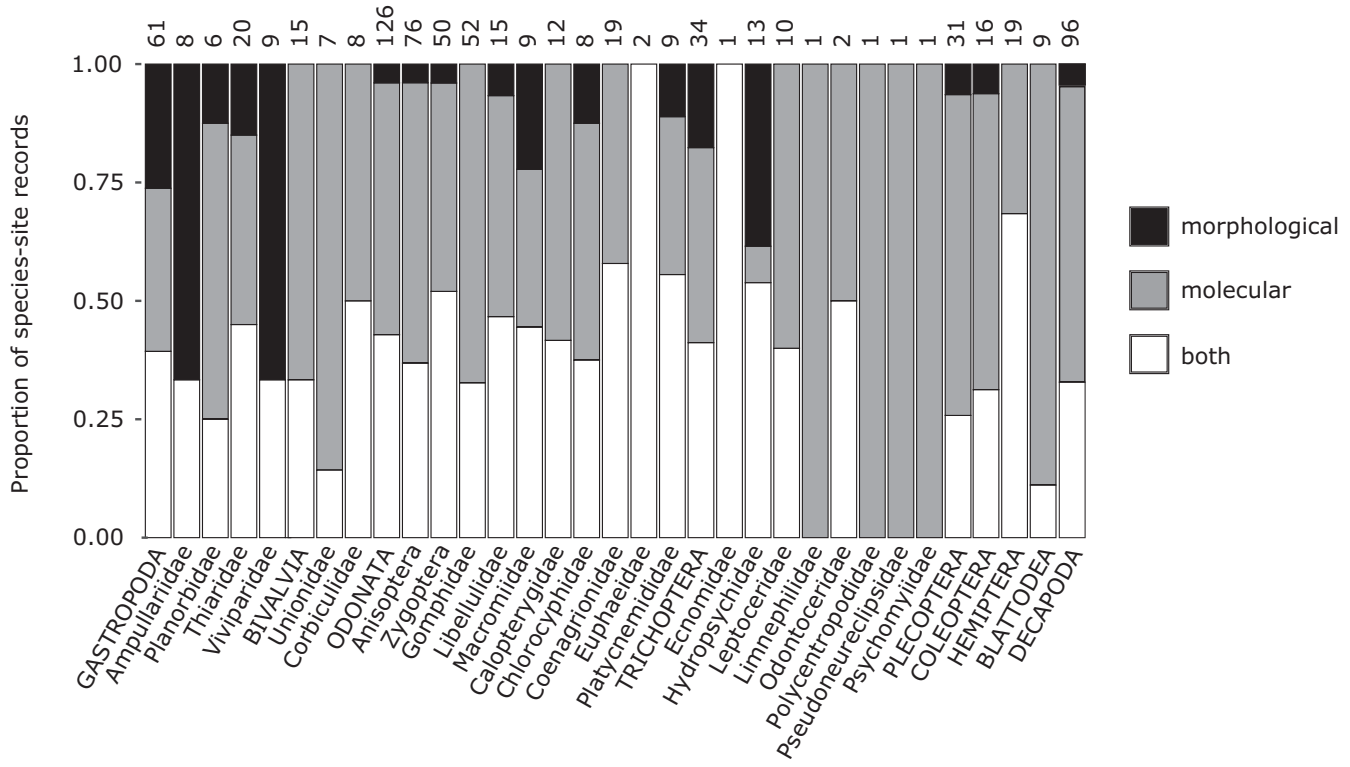

FIGURE 4 Proportion of species-site records for selected higher taxa recorded by both morphological method in large-fraction (>1 mm) samples and DNA metabarcoding, or only either of these two methods. Numbers above columns indicate total number of species-site records

taxa, such as Ephemeroptera, Diptera, Hemiptera, and Plecoptera. By contrast, Tissue-Kit appeared to work equally well as Soil-Kit for certain groups, including Mollusca, Decapoda, Odonata, and Trichoptera. These observations are in accordance with, for example, Majaneva et al. (2018), who showed that Qiagen DNeasy PowerPlant Pro Kit produced higher repeatability and more completely described the benthic community of Norwegian rivers compared to either Qiagen DNeasy Blood & Tissue Kit or a HotSHOT approach. For the majority of future studies carried out at similar habitats and conditions, DNA of small-fraction samples should thus be extracted using a DNA Soil-Kit or equivalent method rather than the cheaper DNA Tissue-Kit (or equivalent) even when the focus is on *macrofauna*. For studies focused on assessing meiofaunal diversity, we refer the reader to specialised protocols such as those by Weigand and Macher (2018), Schenk and Fontaneto (2019), and Laforest et al. (2013).

Amplification success rate across the nine orders analysed was high in our dataset (0.92), indicating that, on average, only about 8% of species-site records detected by morphological analysis of large-fraction samples were not detected by DNA metabarcoding. As such, overall ASR in this study exceeded that observed by Elbrecht et al. (2017), where 32% of morphologically identified taxa were not detected after DNA metabarcoding of Finnish stream invertebrates. By contrast, the high proportion of single-site species (58% of macrofaunal and 92% of meiofaunal taxa) in our study suggests a low degree of repeatability of our sampling protocol. Future studies, including those involving parallel processing of replicate samples (from DNA extraction to sequencing), will be needed to determine to what extent this reflects natural distribution patterns or is an artefact of inadequate field sampling (e.g. insufficient replication number) and/or DNA metabarcoding protocol (e.g. primer bias or bias due to relative abundance of taxon in respective samples; Elbrecht et al., 2017).

Amplification success rate differed considerably among taxonomic groups and was particularly low in certain snails, with DNA metabarcoding detecting the non-native *Pomacea canaliculata* (Ampullariidae) and the native *Filopadulina* spp. (Viviparidae) only in about 33% of cases. We suspect that this poor performance is probably due to a combination of factors, including high mucous production, which can lead to PCR inhibition and presence of operculum preventing proper fixation (Jaksch et al., 2016; T.H.N. personal observation). This effect can be minimised by adjusting the storage protocol of large snail samples following Jaksch et al. (2016), who recommend exchanging ethanol after collection, e.g. twice over each of 2 days before storage at −20°C (T.H.N. personal observation). Our protocol also performed relatively poorly for hydropsychid Trichoptera and macromiid Odonata, detecting only 62% and 78% of morphologically confirmed records, respectively. Further studies, comparing specific storage, DNA extraction, and PCR protocols, including primer choice, will be needed to provide detailed recommendations for improving ASRs for these taxa. By contrast, large proportions of records of other macrofaunal groups were detected

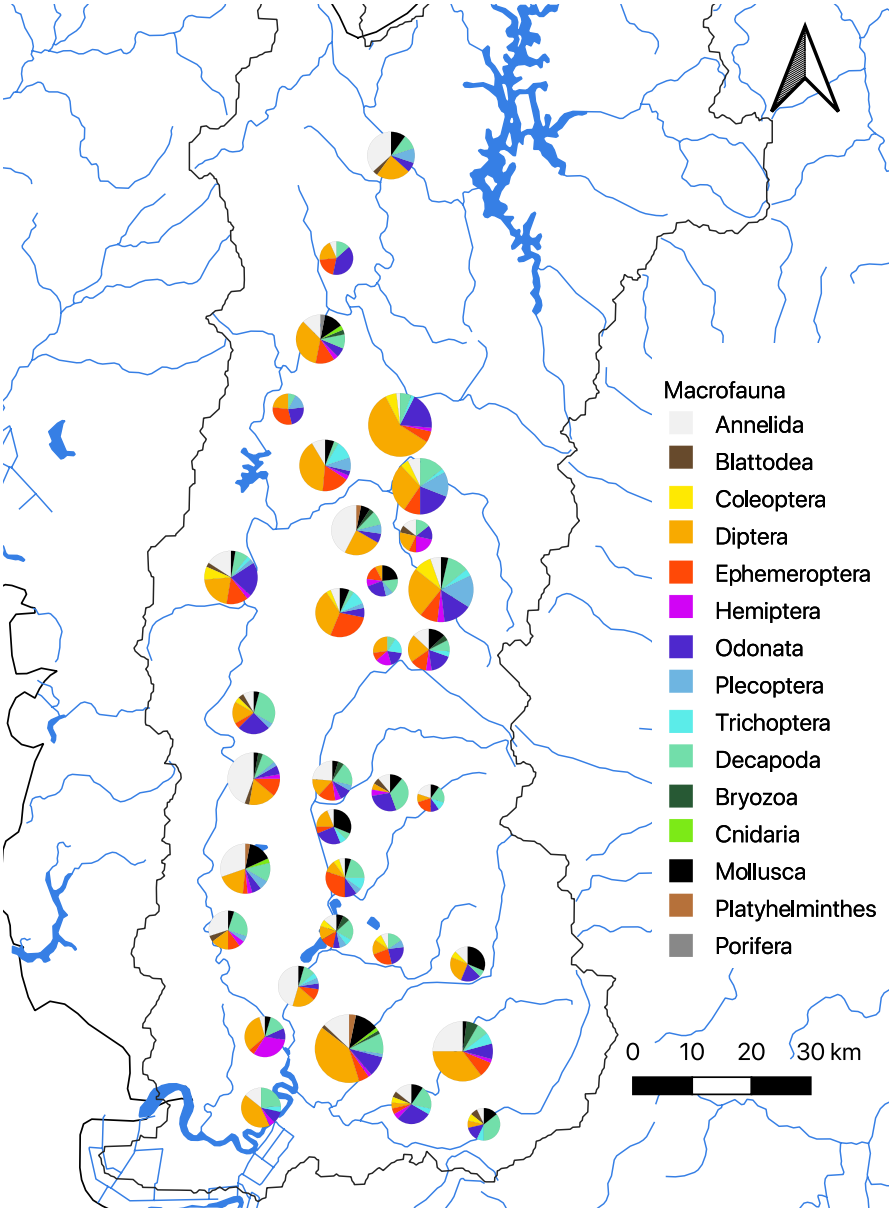

FIGURE 5 Relative species richness across macrofaunal taxa at 34 sites in the Perak River basin, Malaysia, as determined by DNA metabarcoding. Pie size scaled with total number of species, ranging from 10 to 62

by DNA metabarcoding only (i.e. despite no specimens being retained in 1 mm-sieves), including Decapoda, Bivalvia, Blattodea, Plecoptera, and Coleoptera. This suggests that the applied DNA metabarcoding protocol is particularly effective in detecting presence of these taxa at a site, potentially merely based on eggs, larval stages, gut contents, or *trace DNA* (Rossi & Mantelatto, 2013).

4.2 | Molecular identification success rate

Less than one third of the 468 FINAL-OTUs retained in our dataset could be assigned a binominal species name based on DNA metabarcoding, with strongly varying results among taxonomic groups. Large macrofaunal groups performed best, but even for well-studied groups, such as Odonata and Decapoda, a considerable proportion of FINAL-OTUs (17% and 30%, respectively) could not be assigned a binominal species name. For other common groups, such as Coleoptera and Plecoptera, not a single FINAL-OTU could be assigned a binominal species name. Moreover, large proportions of those FINAL-OTUs with an assigned species name (e.g. 26% and 43% for Odonata and Decapoda, respectively) were assigned more than one species name. In other cases, separate FINAL-OTUs were assigned the same species names. The recovery of four *Limnodrilus hoffmeisteri* FINAL-OTUs from our dataset is thereby in accordance with other studies, such as that by Vivien et al. (2017), who recovered six cryptic species within that clade from Switzerland.

In summary, our results indicate that despite recent efforts, including BOLD-initiatives such as the Trichoptera Barcode of Life Initiative/All Caddis DNA Barcoding (Zhou et al., 2016), presently available reference databases for the freshwater invertebrates of Sundaland are largely incomplete and require urgent attention. Ideally, these reference databases should be based not only on

reliable sequence data (i.e. from reliably identified specimens and, ideally, types) but also include photographs of specimens, so that morphological misidentifications can be readily spotted and thus, future misidentifications by DNA barcoding minimised.

For some taxonomic groups, such as the caridean decapods, reference database development will have to be preceded by taxonomic revisions that integrate morphological and molecular data. Shrimps of the caridean families Palaemonidae and Atyidae are common and abundant in freshwater habitats across Southeast Asia, and are an important food source, particularly for low-income, rural communities (Wowor et al., 2004). Despite their ubiquity and importance to humans, however, the taxonomy of especially *Macrobrachium* (Palaemonidae) and *Caridina* (Atyidae) remain poorly resolved, with taxonomic revisions frequently published and several species from Sundaland awaiting formal description (Siriwut et al., 2020; Siriwut et al., 2021; Wowor & Ng, 2007). In addition, morphological identification of caridean species is notoriously difficult, further increasing the chances of errors in DNA barcode reference databases, with identical or highly similar DNA sequences deposited under different taxonomic names. This has resulted in multiple species name assignment for most decapod OTUs of our dataset, with, for example, up to six available names for one of the *Caridina* OTUs. The matter appears to be further complicated by a lack of a clear barcoding gap for COI in this group, as observed by Pileggi and Mantelatto (2010) for certain South American *Macrobrachium* species. This is also suggested in our dataset by the high degree of lumping of BOLD-OTUs to FINAL-OTUs after ABGD analysis (leading to an 88% reduction of OTU number). Future work will be needed to assess the utility of COI in DNA barcoding of freshwater Caridea.

4.3 | Potential of DNA metabarcoding in filling gaps of knowledge on tropical freshwater invertebrates

The combined DNA metabarcoding-morphological approach applied in this study achieved a number of new insights into tropical freshwater invertebrate diversity, including the first Malaysian records of at least eight species. At least two of these (the gastropod *F. fragilis* and the platyhelminth *D. notogaea*) are not native to the region (GBIF, 2020; Lázaro et al., 2009; Sluys et al., 1998). With regard to the other six species, our records from the Perak River basin indicate a wider native range than previously assumed, i.e. *M. jugicostis* and *T. vietnamensis* beyond southern Thailand (Dechruksa et al., 2013; Sutthacharoenthad et al., 2019), *V. sushmae* and *O. risi* beyond India (Cheng et al., 2001; Gupta, 1981), *P. tamasemusi* beyond Japan (Kawai et al., 2014), and *C. yoshimurai* beyond Japan and China (Fu et al., 2019). These observations confirm the significant potential of DNA metabarcoding in both native as well as non-native species monitoring of tropical rivers as has been shown previously for temperate rivers (Blackman et al., 2017; Elbrecht et al., 2017) and tropical reservoirs (Lim et al., 2016).

Comparison of distribution data gathered here through DNA metabarcoding with previous data obtained by traditional sampling suggests that these two techniques are complementary. For example, traditional hand-surveys by Zieritz et al. (2016) recovered six species of unionid bivalves from the Perak river basin, whilst only two of these were detected through DNA metabarcoding. This is not surprising, as unionids commonly show a strongly aggregated distribution and can therefore be effectively surveyed only by targeted hand-sampling (Strayer & Smith, 2003). Of the two species detected by DNA metabarcoding, however, *Rectidens sumatrensis* was detected from six sites without any specimens being retained in large-fraction samples, whilst *Contradens contradens* was recorded only from one site in the main channel of the Perak River by both DNA barcoding and morphology of large-fraction samples. This pattern of distribution is in contrast to the findings of Zieritz et al. (2016), where *C. contradens*, globally and regionally assessed as Least Concern, was found to be relatively common, whilst *R. sumatrensis*, regionally assessed as Near Threatened, was recovered only from a single site. In combination, these observations suggest that the DNA metabarcoding records of *R. sumatrensis*, which were all collected in the months of December and January, probably stem from DNA material of larval and/or juvenile individuals. This in turn provides valuable data on the reproductive seasonality of this species, which was hitherto completely unknown. In addition, these records reveal the presence of *R. sumatrensis* in tributaries of the middle and lower Perak River, including Sungai Geroh, Sungai Sungkai, and Sungai Sengkoh.

AUTHOR CONTRIBUTIONS

Conceptualisation, developing methods: A.Z., P.S.L., J.J.W. Conducting the research: A.Z., P.S.L., W.E.W.H., S.Y.L., S.K.W., W.N.C., J.S.L., F.N.M., T.H.N., D.C.J.Y., L.G.X., J.Y.G., C.G., M.Z.H.Z. Data analysis: A.Z., J.J.W. Data interpretation: A.Z., J.J.W., T.H.N., D.C.J.Y. Writing: A.Z., P.S.L., W.E.W.H., S.Y.L., S.K.W., W.N.C., J.S.L., F.N.M., T.H.N., D.C.J.Y., L.G.X., J.Y.G., C.G., M.Z.H.Z., J.J.W.

ACKNOWLEDGMENTS

This study was funded by the University of Nottingham through an Anne McLaren Fellowship to A.Z.

DATA AVAILABILITY STATEMENT

DNA sequences and sampling location data are available at http://boldsystems.org under BOLD project 'FIM - Freshwater Invertebrates Malaysia metabarcoding'.

ORCID

Alexandra Zieritz https://orcid.org/0000-0002-0305-8270
Ting Hui Ng https://orcid.org/0000-0002-5123-0039

REFERENCES

Andújar, C., Arribas, P., Gray, C., Bruce, C., Woodward, G., Yu, D. W., & Vogler, A. P. (2018). Metabarcoding of freshwater invertebrates to

detect the effects of a pesticide spill. *Molecular Ecology*, *27*, 146–166. https://doi.org/10.1111/mec.14410

Aylagas, E., Borja, Á., & Rodríguez-Ezpeleta, N. (2014). Environmental status assessment using DNA metabarcoding: Towards a genetics based marine biotic index (gAMBI). *PLoS One*, *9*(3), e90529. https://doi.org/10.1371/journal.pone.0090529

Balke, M., Hendrich, L., Toussaint, E. F., Zhou, X., von Rintelen, T., & De Bruyn, M. (2013). Suggestions for a molecular biodiversity assessment of South East Asian freshwater invertebrates. Lessons from the megadiverse beetles (coleoptera). *Journal of Limnology*, *72*, 61–68.

Baloğlu, B., Clews, E., & Meier, R. (2018). NGS barcoding reveals high resistance of a hyperdiverse chironomid (Diptera) swamp fauna against invasion from adjacent freshwater reservoirs. *Frontiers in Zoology*, *15*(1), 1–12.

Beermann, A. J., Zizka, V. M., Elbrecht, V., Baranov, V., & Leese, F. (2018). DNA metabarcoding reveals the complex and hidden responses of chironomids to multiple stressors. *Environmental Sciences Europe*, *30*(1), 26.

Blackman, R. C., Constable, D., Hahn, C., Sheard, A. M., Durkota, J., Hänfling, B., & Lawson Handley, L. (2017). Detection of a new non-native freshwater species by DNA metabarcoding of environmental samples–first record of *Gammarus fossarum* in the UK. *Aquatic Invasions*, *12*(2), 177–189.

Brandon-Mong, G.-J., Gan, H.-M., Sing, K.-W., Lee, P.-S., Lim, P.-E., & Wilson, J.-J. (2015). DNA metabarcoding of insects and allies: An evaluation of primers and pipelines. *Bulletin of Entomological Research*, *105*, 717–727. https://doi.org/10.1017/S0007485315000681

Cheng, L., Yang, C. M., & Andersen, N. M. (2001). Guide to the aquatic Heteroptera of Singapore and Peninsular Malaysia. I. Gerridae and Hermatobatidae. *The Raffles Bulletin of Zoology*, *49*(1), 129–148.

Chowdhury, G. W., Zieritz, A., & Aldridge, D. C. (2016). Ecosystem engineering by mussels supports biodiversity and water clarity in a heavily polluted lake in Dhaka, Bangladesh. *Freshwater Science*, *35*, 188–199. https://doi.org/10.1086/684169

Compson, Z., Monk, W., Hayden, B., Bush, A., O'Malley, Z., Hajibabaei, M., … Baird, D. J. (2019). Network-based biomonitoring: Exploring freshwater food webs with stable isotope analysis and DNA metabarcoding. *Frontiers in Ecology and Evolution*, *7*, 395. https://doi.org/10.3389/fevo

Couton, M., Comtet, T., Le Cam, S., Corre, E., & Viard, F. (2019). Metabarcoding on planktonic larval stages: An efficient approach for detecting and investigating life cycle dynamics of benthic aliens. *Management of Biological Invasions*, *10*(4), 657–689.

Covich, A. P., Ewel, K. C., Hall, R., Giller, P., Goedkoop, W., & Merritt, D. M. (2004). Ecosystem services provided by freshwater benthos. In D. H. Wall (Ed.), *Sustaining biodiversity and ecosystem services in soils and sediments* (Vol. *64*, p. 45). Island Press.

Dechruksa, W., Krailas, D., & Glaubrecht, M. (2013). Evaluating the status and identity of "*Melania*" *jugicostis* Hanley & Theobald, 1876–an enigmatic thiarid gastropod in Thailand (Caenogastropoda, Cerithioidea). *Zoosystematics and Evolution*, *89*(2), 293–310.

Dudgeon, D. (2000). The ecology of tropical Asian rivers and streams in relation to biodiversity conservation. *Annual Review of Ecology and Systematics*, *31*, 239–263. https://doi.org/10.1146/annurev.ecolsys.31.1.239

Dudgeon, D. (2003). The contribution of scientific information to the conservation and management of freshwater biodiversity in tropical Asia. *Hydrobiologia*, *500*, 295–314.

Dudgeon, D. (2019). Multiple threats imperil freshwater biodiversity in the Anthropocene. *Current Biology*, *29*(19), R960–R967.

Dudgeon, D., Arthington, A. H., Gessner, M. O., Kawabata, Z., Knowler, D., Lévêque, C., … Sullivan, C. A. (2006). Freshwater biodiversity: Importance, status, and conservation challenges. *Biological Reviews*, *81*, 163–182.

Elbrecht, V., Vamos, E. E., Meissner, K., Aroviita, J., & Leese, F. (2017). Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. *Methods in Ecology and Evolution*, *8*, 1265–1275. https://doi.org/10.1111/2041-210X.12789

Emilson, C. E., Thompson, D. G., Venier, L. A., Porter, T. M., Swystun, T., Chartrand, D., … Hajibabaei, M. (2017). DNA metabarcoding and morphological macroinvertebrate metrics reveal the same changes in boreal watersheds across an environmental gradient. *Scientific Reports*, *7*(1), 12777. https://doi.org/10.1038/s41598-017-13157-x

Fu, Y., Fang, X., & Wang, X. (2019). *Taxonomy of Corynoneura winnertz (Diptera: Chironomidae)*. Academic Press.

Gardham, S., Hose, G. C., Stephenson, S., & Chariton, A. A. (2014). Chapter three - DNA metabarcoding meets experimental ecotoxicology: Advancing knowledge on the ecological effects of copper in freshwater ecosystems. In G. Woodward, A. J. Dumbrell, D. J. Baird, & M. Hajibabaei (Eds.), *Advances in ecological research* (Vol. *51*, pp. 79–104). Academic Press.

GBIF. (2020). GBIF – Global Biodiversity Information Facility, Data Portal. Retrieved from https://www.gbif.org

Gupta, Y. (1981). A new species of *Ventidius* distant (Hemiptera: Gerridae) from Indi. *Oriental Insects*, *15*(1), 97–102.

Hashim, Z. H., Zainuddin, R. Y., Shah, A. S. R. M., Sah, S. A. M., Mohammad, M. S., & Mansor, M. (2012). Fish checklist of Perak River, Malaysia. *Check List*, *8*(3), 6. https://doi.org/10.15560/8.3.408

Holland, R. A., Darwall, W. R. T., & Smith, K. G. (2012). Conservation priorities for freshwater biodiversity: The key biodiversity area approach refined and tested for continental Africa. *Biological Conservation*, *148*(1), 167–179. https://doi.org/10.1016/j.biocon.2012.01.016

Jackson, J. K., Battle, J. M., White, B. P., Pilgrim, E. M., Stein, E. D., Miller, P. E., & Sweeney, B. W. (2014). Cryptic biodiversity in streams: A comparison of macroinvertebrate communities based on morphological and DNA barcode identifications. *Freshwater Science*, *33*(1), 312–324.

Jaksch, K., Eschner, A., Rintelen, T. V., & Haring, E. (2016). DNA analysis of molluscs from a museum wet collection: A comparison of different extraction methods. *BMC Research Notes*, *9*, 348. https://doi.org/10.1186/s13104-016-2147-7

Jeratthitikul, E., Paphatmethin, S., Zieritz, A., Lopes-Lima, M., & Ngor, P. B. (2021). *Hyriopsis panhai*, a new species of freshwater mussel from Thailand (Bivalvia: Unionidae). *Raffles Bulletin of Zoology*, *69*, 124–136. https://doi.org/10.26107/RBZ-2021-0011

Kawai, K., Hara, S., & Saito, H. (2014). Usefulness of chironomid larvae as physicochemical and biological indicators. *Bulletin of the Hiroshima University Museum*, *6*, 7–13.

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, *35*(6), 1547–1549. https://doi.org/10.1093/molbev/msy096

Kutty, S., Wang, W., Ang, Y., Tay, Y., Ho, J., & Meier, R. (2018). Next-generation identification tools for nee soon freshwater swamp forest, Singapore. *Gardens' Bulletin Singapore*, *70*(Suppl 1), 155–173.

Kvist, S. (2013). Barcoding in the dark?: A critical view of the sufficiency of zoological DNA barcoding databases and a plea for broader integration of taxonomic knowledge. *Molecular Phylogenetics and Evolution*, *69*(1), 39–45. https://doi.org/10.1016/j.ympev.2013.05.012

Laforest, B. J., Winegardner, A. K., Zaheer, O. A., Jeffery, N. W., Boyle, E. E., & Adamowicz, S. J. (2013). Insights into biodiversity sampling strategies for freshwater microinvertebrate faunas through bioblitz campaigns and DNA barcoding. *BMC Ecology*, *13*(1), 13. https://doi.org/10.1186/1472-6785-13-13

Lázaro, E. M., Sluys, R., Pala, M., Stocchino, G. A., Baguñà, J., & Riutort, M. (2009). Molecular barcoding and phylogeography of sexual and asexual freshwater planarians of the genus *Dugesia* in the Western Mediterranean (Platyhelminthes, Tricladida, Dugesiidae).

Molecular Phylogenetics and Evolution, *52*(3), 835–845. https://doi.org/10.1016/j.ympev.2009.04.022

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*, 34. https://doi.org/10.1186/1742-9994-10-34

Liew, J. H., Lim, R. B. H., Low, B. W., Mowe, M. A. D., Ng, T. H., Zeng, Y., & Yeo, D. C. J. (2020). Tropical freshwater ecosystems, biota, and anthropogenic activities with reference to South-East Asia. In *Climate change and infectious fish diseases* (pp. 19–43). CABI.

Lim, N. K. M., Tay Ywee, C., Srivathsan, A., Tan Jonathan, W. T., Kwik Jeffrey, T. B., Baloğlu, B., … Yeo, D. C. J. (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

Macadam, C. R., & Stockan, J. A. (2015). More than just fish food: Ecosystem services provided by freshwater insects. *Ecological Entomology*, *40*(S1), 113–123. https://doi.org/10.1111/een.12245

Majaneva, M., Diserud, O. H., Eagle, S. H., Hajibabaei, M., & Ekrem, T. (2018). Choice of DNA extraction method affects DNA metabarcoding of unsorted invertebrate bulk samples. *Metabarcoding and Metagenomics*, *2*, e26664.

Mendoza, J. C., & Yeo, D. C. (2014). A new species of *Isolapotamon* Bott, 1968 (Decapoda, Brachyura, Potamidae) from Mindanao, with notes on the Philippine *Isolapotamon* species. In *Advances in freshwater decapod systematics and biology* (pp. 135–159). Brill.

Mittermeier, R. A., Turner, W. R., Larsen, F. W., Brooks, T. M., & Gascon, C. (2011). Global biodiversity conservation: The critical role of hotspots. In F. E. Zachos & J. C. Habel (Eds.), *Biodiversity hotspots* (pp. 3–22). Springer.

Morse, J. C., Bae, Y. J., Munkhjargal, G., Sangpradub, N., Tanida, K., Vshivkova, T. S., Wang, B., Yang, L., & Yule, C. M. (2007). Freshwater biomonitoring with macroinvertebrates in East Asia. *Frontiers in Ecology and the Environment*, *5*, 33–42. https://doi.org/10.1890/1540-9295(2007)5[33:FBWMIE]2.0.CO;2

Mulligan, M., van Soesbergen, A., & Sáenz, L. (2020). GOODD, a global dataset of more than 38,000 georeferenced dams. *Scientific Data*, *7*(1), 31. https://doi.org/10.1038/s41597-020-0362-5

Pileggi, L. G., & Mantelatto, F. L. (2010). Molecular phylogeny of the freshwater prawn genus *Macrobrachium* (Decapoda, Palaemonidae), with emphasis on the relationships among selected American species. *Invertebrate Systematics*, *24*(2), 194–208.

Port, J. A., O'Donnell, J. L., Romero-Maraccini, O. C., Leary, P. R., Litvin, S. Y., Nickols, K. J., … Kelly, R. P. (2016). Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. *Molecular Ecology*, *25*(2), 527–541.

Porter, T. M., & Hajibabaei, M. (2018). Over 2.5 million COI sequences in GenBank and growing. *PLoS One*, *13*(9), e0200177. https://doi.org/10.1371/journal.pone.0200177

Ratnasingham, S. (2019). mBRAVE: The multiplex barcode research and visualization environment. *Biodiversity Information Science and Standards*, *3*, e37986.

Ratnasingham, S., & Hebert, P. D. N. (2013). A DNA-based registry for all animal species: The barcode index number (BIN) system. *PLoS One*, *8*, e66213. https://doi.org/10.1371/journal.pone.0066213

Reid, A. J., Carlson, A. K., Creed, I. F., Eliason, E. J., Gell, P. A., Johnson, P. T. J., … Cooke, S. J. (2019). Emerging threats and persistent conservation challenges for freshwater biodiversity. *Biological Reviews*, *94*(3), 849–873. https://doi.org/10.1111/brv.12480

Rossi, N., & Mantelatto, F. L. (2013). Molecular analysis of the freshwater prawn *Macrobrachium olfersii* (Decapoda, Palaemonidae) supports the existence of a single species throughout its distribution. *PLoS One*, *8*(1), e54698.

Schenk, J., & Fontaneto, D. (2019). Biodiversity analyses in freshwater meiofauna through DNA sequence data. *Hydrobiologia*, *847*, 2597–2611. https://doi.org/10.1007/s10750-019-04067-2

Sharma, D., & Tisen, O. (2000). Freshwater turtle and tortoise utilization and conservation status in Malaysia. *Chelonian Research Monographs*, *2*, 120–128.

Siriwut, W., Jeratthitikul, E., Panha, S., Chanabun, R., Ngor, P. B., & Sutcharit, C. (2021). Evidence of cryptic diversity in freshwater *Macrobrachium* prawns from Indochinese riverine systems revealed by DNA barcode, species delimitation and phylogenetic approaches. *PLoS One*, *16*(6), e0252546.

Siriwut, W., Jeratthitikul, E., Panha, S., Chanabun, R., & Sutcharit, C. (2020). Molecular phylogeny and species delimitation of the freshwater prawn *Macrobrachium pilimanus* species group, with descriptions of three new species from Thailand. *PeerJ*, *8*, e10137.

Sluys, R., Kawakatsu, M., & Winsor, L. (1998). The genus *Dugesia* in Australia, with its phylogenetic analysis and historical biogeography (Platyhelminthes, Tricladida, Dugesiidae). *Zoologica Scripta*, *27*(4), 273–290.

Strayer, D. L., & Smith, D. R. (2003). A guide to sampling freshwater mussel populations. *American Fisheries Society Monograph*, *8*, 1–103.

Sutthacharoenthad, W., Sartori, M., & Boonsoong, B. (2019). Integrative taxonomy of *Thalerosphyrus* Eaton, 1881 (Ephemeroptera, Heptageniidae) in Thailand. *Journal of Natural History*, *53*(23–24), 1491–1514.

Taberlet, P., Coissac, E., Popmanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*, 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Theissinger, K., Kästel, A., Elbrecht, V., Makkonen, J., Michiels, S., Schmidt, S. I., …, Brühl, C. A. (2018). Using DNA metabarcoding for assessing chironomid diversity and community change in mosquito controlled temporary wetlands. *Metabarcoding and Metagenomics*, *2*, e21060.

UNEP-WCMC, & IUCN. (2014–2020). Protected Planet: The World Database on Protected Areas (WDPA). Retrieved from https://www.protectedplanet.net/

Vivien, R., Holzmann, M., Werner, I., Pawlowski, J., Lafont, M., & Ferrari, B. J. D. (2017). Cytochrome c oxidase barcodes for aquatic oligochaete identification: Development of a Swiss reference database. *PeerJ*, *5*, e4122. https://doi.org/10.7717/peerj.4122

Weigand, A. M., & Macher, J.-N. (2018). A DNA metabarcoding protocol for hyporheic freshwater meiofauna: Evaluating highly degenerate COI primers and replication strategy. *Metabarcoding and Metagenomics*, *2*, e26869.

Wilson, J.-J., Sing, K.-W., Lee, P.-S., & Wee, A. K. S. (2016). Application of DNA barcodes in wildlife conservation in tropical East Asia. *Conservation Biology*, *30*, 982–989. https://doi.org/10.1111/cobi.12787

Wowor, D., Cai, Y., & Ng, P. K. L. (2004). Crustacea: Decapoda, Caridea. In C. M. Yule & H.-S. Yong (Eds.), *Freshwater invertebrates of the Malaysian region* (pp. 337–357). Academy of Sciences Malaysia.

Wowor, D., & Ng, P. K. (2007). The giant freshwater prawns of the *Macrobrachium rosenbergii* species group (crustacea: Decapoda: Caridea: Palaemonidae). *The Raffles Bulletin of Zoology*, *55*(2), 321–336.

Yule, C. M., & Yong, H.-S. (2004). *Freshwater invertebrates of the Malaysian region*. Academy of Sciences Malaysia.

Zhou, X., Frandsen, P. B., Holzenthal, R. W., Beet, C. R., Bennett, K. R., Blahnik, R. J., … Kjer, K. M. (2016). The Trichoptera barcode initiative: A strategy for generating a species-level tree of life. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1702), 20160025. https://doi.org/10.1098/rstb.2016.0025

Zieritz, A., Jainih, L., Pfeiffer, J., Rahim, K. A., Prayogo, H., Anwari, M. S., … Lopes-Lima, M. (2021a). A new genus and two new, rare freshwater mussel (Bivalvia: Unionidae) species endemic to Borneo are threatened by ongoing habitat destruction. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *31*, 3169–3183.

Zieritz, A., Lee, P. S., Han, W. E. W., Lim, S. Y., Wah, S. K., Zoqratt, M. Z. H. M., & Wilson, J.-J. (2021b). FIM – Freshwater Invertebrates Malaysia metabarcoding. *BOLD – Barcode of Life Data System*. Retrieved from https://www.boldsystems.org

Zieritz, A., & Lopes-Lima, M. (2018). *Handbook and National red-List of the freshwater mussels of Malaysia*. IUCN.

Zieritz, A., Lopes-Lima, M., Bogan, A. E., Sousa, R., Walton, S., Rahim, K. A. A., … McGowan, S. (2016). Factors driving changes in freshwater mussel (Bivalvia, Unionida) diversity and distribution in Peninsular Malaysia. *Science of the Total Environment*, *571*, 1069–1078. https://doi.org/10.1016/j.scitotenv.2016.07.098

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Zieritz, A., Lee, P. S., Eng, W. W. H., Lim, S. Y., Sing, K. W., Chan, W. N., Loo, J. S., Mahadzir, F. N., Ng, T. H., Yeo, D. C., Gan, L. X., Gan, J. Y., Gibbins, C., Zoqratt, M. Z. H., & Wilson, J-J (2022). DNA metabarcoding unravels unknown diversity and distribution patterns of tropical freshwater invertebrates. *Freshwater Biology*, *00*, 1–17. https://doi.org/10.1111/fwb.13926