Received: 23 November 2020 | Revised: 18 August 2021 | Accepted: 9 September 2021

DOI: 10.1111/1755-0998.13506

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Using metatranscriptomics to estimate the diversity and composition of zooplankton communities

Mark Louie D. Lopez[1,2] | Ya-ying Lin[3] | Mitsuhide Sato[4] | Chih-hao Hsieh[5,6] | Fuh-Kwo Shiah[6] | Ryuji J. Machida[3]

[1]Biodiversity Program, Taiwan International Graduate Program, Academia Sinica and National Taiwan Normal University, Taipei, Taiwan

[2]Department of Life Science, National Taiwan Normal University, Taipei, Taiwan

[3]Biodiversity Research Center, Academia Sinica, Taipei, Taiwan

[4]Department of Environment and Fisheries Resources, Nagasaki University, Nagasaki, Japan

[5]Institute of Oceanography, National Taiwan University, Taipei, Taiwan

[6]Environmental Change Research Center, Academia Sinica, Taipei, Taiwan

Correspondence
Ryuji J. Machida, Biodiversity Research Centre, Academia Sinica, Nankang Taipei 115, Taiwan.
Email: ryujimachida@gmail.com

Funding information
Ministry of Science and Technology, Taiwan, Grant/Award Number: 108-2611-M-001 and 109-2611-M-001; Scientific Committee on Oceanic Research; Academia Sinica, Taiwan; National Taiwan University, Grant/Award Number: 109L8836

Abstract

DNA metabarcoding is a rapid, high-resolution tool used for biomonitoring complex zooplankton communities. However, diversity estimates derived with this approach can be biased by the co-detection of sequences from environmental DNA (eDNA), nuclear-encoded mitochondrial (NUMT) pseudogene contamination, and taxon-specific PCR primer affinity differences. To avoid these methodological uncertainties, we tested the use of metatranscriptomics as an alternative approach for characterizing zooplankton communities. Specifically, we compared metatranscriptomics with PCR-based methods using genomic (gDNA) and complementary DNA (cDNA) amplicons, and morphology-based data for estimating species diversity and composition for both mock communities and field-collected samples. Mock community analyses showed that the use of gDNA mitochondrial cytochrome c oxidase I (mtCO1) amplicons inflates species richness due to the co-detection of extra-organismal eDNA. Significantly more amplicon sequence variants, nucleotide diversity, and indels were observed with gDNA amplicons than with cDNA, indicating the presence of putative NUMT pseudogenes. Moreover, PCR-based methods failed to detect the most abundant species in mock communities due to priming site mismatch. Overall, metatranscriptomics provided estimates of species richness and composition that closely resembled those derived from morphological data. The use of metatranscriptomics was further tested using field-collected samples, with the results showing consistent species diversity estimates among biological and technical replicates. Additionally, temporal zooplankton species composition changes could be monitored using different mitochondrial markers. These findings demonstrate the advantages of metatranscriptomics as an effective tool for monitoring diversity in zooplankton research.

KEYWORDS
metatranscriptomics, mitochondrial transcripts, NUMT pseudogenes, PCR bias

1 | INTRODUCTION

DNA metabarcoding is a widely used molecular method for biomonitoring of zooplankton communities (Braukmann et al., 2019; Santoferrara, 2019; Yang et al., 2017). This PCR-based method amplifies a target gene region from genomic DNA (gDNA) before next-generation sequencing (NGS) (Cristescu, 2014; Elbrecht & Leese, 2015; Zhang et al., 2018). However, DNA metabarcoding has methodological uncertainties that limit its capacity to reflect the true diversity of complex zooplankton communities (van der Loos & Nijland, 2020). Methodological biases related to the co-detection of environmental DNA (eDNA) when using gDNA templates may also affect the accuracy of downstream analyses by adding more background noise to the sequence data (Piper et al., 2019). The co-amplification of mitochondrial-like templates in the nuclear genome, also called nuclear-encoded mitochondrial (NUMT) pseudogenes, when using genomic DNA templates may lead to inflated diversity estimates (Creedy et al., 2019; Shokralla et al., 2014; Song et al., 2008). NUMT pseudogenes are well documented in various metazoan taxa, especially in copepods, which have large nuclear genomes (Bensasson et al., 2003, 2011; Machida & Lin, 2017; Richly & Leister, 2004). Most importantly, PCR-amplification bias may happen when primers fail to bind effectively to the target gene sequences of specific taxa, leading to inaccurate estimates of community diversity and composition (Elbrecht & Leese, 2015; Krehenwinkel et al., 2017; Piñol et al., 2019).

Given these methodological uncertainties, a molecular approach bypassing PCR and using RNA instead of gDNA as a template is an ideal alternative method for monitoring zooplankton communities (Machida et al., 2021; Semmouri et al., 2019). RNA-based methods that capture messenger RNA (mRNA), such as metatranscriptomics, may be less prone to biases when characterizing zooplankton communities. The isolation of mRNA transcripts rather than gDNA reduces the chance of NUMT pseudogene contamination because pseudogenes are not transcribed into mature mRNA (Collura et al., 1996; Hlaing et al., 2009; Valdes & Capobianco, 2014). Moreover, metatranscriptomics does not require amplification of a target gene region using PCR, avoiding biases related to primer binding efficiency. Metatranscriptomics has been useful in advancing many aspects of plankton research but remains underused in zooplankton studies (Bucklin et al., 2018), which has hampered progress in this field compared with phytoplankton and microbial research. The doubts regarding the use of RNA-based methods may be rooted in misconceptions regarding the difficulty of sample preservation, storage, and bioinformatics procedures (Lenz et al., 2021). Consequently, the performance of metatranscriptomics at characterizing zooplankton community samples has not been rigorously validated, especially in comparison to DNA metabarcoding and morphological analysis.

Here, we assess the ability of metatranscriptomics to estimate the diversity and composition of freshwater zooplankton using both mock communities and field-collected samples. We predict that metatranscriptomics excludes NUMT pseudogene contamination and PCR-related biases (Figure 1), providing a community diversity estimate that is closer to that derived from morphology data compared with other approaches. To examine the possible presence of NUMT pseudogene contamination in DNA-based methods, we compared mtCOI amplicons from gDNA, which is expected to include amplified pseudogenes, with mtCOI amplicons from RT-PCR complementary DNA (cDNA) as a template, with which the possibility of pseudogene co-amplification is excluded. Then, the accuracy of metatranscriptomics in diversity estimation was compared with both gDNA and cDNA mtCOI amplicons in terms of species richness detection, diversity index estimation, and constructed community composition. We also evaluated the suitability of using metatranscriptomics for monitoring temporal changes in the composition of microcrustacean zooplankton communities in a subtropical reservoir, the Fei Tsui Reservoir, in Taiwan. The results of this study may help advance the use of metatranscriptomics in zooplankton monitoring.

2 | MATERIALS AND METHODS

2.1 | Sample collection

Freshwater microcrustacean zooplankton (Arthropoda: Cladocera and Copepoda) were collected from Fei Tsui Reservoir, a subtropical reservoir located in northeastern Taiwan (24°54'34.9"N, 121°34'53.0"E; altitude 160 m). A conical plankton net (50 μm mesh size) with a 45 cm-wide mouth and an attached flow meter was hauled vertically from 50 m depth to the surface. The samples were filtered with a 100 μm mesh bag to remove lake water and small taxa such as rotifers and phytoplankton. The samples were then immediately soaked in 10× the sample volume of RNAlater (Invitrogen) for 15 min to allow the remaining lake water to mix with the solution (Gorokhova, 2005). Then, each sample was transferred to a new container with the same volume of RNAlater to ensure the preservation of both RNA and DNA. The preserved samples were transported to the laboratory at room temperature within 2–3 h, stored at 4°C for 24 h, and then kept at −20°C for longer storage until DNA/RNA extraction. Individuals used to prepare mock communities were isolated from the RNAlater-preserved samples collected on 20 August 2019.

For field-collected samples, biological replicates (three different plankton hauls at the same site) and technical replicates (three independent total RNA aliquots from the same biological sample) were prepared to check the consistency of metatranscriptomics analysis (for samples collected on 24 December 2019) in characterizing field-collected samples. Samples collected from 2 July to 24 December 2019 were used to validate the capacity of metatranscriptomics to monitor temporal changes in the species composition of microcrustacean zooplankton in the field.

2.2 | Mock community preparation

Five mock communities were constructed using zooplankton samples collected from the field: a cladoceran-dominated community, a

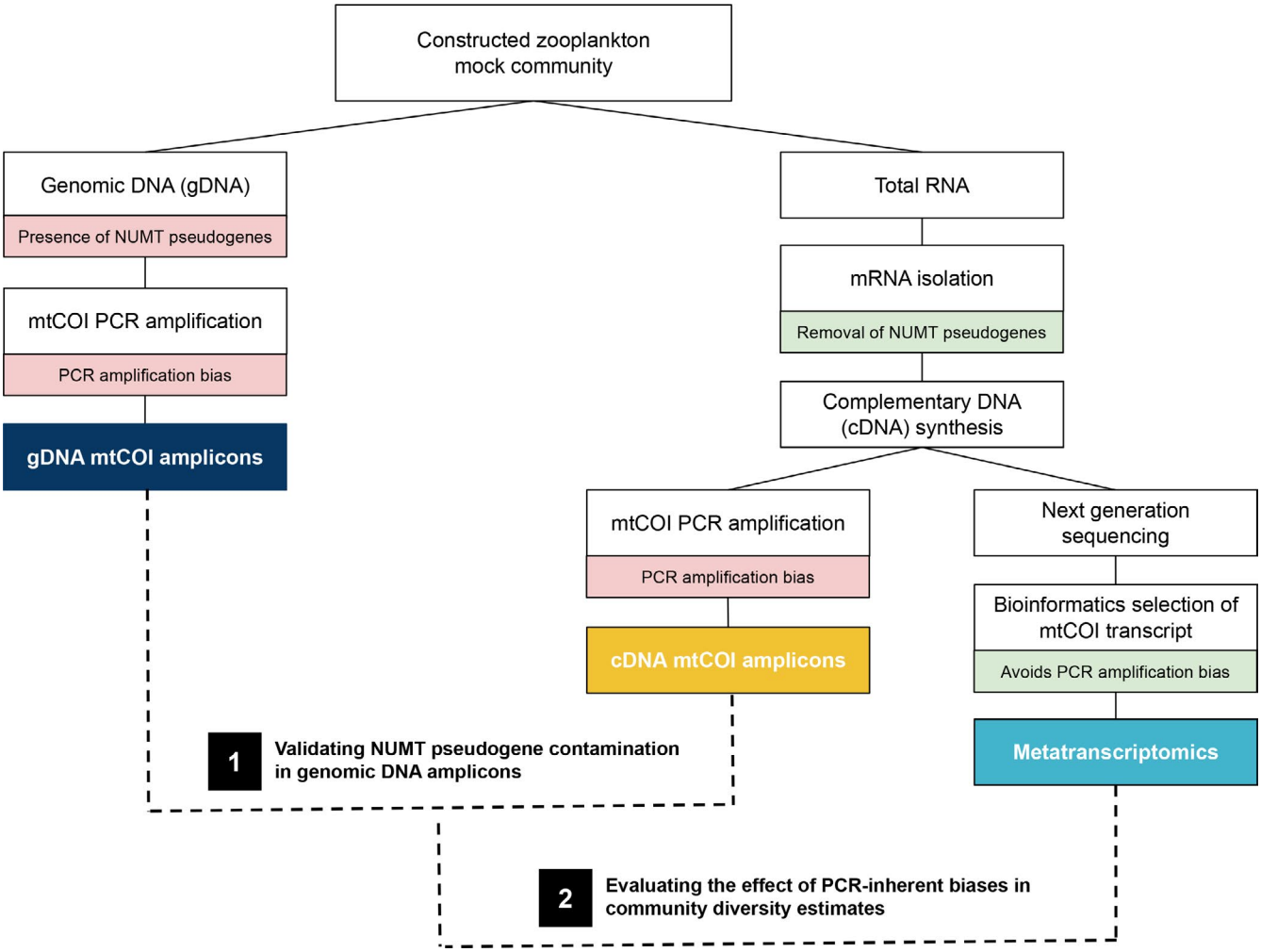

FIGURE 1 Assumptions made when evaluating the capacity of metatranscriptomics to estimate zooplankton mock community diversity

TABLE 1 The compositions of the mock communities constructed in this study

			Mock communities (number of individuals [dry weight biomass: μg])				
Taxa	Species	Individual species dry weight (μg)	Cladoceran dominated	Copepod dominated	Equal biomass	Natural assembly	With rare species
Copepoda	*Mongolodiaptomus birulai*	6.788	5 (33.94)	50 (339.43)	3 (20.37)	50 (339.43)	10 (67.89)
	Mesocyclops leuckartii	7.563	5 (37.82)	10 (75.63)	3 (22.69)	39 (22.69)	10 (75.63)
Cladocera	*Bosmina longirostris*	0.995	20 (19.90)	5 (4.97)	20 (19.90)	10 (9.95)	10 (9.95)
	Ceriodaphnia cornuta	0.726	7 (5.09)	5 (3.63)	28 (20.35)	7 (5.09)	1 (0.73)
	Daphnia galeata	7.491	30 (224.74)	5 (37.46)	3 (22.47)	20 (149.83)	10 (74.91)
	Diaphanosoma dubium	1.245	2 (2.49)	2 (92.49)	16 (19.93)	2 (2.49)	10 (12.46)
	Moina micrura	4.787	50 (239.35)	5 (23.93)	4 (19.15)	20 (95.74)	10 (47.87)

Note: Values outside the parentheses are the number of individuals per species present in each mock community, whereas the values in parentheses reflect the dry weight biomass (μg) per species calculated using a weight–length regression equation (Dumont et al., 1975).

copepod-dominated community, a community with equal species biomass, a natural assembly mimicking the actual reservoir community composition, and a community with rare species. Table 1 presents the species and dry biomass composition of each mock community.

For morphological analysis, selected individuals from each morpho-species were dissected to examine important taxonomic characters under dissecting (Nikon SMZ1500, Japan) and compound (Olympus CX41, Japan) microscopes using the techniques of

Fernando (2002). Whole mounts of the dissected specimens were prepared using a drop of glycerin as the mounting medium and sealed with transparent nail polish. Specimens were identified up to species level using annotated checklists (Lopez et al., 2017) and taxonomic keys (Dumont & Tundisi, 1984; Fernando, 1994). The body length of each specimen was measured to calculate the dry weight biomass (in μg) based on the length–weight regression equation (Dumont et al., 1975).

2.3 | DNA and RNA extractions

Figure 2 shows the workflow from nucleic acid extraction to bioinformatic analyses of the zooplankton mock communities. All sorted individuals for each mock community were placed in 1.5 ml tubes for the simultaneous extraction of RNA and DNA from the same biological sample (Triant & Whitehead, 2009). Total RNA was extracted using TriPure isolation reagent (Roche) with a PureLink RNA mini kit (Invitrogen). The quality and concentration of all extracted total RNA samples were analysed using Bioanalyser RNA 6000 Nano (Agilent Technologies) to measure the RNA integrity number (RIN), which is calculated based on the areas of 18S rRNA and 28S rRNA, with RIN =1 indicating the most degraded profile and RIN = 10 indicating the most intact profile (Schroeder et al., 2006). All samples with RIN > 7 were processed and stored at −80°C until the next procedure (Table S1).

Genomic DNA was extracted from the same mock community samples using a DNeasy kit (Qiagen) with back extraction buffer (BEB; 4 M guanidine thiocyanate, 50 mM sodium citrate, and 1 M Tris [free base]) based on the protocol of Triant and Whitehead (2009) with modifications. The extracted gDNA was purified using Agencourt AMPure XP (Beckman Coulter), following the manufacturer's protocol. The concentration and quality of the purified gDNA were measured using a NanoDrop 2000 instrument (Thermo Fisher Scientific) and a Qubit fluorometric quantitation fluorometer (Thermo Fisher Scientific). Additional details on the DNA and RNA extraction were presented in the Appendix S1.

In processing the field-collected samples for testing biological and technical replicates and zooplankton monitoring in Fei Tsui Reservoir using metatranscriptomics, the preserved samples were carefully taken out of the mesh bags and weighed using a microbalance (Denver Instrument) to determine the wet weight. Then, the weighed zooplankton samples were processed using the protocol employed to extract total RNA from the mock community samples (Tables S2 and S3). All RNA samples were stored at −80°C until the next procedure.

2.4 | PCR amplification and sequencing

The gDNA used for PCR amplification was the direct product of the DNA extraction and purification process described earlier. By contrast, cDNA was prepared using mRNA purified from total RNA with Dynabeads mRNA purification kit (Invitrogen). One hundred fifty ng of isolated mRNA was used for the reverse transcription with the standard SuperScript IV VILO Master Mix (Invitrogen) protocol.

The mtCOI from the gDNA and cDNA templates was amplified by preparing a 50 μl reaction volume containing 10 ng of template (gDNA or cDNA), 5 μl of PCR buffer, 4.0 μl of dNTP, 1.0 μl of each primer (5 μM), 1.0 μl of Advantage 2 polymerase mix (Takara Bio), and nuclease-free water. PCR was carried out using a Veriti Thermal Cycler (Applied Biosystems) with the following touchdown PCR conditions: an initial 95°C for 10 min, followed by 95°C for 10 s, 62°C for 30 s, and 72°C for 60 s. The annealing temperature was progressively reduced by 1.0°C per cycle from 62°C to 46°C during the first 16 cycles and then kept constant at 46°C for 20 additional cycles. A PCR mixture without a template was prepared as a negative control. The mtCOI primers used in this PCR were mlCOIintF (GGWACWGGWTGAACWGTWTAYCCYCC) and jgHCO2198 (TAIACYTCIGGRTGICCRAARAAYCA), which target a 313-bp fragment (Leray et al., 2013). After the PCR, the length of the amplicon band from each sample and the absence of amplicons in the negative control were confirmed using gel electrophoresis. Size selection and purification of the amplicons were performed using Agencourt AMPure XP (Beckman Coulter).

A second PCR was done to attach the barcode adapter, using different barcoding primers for each mock community reaction (Table S4). The same amount of template (10 ng) was used for each reaction, with the PCR program comprising an initial 10 min at 95°C and 20 cycles of 95°C for 10 s, 62°C for 30 s, and 72°C for 60 s. After the PCR, the amplicon size was again selected using Agencourt AMPure XP (Beckman Coulter). The DNA concentration was measured using a Qubit Fluorometric Quantitation fluorometer (Thermo Fisher Scientific). Then, 100 ng of each of the purified samples was pooled, purified with 0.9× Agencourt AMPure XP (Beckman Coulter), and eluted with 30 μl of nuclease-free water. The prepared libraries were sent for Illumina MiSeq 300 PE sequencing (1% PhiX spike-in, 10 pM loading concentration for all libraries) at the NGS High Throughput Genomics Core at the Biodiversity Research Centre, Academia Sinica, Taiwan.

2.5 | Metatranscriptome library preparation and sequencing

Metatranscriptome libraries were prepared using NEBNext mRNA Library Prep Reagent Set for Illumina (E6110) with the NEBNext Poly(A) mRNA Magnetic Isolation Module (E7490) and NEBNext Multiplex Oligos for Illumina (New England BioLabs), following the manufacturer's protocol. One thousand ng of total RNA was used for poly(A) mRNA isolation. The isolated mRNA was fragmented to *ca.* 500 bp lengths for library preparation. Final enrichment was performed for 15 PCR cycles. After purifying the enriched product using 0.9 × Agencourt AMPure XP, equal amounts of the products were pooled and sent for Illumina MiSeq 300 PE sequencing (1% PhiX spike-in, 10 pM loading concentration for all libraries) as above.

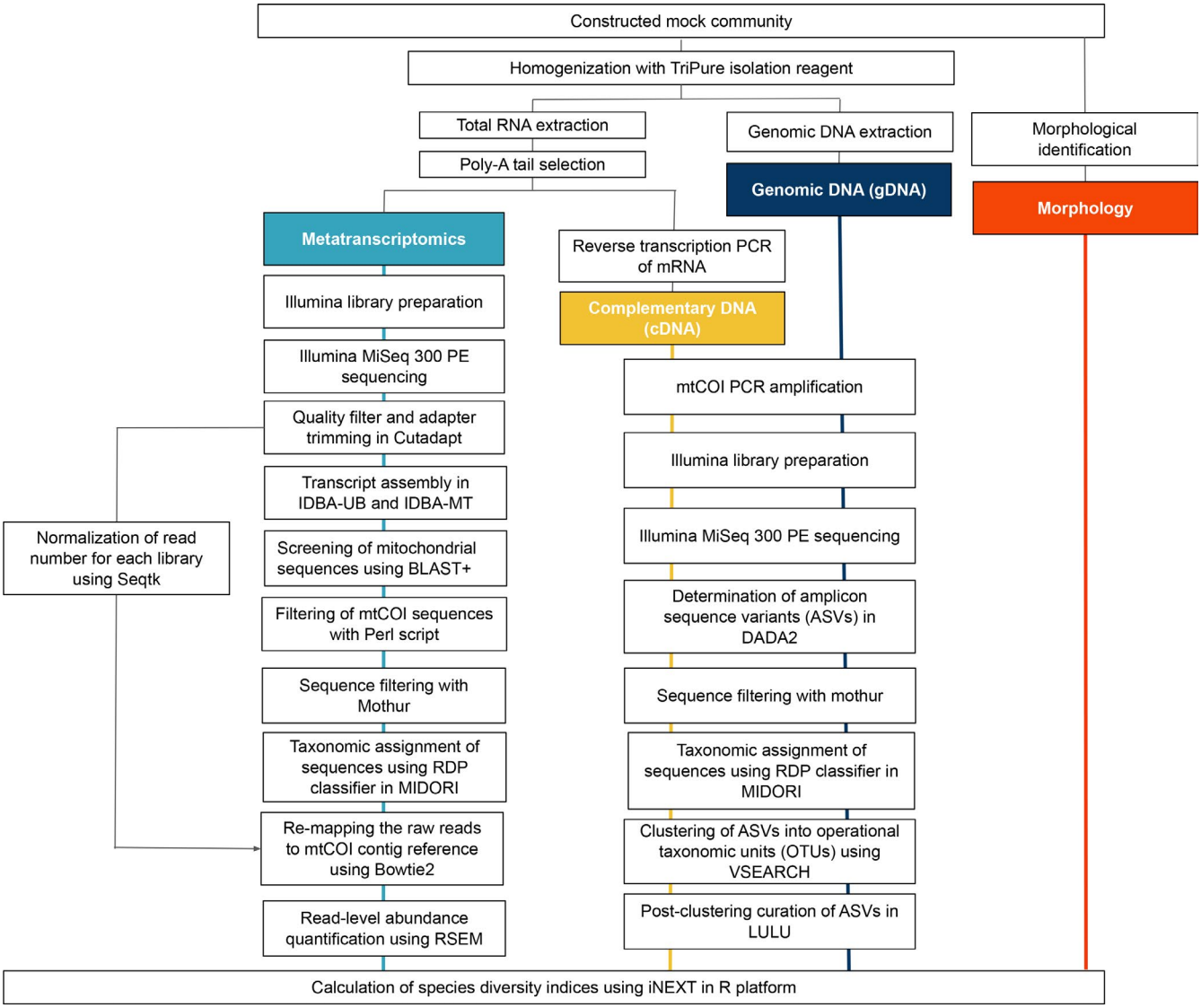

FIGURE 2 Methodology workflow for the mock community analysis

2.6 | Bioinformatic analysis of the mtCOI amplicons

From the mock community cDNA and gDNA mtCOI amplicon libraries, 8,717,291 and 8,815,608 raw reads were generated, respectively. For both gDNA and cDNA mtCOI amplicons, sequences were processed via quality filtering and adapter removal using a quality threshold of 15 in Cutadapt (version 2.10, Martin, 2011). To account for the varying sequencing depth of the community samples, the number of quality-filtered reads for each sample was normalized (equalizing the number of reads) with Seqtk (https://github.com/lh3/seqtk). The sequences were then subjected to DADA2 pipeline processing for further quality filtering, merging of paired reads, and removal of chimeras using default commands (Callahan et al., 2016), after which 70%–80% of the raw reads were retained (Table S5). Normalization of read number was done before DADA2 analysis as its pipeline does not allow extraction of sequence fasta file of the processed non-chimera amplicons. The resulting unique amplicon sequence variant (ASV) sequences per sample were extracted from the DADA2 pipeline. All arthropod-unique ASV sequences were then filtered from the fasta file using the classify.seqs and get.lineage commands in Mothur (version 1.44.3; Schloss et al., 2009), based on the mtCOI reference data set from MIDORI Longest 1.1 (Machida et al., 2017). Both the filtered and unfiltered sequences were used when comparing the methods in terms of species richness detection; however, only filtered sequences for the target taxa were used for estimating species diversity and species composition to allow a more thorough analysis of the actual mock community. Considering the difference in evolutionary rate of genes among taxonomic lineages and genetic diversity of focal species, refinement of the similarity threshold for OTU clustering is needed (Zhao et al., 2019). Here, we adopted the use of mismatch frequency distribution estimated from the gDNA and cDNA mtCOI amplicon sequences used by Machida et al. (2009). From this, we identified 94% as the optimum similarity threshold for species delineation for our mock communities, which has been also used in other complex communities (Forster et al.,

2019). With this, the ASVs were clustered into operational taxonomic units (OTUs) using the cluster_fast command in VSEARCH (version 2.15; Rognes et al., 2016). The generated VSEARCH centroid sequences were then used for taxonomic assignment of the OTUs with the RDP Classifier function (Wang et al., 2007) in the MIDORI server (Leray et al., 2018), with MIDORI Longest 1.1 (Machida et al., 2017) taken as the reference dataset with a confidence threshold of 80% at the species level as a significance cut-off. Post-clustering of the OTUs was also done using the default LULU command to remove erroneous molecular OTUs (ver. 1.2.3; Frøslev et al., 2017). After sequence curation, the OTU tables were used to calculate species diversity indices with the iNEXT package (Hsieh et al., 2020) within the R platform (R Core Team, 2017). The iNEXT calculates rarefied and extrapolated diversity indices up to double the reference sample size, which accounts for the potential bias caused by the varying number of non-chimeral sequences per community sample after DADA2 processing (Table S5).

2.7 | Bioinformatic analysis of metatranscriptome sequences

The metatranscriptome sequences of the mock communities yielded 23,382,940 reads that were demultiplexed into five different mock community libraries (Table S1). The raw sequences were subjected to quality filtering and adapter removal with the same criteria used for the amplicon sequences. The cleaned paired-end reads were used for transcript assembly in IDBA-UD (ver. 1.1.3; Peng et al., 2012) with long read (–l) parameters. The transcript assembly was then subjected to IDBA-MT analysis (Leung et al., 2013) to remove chimeric contigs. The IDBA-UD and IDBA-MT assembly workflow were specifically designed for metatranscriptomic data. From the assembled transcript, contigs with high similarity to mtCOI were screened by querying the assembly fasta file against the indexed local BLAST database (Camacho et al., 2009) containing mitochondrial reference sequences (13 proteins and 2 rRNAs) from the MIDORI Longest 1.1 datasets (Machida et al., 2017). All mtCOI transcripts were then extracted using a constructed Perl script (https://bit.ly/3lDPSfd) based on the BLAST result. Table S1 gives the assembly statistics and BUSCO completeness values (v5.1.2, Seppey et al., 2019) for the community transcripts and the quality of the extracted mtCOI markers. The mtCOI transcripts were filtered using the classify.seqs and get.lineage functions of Mothur (version 1.44.3; Schloss et al., 2009) based on the COI reference dataset from MIDORI Longest 1.1 (Machida et al., 2017) to get the sequences of the target species in the mock communities. As in the amplicon processing, only the extracted sequences assigned to the target species present in the mock communities were used for species diversity estimation and community composition construction. The mtCOI transcript reference was then indexed using the bowtie2-build command (Langmead & Salzberg, 2012) to serve as the reference in mapping back the normalized paired-end reads of each community (using Seqtk to subsample equal number of raw reads; Table S1). Then, the read-level abundance was quantified in transcripts per million (TPM) with RSEM (version 1.3.3; Li & Dewey, 2011). The read-level species diversity indices were calculated with the iNEXT package (Hsieh et al., 2020) within the R platform (R Core Team, 2017) using the output file from RSEM.

For the field-collected samples, 8,468,448 raw reads were generated for the replication test (Table S2) and 15,728,180 for temporal monitoring of the zooplankton community (Table S3). The same bioinformatics workflow was used for processing field community samples for both metatranscriptomics replication testing and monitoring of temporal changes in the community composition of microcrustacean zooplankton in the reservoir: quality filtering, transcript assembly, extraction of selected mitochondrial genes for reference construction, mapping back raw reads to the assembled reference, read-level abundance quantification, and calculation of species diversity indices. The numbers of processed reads and assembly quality scores for the biological and technical replicates are in Table S2. Table S3 contains supplemental information on the use of different mitochondrial transcripts (16S, COI, and CytB) from metatranscriptomics for monitoring temporal changes in zooplankton composition from July to December of 2019. To address the lack of 16S and CytB reference sequences for the target species in GenBank, the taxonomy was assigned using a modified MIDORI Longest 1.1 reference dataset (Machida et al., 2017) with added sequences from unpublished transcriptomic sequences of *Mongolodiaptomus birulai*, *Mesocyclops leuckarti*, *Bosmina longirostris*, *Ceriodaphnia cornuta*, and *Moina micrura*.

2.8 | NUMT pseudogene analyses

Most NUMT pseudogenes can be detected using MACSE (Ranwez et al., 2011), which detects pseudogenes by screening for frameshift and stop codons caused by the presence of indels (Leray & Knowlton, 2015). However, some NUMT pseudogene variants do not possess these obvious features, making it difficult to distinguish them from a normal mitochondrial copy in DNA barcoding amplicons (Andújar et al., 2021; Creedy et al., 2019; Shokralla et al., 2014). Therefore, pseudogenes were detected in the DNA-based method by comparing the gDNA and cDNA sequences based on calculating nucleotide diversity, synonymous (π [S]) and nonsynonymous (π [N]) substitutions, and indels using DnaSP (Rozas et al., 2017). The presence of putative NUMT pseudogenes tends to increase sequence variability (ASVs, nucleotide diversity, and indels), with synonymous substitutions predominating (Machida & Lin, 2017; Perna & Kocher, 1996; Zischler et al., 1995).

2.9 | Statistical analysis

The similarity in the species richness detected by each method was assessed via a Venn diagram constructed using the VennDiagram package (Chen, 2018). The independent *t*-test was used to determine

statistical differences in the parameters used for detecting pseudogenes in the vegan package (Oksanen et al., 2019). The statistical differences between the species diversity indices (Shannon and Simpson's indices) calculated with each method were tested using analysis of variance (ANOVA) with the ggpubr package (Kassambara, 2020). Both non-metric multidimensional scaling (NMDS) clustering and permutational multivariate analysis of variance (PERMANOVA) were performed to assess similarities in the species composition derived from each method using the vegan package (Oksanen et al., 2019). All these statistical analyses were done using the R platform (R Core Team, 2017).

3 | RESULTS

3.1 | Species richness detection

A total of 13 mtCOI transcripts from metatranscriptomics (Figure 3a) were assigned to the seven microcrustaceans species present in the mock communities, small freshwater rotifers, and *Alexandrium* species (Table S6). With PCR-based methods, 44 and 21 OTUs were detected from gDNA and cDNA, respectively. The cDNA amplicons represented a subset of gDNA amplicons, with 20 shared and one exclusive (Table S8, Rotifera: *Conochilus unicornis* with 33 sequence reads in one mock community) OTUs. The analysis with gDNA amplicons produced greater number of recovered OTUs, including 23 OTUs not observed with the RNA-based methods. These gDNA-exclusive OTUs can be due to the co-detection of extra-organismal eDNA together with sequences from epiphytes and possible zooplankton gut contents. At the same time, erroneous sequences from the gDNA amplicon-based analysis caused unusual taxonomic assignments including species that are not expected to be present in the reservoir's limnetic area, such as a marine hydrozoan (*Aglaophenia* sp.) and a spider (Figure S1 and Table S7). The presence of erroneous sequences can be attributed to the co-detection of eDNA or the presence of chimeric artifacts that remain after denoising with DADA2 (Prodan et al., 2020).

To compare the species richness estimates of mock communities across methods, the target species' sequences were filtered (Figure 3b) and used in the subsequent analyses. Note that only the non-PCR-based method, metatranscriptomics, detected all species present in the actual mock communities. Both methods based on cDNA and gDNA mtCOI amplicons failed to detect *Mongolodiaptomus birulai* (the most abundant species in Fei Tsui Reservoir) in all mock communities. The universal primers do not match the DNA template for this species perfectly, with four primer mismatches with the species' priming site sequences (Figure S2).

3.2 | NUMT pseudogene detection

To examine NUMT pseudogene contamination, the sequences of six species in the mock communities detected using both gDNA and cDNA amplicons were compared (Table 2). The sequence variation was significantly greater for the gDNA amplicon-based method than for the cDNA-based method in all six species. The gDNA-based method produced 1.3–11.2 times more ASVs than did the cDNA-based method among examined species ($p < .05$). There was also a significant ($p < .05$) difference in nucleotide diversity (π) due to relatively greater variation in the mtCOI sequences derived from gDNA (increase of 0.002–0.025). No indels were seen in the cDNA sequences, whereas 24, five, and six indels were observed in the gDNA of *Mesocyclops leuckarti*, *Daphnia galeata*, and *Moina micrura*, respectively. Lastly, a predominance of synonymous substitutions (π [S]) was observed in *Mesocyclops leuckarti* gDNA mtCOI sequences, which had 4.5 times more synonymous substitutions than observed nonsynonymous substitutions.

3.3 | Diversity indices and community composition estimates

The Shannon and Simpson's diversity indices derived from the three molecular-based methods (Figure 3c) did not differ significantly ($p > .05$) from those derived from morphological data, despite the absence of one species for which neither cDNA nor gDNA was amplified. In terms of inferred community composition based on read-level abundance, both cDNA and gDNA amplicon-based methods produced highly similar species compositions for all mock communities. In comparison, the species composition inferred from metatranscriptomics was highly similar to that derived from the morphological data, as shown in Figure 4a (details in Table S6). The NMDS clustering result (PERMANOVA; $p < .05$; ordination stress value: 0.0873) supports this, showing the cDNA and gDNA amplicon data clustered together, whereas both the metatranscriptomics and morphology data are distributed on the other side of the plot (Figure 4b).

3.4 | Application of metatranscriptomics to the field-collected samples

In terms of the consistency of metatranscriptomics at estimating the species diversity of actual field samples, replication testing revealed that mtCOI transcripts from biological and technical replicates provided fairly consistent results. There were no significant differences ($p > .05$) among the biological and technical replicates in terms of species diversity indices (Figure 5a,b) and composition (Figure 5c and Table S9). Succession in the temporal samples, in terms of changes in microcrustacean species composition, was successfully monitored using different mitochondrial transcript markers. Similar patterns were observed for each sampling date, and all known species at the sampling site, as reported by Chang et al. (2014), were documented (Figure 6 and Table S10). This reflects the versatility of metatranscriptomics at providing consistent taxonomic information for community ecology studies with the convenience of using various taxonomically important markers.

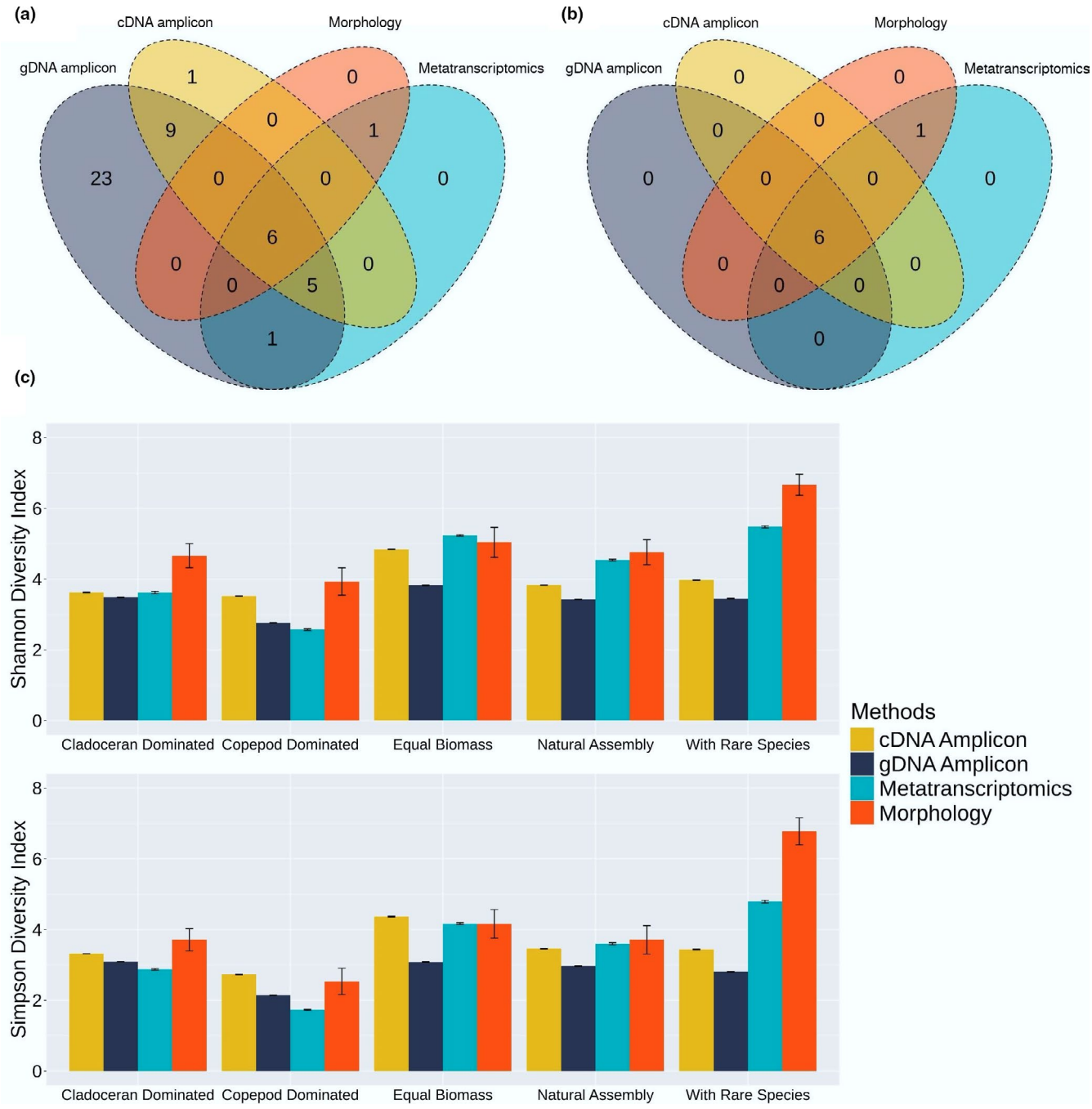

FIGURE 3 Comparison of the species diversity estimates derived from molecular-based approaches (cDNA, gDNA, and metatranscriptomics) and morphological data. Venn diagram showing (a) the number of shared observed species across the methods used with environmental contamination and (b) the number of shared species after extracting only target species sequences using Mothur (Schloss et al., 2009) and the MIDORI data set (Machida et al., 2017). (c) Estimation of diversity using Shannon and Simpson indices (ANOVA: $p > .05$ across methods)

4 | DISCUSSION

This study evaluated the use of metatranscriptomics for estimating the diversity and composition of zooplankton mock communities and field samples by comparing metatranscriptomics with PCR-based (gDNA and cDNA amplicons) and morphological-based methods. To do this, we assembled mock zooplankton communities with known taxonomic compositions. This approach has proven efficient for examining factors influencing successful species recovery and the accuracy of diversity estimation of molecular-based approaches (Braukmann et al., 2019). The use of mtCOI is ideal for benchmarking method accuracy due to its extensive reference library (Leray et al., 2019), taxonomic discriminatory power, and predictable sequence variation (Elbrecht et al., 2019). The mtCOI marker could delineate the seven freshwater zooplankton species studied here; however, a multi-marker approach is suggested for analysing more complex

communities, such as actual marine zooplankton samples (Stefanni et al., 2018).

For mock community analysis, metatranscriptomics provided the most reliable species diversity and community composition estimates, which closely resembled those derived from morphological data. The use of metatranscriptomics avoided the co-detection of extra-organismal eDNA and minimized background noise encountered during PCR-based methods, which may cause inflated estimates of species richness and complicated taxonomic assignment of sequences, especially with the absence of good quality reference databases (Molik et al., 2020). Another advantage of using RNA for monitoring zooplankton is that it avoids the bias related to NUMT pseudogene contamination (Collura et al., 1996). Comparison of the gDNA and cDNA mtCOI amplicon sequences revealed much higher sequence diversity (Table 2) in the gDNA amplicons; however, the degree of the observed differences was not consistent among species. For example, the ASV number from gDNA mtCOI amplicons of *Moina micrura* was 10 times greater than its cDNA counterpart, whereas only a minimal (1.3-fold) difference was noted for *Ceriodaphnia cornuta*. This shows the difficulty in estimating the real impact of NUMT pseudogene contamination on DNA-based diversity estimates using nucleotide diversity as a basis. The use of the ratio of nonsynonymous to synonymous substitutions (π[N]/π[S]) in amplicon sequences may provide additional insight into the presence of putative NUMT pseudogenes (i.e., *Mesocyclops leuckarti* gDNA mtCOI amplicons). The repeated transfer and "fossilization" of the continuously evolving mtDNA segments inserted in the nuclear genome creates multiple NUMT haplotypes with a predominance of synonymous substitutions (Perna & Kocher, 1996; Zischler et al., 1995). The presence of NUMT pseudogene sequences with predominant synonymous substitutions in gDNA amplicons can only be noted through comparison with cDNA sequences as a control group that excludes the presence of NUMT sequences. Using this additional criterion enables the detection of NUMT pseudogene sequences that resembles the normal and functional mitochondrial gene copy (Machida & Lin, 2017). These findings emphasize the importance of careful interpretation of amplicon sequences generated from a DNA-based approach. Lastly, bypassing PCR amplification in the metatranscriptomics workflow excludes any potential PCR-inherent biases that could lead to inaccurate inferences of zooplankton community composition (Piñol et al., 2019). A good example is the case of *Mongolodiaptomus birulai*, the most abundant species in our mock community samples (Figure 2b and Figure S2), which was not detected by either PCR-based method due to variable primer–template mismatches for this species (Piñol et al., 2014) despite the use of broadly used universal primer for metazoan communities. Having a variable number of primers–template mismatches across species may lead to inaccurate community analysis, with some species better amplified than others, such that the proportions in the final mixture do not reflect the original proportion of each species (Bista et al., 2018; Elbrecht & Leese, 2015; Leray et al., 2013).

With field-collected samples, metatranscriptomics provided highly similar community composition estimates for both technical

TABLE 2 Comparison of the number of amplicon sequence variants (ASVs), nucleotide diversity, and the number of substitutions and indels between methods based on genomic DNA (gDNA) and complement DNA (cDNA) mtCOI amplicons from six microcrustacean species

Taxa	Species (Total number of individuals used in the mock communities)	gDNA amplicons				cDNA amplicons			
		Number of ASVs	π	π (N)/π (S)	Number of indels	Number of ASVs	π	π (N)/π (S)	Number of indels
Copepoda	*Mesocyclops leuckartii* (118)	902	0.021	0.012/0.054	24	91	0.015	0.017/0.005	0
Cladocera	*Bosmina longirostris* (65)	42	0.036	0.043/0.022	0	19	0.011	0.014/0	0
	Ceriodaphnia cornuta (48)	65	0.031	0.033/0.024	0	50	0.012	0.012/0.016	0
	Daphnia galeata (68)	118	0.022	0.024/0.012	5	13	0.002	0/0.008	0
	Diaphanosoma dubium (32)	158	0.034	0.045/0.003	0	40	0.032	0.034/0.026	0
	Moina micrura (89)	449	0.025	0.022/0.023	6	41	0.003	0.047/0.008	0

Note: π, nucleotide diversity; (N), nonsynonymous substitution; (S), synonymous substitution; indel, insertion/deletion event. Values were derived from the combined sequences of all five mock communities constructed in this study.

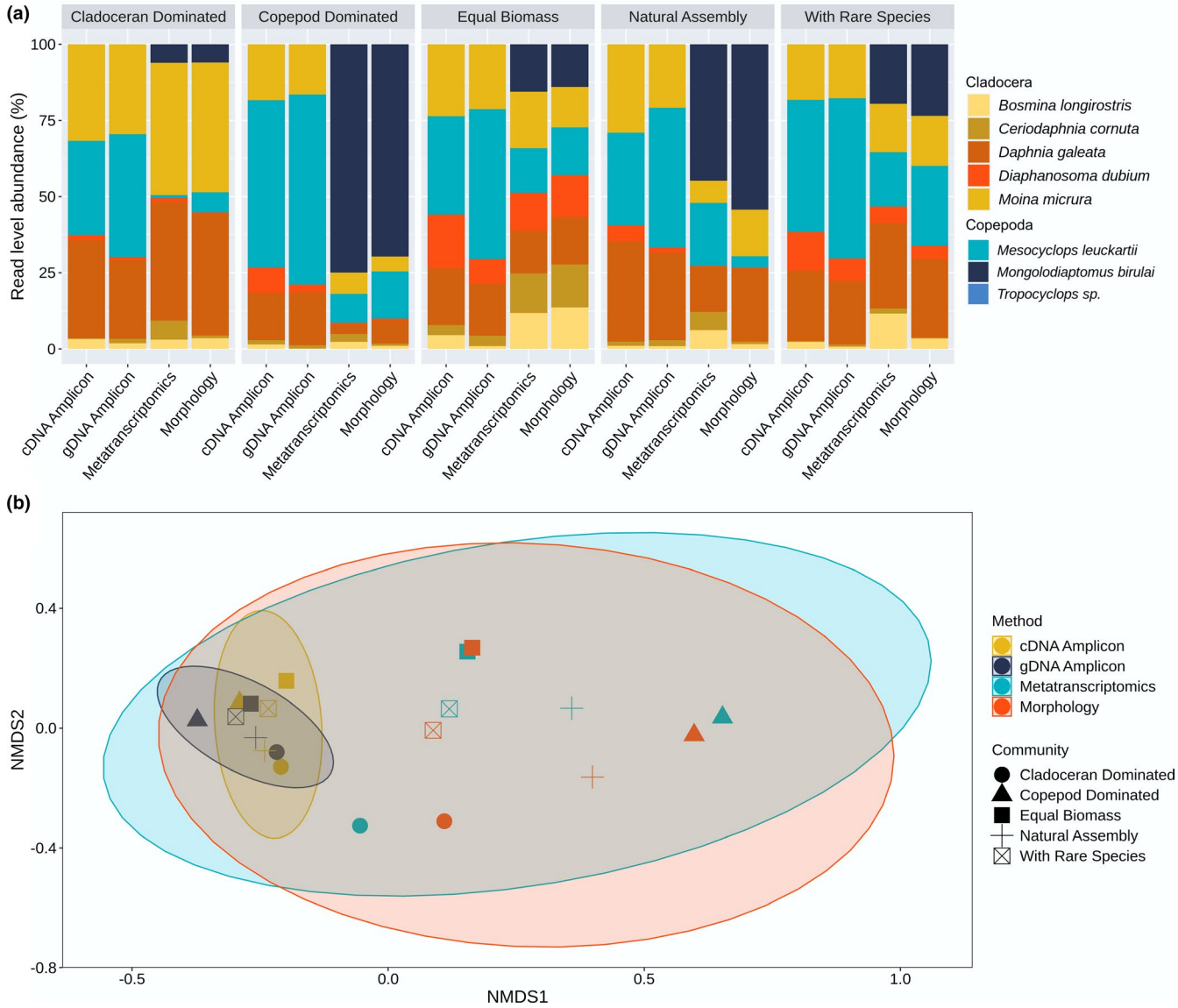

FIGURE 4 Comparison of the compositions of mock communities derived from molecular approaches (cDNA, gDNA, and metatranscriptomics) and morphology data. (a) Percentage read-level abundance (cDNA, gDNA, and metatranscriptomics) and relative dry weight biomass (morphology) of each species. (b) Nonmetric multidimensional scaling (NMDS) plot of the community compositions derived using each method (PERMANOVA: $p < .05$, ordination stress value: 0.0873)

and biological replicates. This means that researchers may consider limiting the number of replicates in future RNA-based zooplankton monitoring studies. This is particularly useful for cost-effective long-term zooplankton biomonitoring. Different mitochondrial markers (16S, COI, and CytB), which are easily retrievable from metatranscriptome sequences, are ideal for multi-marker approaches and can effectively discriminate closely related species in the community (Stefanni et al., 2018). Overall, this study demonstrates the potential use of metatranscriptomics for long-term ecological monitoring of complex metazoan communities, such as freshwater and marine zooplankton.

Despite the demonstrated advantages of using metatranscriptomics to analyse both mock communities and field-collected samples, several things must still be considered before applying this technique for zooplankton biomonitoring. Importantly, samples have to be handled carefully for high-quality RNA extraction. Unlike DNA, RNA can degrade if the samples are not preserved correctly in the field. RNAlater (Invitrogen) can prevent RNA degradation at 4°C or even room temperature (Gorokhova, 2005). Quality checking using the RIN can help ensure the use of only high-quality RNA (RIN > 7 or 8 indicates non-degraded, usable RNA) (Pérez-Portela & Riesgo, 2013). Additionally, unlike DNA metabarcoding that targets specific amplicon sequences, metatranscriptomics requires sequencing the overall community mRNA profile, which needs higher sequencing coverage. It is demonstrated in this study that Illumina MiSeq sequencing seems to be sufficient for accurate diversity estimation of both mock and field-collected communities; however, more advanced sequencing platforms (e.g., Illumina HiSeq, Illumina

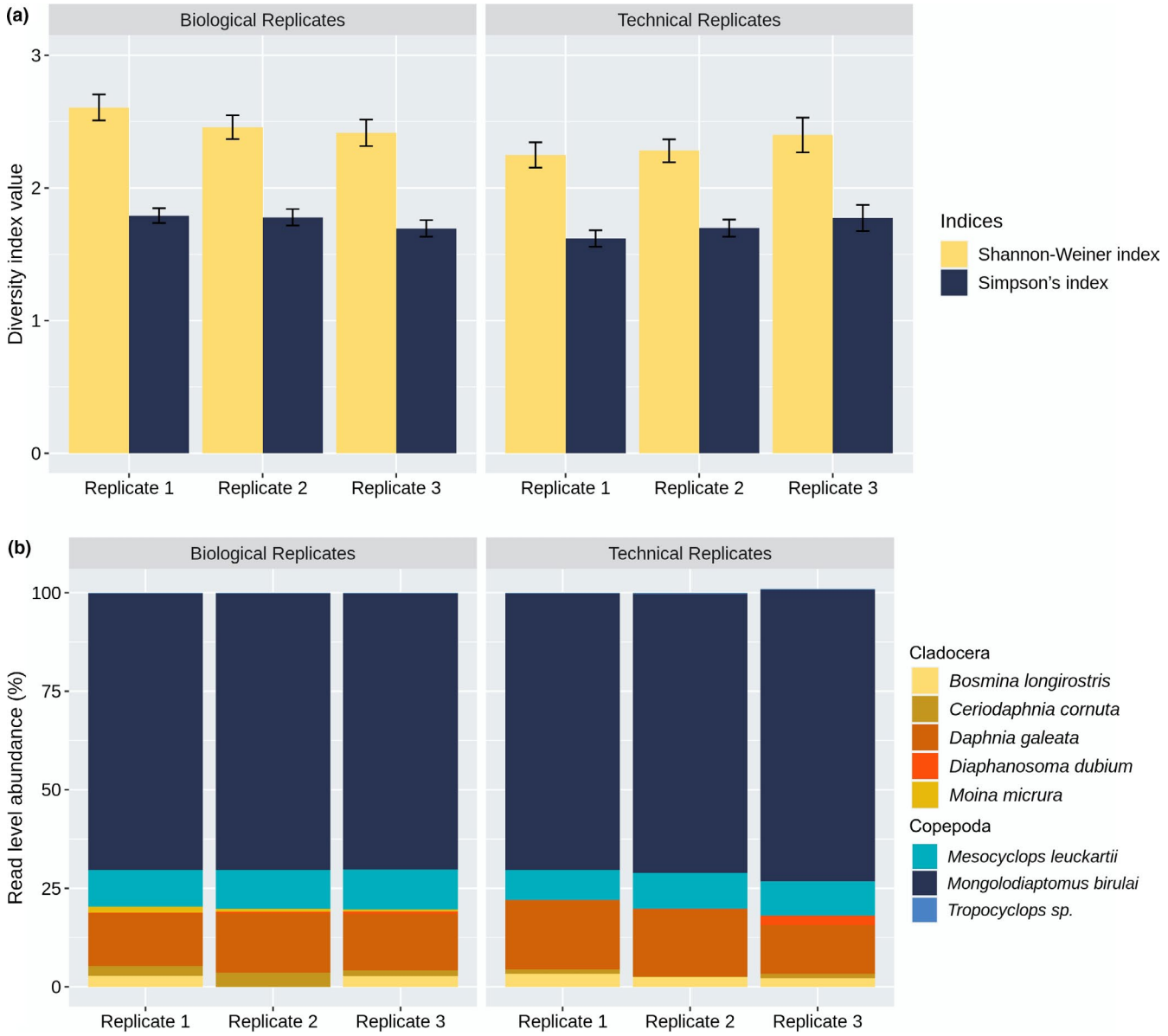

FIGURE 5 Comparison of diversity indices and community composition across biological and technical replicates of microcrustacean zooplankton samples collected from Fei Tsui Reservoir as inferred from mtCOI transcripts. Estimation of (a) diversity using the Shannon and Simpson diversity indices ($p > .05$) and (b) community composition based on percentage read-level abundance per species. Biological replicates comprised zooplankton samples from three independent vertical plankton net tows. Technical replicates comprised three independent metatranscriptome sequencing libraries prepared from RNA extracted from a single zooplankton community. Technical replicates were prepared from the third biological replicate

NextSeq, or PacBio SMRT sequencing) can be explored for successful recovery of marker sequences from all species, for example, mtCOI transcripts, in more complex metazoan communities like marine zooplankton. This may require additional costs, especially when multiplexing several libraries for sequencing. In terms of the bioinformatics workflow, the main technical limitations to using metatranscriptomics are related to the limited progress in assemblers specially designed for metatranscriptomics. Most assemblers were designed for genomic and transcriptomic data and are not ideal for use in metatranscriptomics due to uneven sequencing depth and repeat patterns occurring in different mRNAs (Leung et al., 2013). Furthermore, metatranscriptomics often involves mapping short-sequence reads to taxon-specific transcript references. The taxonomic assignments for mapped reads from mixed-species sequences can be challenging, especially if closely related taxa are present (Lenz et al., 2021). A high cross-species mapping rate may significantly affect the interpretation of metatranscriptomics data. Currently, there is no consensus on which bioinformatics tools or parameter settings should be used to address these concerns. But with advanced sequencing technologies, longer reads and greater depth sequencing are becoming more easily available at a much lower cost. Longer reads can currently be assembled and mapped more accurately

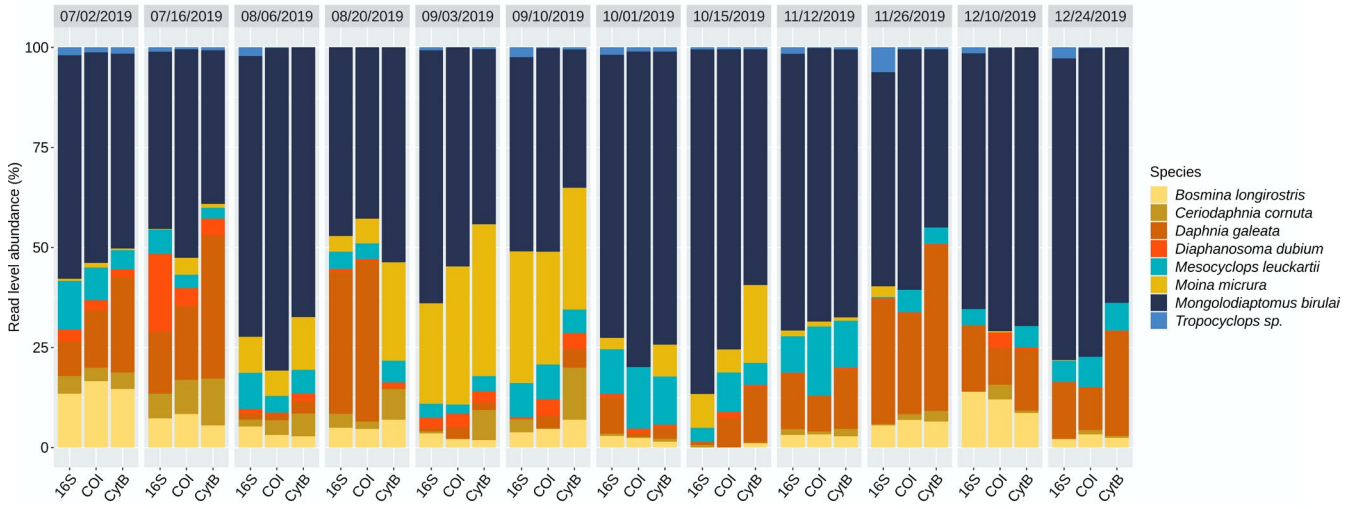

FIGURE 6 Temporal changes in community composition of the freshwater microcrustacean zooplankton in Fei Tsui Reservoir were tracked using three mitochondrial markers (mt16S, mtCOI, and mtCytB) in a metatranscriptomics analysis

(Kuosmanen et al.,2018). Overall, the use of metatranscriptomics in zooplankton monitoring is still in its infancy. The development of relevant field and laboratory protocols together with advances in bioinformatics tools are needed to allow metatranscriptomics to become a more optimized approach for zooplankton research.

5 | CONCLUSIONS

Our results demonstrated that metatranscriptomics can be used as an alternative approach to PCR-based methods (e.g., DNA metabarcoding) in characterizing zooplankton communities. Metatranscriptomics provided accurate diversity estimates for both mock community and field-collected zooplankton samples. With the isolation of mRNA transcripts and by-passing target gene amplification, metatranscriptomics avoided the co-detection of extra-organismal eDNA, minimized the presence of background noise, minimized co-amplification of putative NUMT pseudogenes, and minimized PCR-related biases that may contribute to inaccurate diversity estimation. Moreover, consistent species diversity estimates among replicates were observed from metatranscriptomics in actual zooplankton communities, while allowing the use of different mitochondrial transcripts as markers. Overall, these findings demonstrate that metatranscriptomics can be an effective tool for monitoring zooplankton diversity and community composition in given ecological contexts.

ACKNOWLEDGEMENTS

The first author was supported by a Taiwan International Graduate Programme scholarship for his PhD degree. This project was supported by Academia Sinica, Taiwan (RJM), the Ministry of Science and Technology, Taiwan (108-2611-M-001 and 109-2611-M-001; RJM), the Scientific Committee on Oceanic Research working group 157 (RJM), and National Taiwan University (109L8836; CHH). The funding agencies played no part in the study design, data collection, analysis, decision to publish, or manuscript preparation. The fieldwork assistance from Hsiang Yi Kuo, Chao Chen Lai, Kuo-yuan Li, Chin Chou Ye, Szulun Huang, and the Fei-Tsui Reservoir Administration Bureau is deeply appreciated. The authors also thank the NGS High Throughput Genomics Core at the Biodiversity Research Centre. The helpful comments by Matthieu Leray, Michael T. Monaghan, and four anonymous reviewers that improved the quality of the manuscript's first draft are also highly appreciated.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHOR CONTRIBUTIONS

Mark Louie D. Lopez, Ryuji J. Machida, Mitsuhide Sato, Chih-hao Hsieh, and Fuh-Kwo Shiah conceived the ideas and designed the methodology; Chih-hao Hsieh and Fuh-Kwo Shiah provided all the means for fieldwork to collect zooplankton samples from Fei Tsui Reservoir; Mark Louie D. Lopez and Ya-ying Lin conducted the molecular experiments; Mark Louie D. Lopez and Ryuji J. Machida analysed the data; and Mark Louie D. Lopez and Ryuji J. Machida led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

DATA AVAILABILITY STATEMENT

The raw sequences are accessible from the DNA Data Bank of Japan (DDBJ) under the accession numbers: DRA011141–DRA011142. The rest of the metadata for the mock communities and field samples are available in Appendix S1. All codes used for the bioinformatics procedures can be found at https://github.com/mldlopez/Using-metatranscriptomics-to-estimate-the-diversity-and-composition-of-zooplankton-communities.git.

ORCID

Mark Louie D. Lopez https://orcid.org/0000-0003-4288-4871
Ya-ying Lin https://orcid.org/0000-0002-6630-0837
Mitsuhide Sato https://orcid.org/0000-0002-4449-7050
Chih-hao Hsieh https://orcid.org/0000-0001-5935-7272
Fuh-Kwo Shiah https://orcid.org/0000-0001-5794-115X
Ryuji J. Machida https://orcid.org/0000-0003-1687-4709

REFERENCES

Andújar, C., Creedy, T. J., Arribas, P., López, H., Salces-Castellano, A., Pérez-Delgado, A. J., Vogler, A. P., & Emerson, B. C. (2021). Validated removal of nuclear pseudogenes and sequencing artefacts from mitochondrial metabarcode data. *Molecular Ecology Resources*, *21*(6), 1772–1787. https://doi.org/10.1111/1755-0998.13337

Bensasson, D., Feldman, M. W., & Petrov, D. A. (2003). Rates of DNA duplication and mitochondrial DNA insertion in the human genome. *Journal of Molecular Evolution*, *57*(3), 343–354. https://doi.org/10.1007/s00239-003-2485-7

Bensasson, D., Zhang, D., Hartl, D. L., & Hewitt, G. M. (2011). Mitochondrial pseudogenes: Evolution's misplaced witnesses. *Trends in Ecology & Evolution*, *16*, 314–321. https://doi.org/10.1016/S0169-5347(01)02151-6

Bista, I., Carvalho, G. R., Tang, M., Walsh, K., Zhou, X., Hajibabaei, M., Shokralla, S., Seymour, M., Bradley, D., Liu, S., Christmas, M., & Creer, S. (2018). Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. *Molecular Ecology Resources*, *18*(5), 1020–1034. https://doi.org/10.1111/1755-0998.12888

Braukmann, T. W. A., Ivanova, N. V., Prosser, S. W. J., Elbrecht, V., Steinke, D., Ratnasingham, S., de Waard, J. R., Sones, J. E., Zakharov, E. V., & Hebert, P. D. N. (2019). Metabarcoding a diverse arthropod mock community. *Molecular Ecology Resources*, *19*, 711–727. https://doi.org/10.1111/1755-0998.13008

Bucklin, A., Divito, K. R., Smolina, I., Choquet, M., Questel, J. M., Hoarau, G., & O'neill, R. J. (2018). Population genomics of marine zooplankton. In O. M. Rajora, & M. Oleksiak (Eds.), *Population genomics: Marine organisms* (pp. 61–102). Springer.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*, 581–583. https://doi.org/10.1038/nmeth.3869

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 421. https://doi.org/10.1186/1471-2105-10-421

Chang, C. W., Shiah, F. K., Wu, J. T., Miki, T., & Hsieh, C. H. (2014). The role of food availability and phytoplankton community dynamics in the seasonal succession of the zooplankton community in a subtropical reservoir. *Limnologica*, *46*, 131–138. https://doi.org/10.1016/j.limno.2014.01.002

Chen, H. (2018). *VennDiagram: Generate high-resolution Venn and Euler plots*. R package version 1.6.20. https://CRAN.R-project.org/package=VennDiagram

Collura, R. V., Auerbach, M. R., & Stewart, C. B. (1996). A quick, direct method that can differentiate expressed mitochondrial genes from their nuclear pseudogenes. *Current Biology*, *6*, 1337–1339. https://doi.org/10.1016/S0960-9822(02)70720-3

Creedy, T. J., Norman, H., Tang, C. Q., Qing Chin, K., Andujar, C., Arribas, P., O'Connor, R. S., Carvell, C., Notton, D. G., & Vogler, A. P. (2019). A validated workflow for rapid taxonomic assignment and monitoring of a national fauna of bees (Apiformes) using high throughput DNA barcoding. *Molecular Ecology Resources*, *20*(1), 40–53. https://doi.org/10.1111/1755-0998.13056

Cristescu, M. E. (2014). From barcoding single individuals to metabarcoding biological communities: Towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution*, *29*, 566–571. https://doi.org/10.1016/j.tree.2014.08.001

Dumont, H. J., & Tundisi, J. G. (1984). Tropical zooplankton. *Hydrobiologia*, *113*, 1–332.

Dumont, H. J., van de Velde, I., & Dumont, S. (1975). The dry weight estimate of biomass in a selection of Cladocera, Copepoda, and Rotifera from the plankton, periphyton, and benthos of continental waters. *Oecologia*, *19*(1), 75–97. https://doi.org/10.1007/BF00377592

Elbrecht, V., Braukmann, T., Ivanova, N. V., Prosser, S., Hajibabaei, M., Wright, M., Zakharov, E. V., Hebert, P., & Steinke, D. (2019). Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ*, *7*, e7745. https://doi.org/10.7717/peerj.7745

Elbrecht, V., & Leese, F. (2015). Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass-sequence relationships with an innovative metabarcoding protocol. *PLoS One*, *10*, e0130324. https://doi.org/10.1371/journal.pone.0130324

Fernando, C. H. (1994). Zooplankton, fish and fisheries in tropical freshwaters. *Hydrobiologia*, *272*, 105–123. https://doi.org/10.1007/BF00006516

Fernando, C. H. (2002). *A guide to tropical freshwater zooplankton – identification, ecology and impact on fisheries*. Backhuys Publishers.

Forster, D., Lentendu, G., Filker, S., Dubois, E., Wilding, T. A., & Stoeck, T. (2019). Improving eDNA-based protist diversity assessments using networks of amplicon sequence variants. *Environmetnal Microbiology*, *21*, 4109–4124. https://doi.org/10.1111/1462-2920.14764

Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A. J. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communication*, *8*, 1188. https://doi.org/10.1038/s41467-017-01312-x

Gorokhova, E. (2005). Effects of preservation and storage of microcrustaceans in RNAlater on RNA and DNA degradation. *Limnology and Oceanography: Methods*, *3*(2), 143–148.

Hlaing, T., Tun-Lin, W., Somboon, P., Socheat, D., Setha, T. O., Min, S., Chang, M. S., & Walton, C. (2009). Mitochondrial pseudogenes in the nuclear genome of *Aedes aegypti* mosquitoes: Implications for past and future population genetic studies. *BMC Genetics*, *10*, 11. https://doi.org/10.1186/1471-2156-10-11

Hsieh, C., Ma, K. H., & Chao, A. (2020). *iNEXT: Interpolation and extrapolation for species diversity*. R package version 2.0.20. Retrieved from http://chao.stat.nthu.edu.tw/wordpress/software_download/

Kassambara, A. (2020). *ggpubr: 'Ggplot2' based publication ready plots*. R package version 0.4.0. Retrieved from https://CRAN.R-project.org/package=ggpubr

Krehenwinkel, H., Wolf, M., Lim, J. Y., Rominger, A. J., Simison, W. B., & Gillespie, R. G. (2017). Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Scientific Reports*, *7*, 17668. https://doi.org/10.1038/s41598-017-17333-x

Kuosmanen, A., Norri, T., & Mäkinen, V. (2018). Evaluating approaches to find exon chains based on long reads. *Brief Bioinformatics*, *19*, 04–14.

Langmead, B., & Salzberg, S. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*, 357–359. https://doi.org/10.1038/nmeth.1923

Lenz, P. H., Lieberman, B., Cieslak, M. C., Roncalli, V., & Hartline, D. K. (2021). Transcriptomics and metatranscriptomics in zooplankton: wave of the future? *Journal of Plankton Research*, *43*(1), 3–9. https://doi.org/10.1093/plankt/fbaa058

Leray, M., Ho, S. L., Lin, I. J., & Machida, R. J. (2018). MIDORI server: A webserver for taxonomic assignment of unknown metazoan mitochondrial-encoded sequences using a curated database.

Bioinformatics, *34*, 3753–3754. https://doi.org/10.1093/bioinformatics/bty454

Leray, M., & Knowlton, N. (2015). DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, 2076–2081. https://doi.org/10.1073/pnas.1424997112

Leray, M., Knowlton, N., Ho, S. L., Nguyen, B. N., & Machida, R. J. (2019). GenBank is a reliable resource for 21st-century biodiversity research. *Proceedings of the National Academy of Sciences of the United States of America*, *116*, 22651–22656. https://doi.org/10.1073/pnas.1911714116

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., Boehm, J. T., & Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, *10*, 34. https://doi.org/10.1186/1742-9994-10-34

Leung, H. C., Yiu, S. M., Parkinson, J., & Chin, F. Y. (2013). IDBA-MT: *de novo* assembler for metatranscriptomic data generated from next-generation sequencing technology. *Journal of Computational Biology*, *20*(7), 540–550. https://doi.org/10.1089/cmb.2013.0042. PMID: 23829653.

Li, B., & Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*, 323. https://doi.org/10.1186/1471-2105-12-323

Lopez, M. L. D., Pascual, J. A., Dela Paz, E. S., Rizo, E. Z., Tordesillas, D., Guinto, S. K., Dumont, H., Mamaril, A. C., & Papa, R. D. S. (2017). Annotated checklist and insular distribution of freshwater microcrustaceans (Copepoda: Calanoida & Cyclopoida; Cladocera: Anomopoda & Ctenopoda) in the Philippines. *Raffles Bulletin of Zoology*, *65*, 623–654.

Machida, R. J., Hashiguchi, Y., Nishida, M., & Nishida, S. (2009). Zooplankton diversity analysis through single-gene sequencing of a community sample. *BMC Genomics*, *10*, 438. https://doi.org/10.1186/1471-2164-10-438

Machida, R. J., Kurihara, H., Nakajima, R., Sakamaki, T., Lin, Y. Y., & Furusawa, K. (2021). Comparative analysis of zooplankton diversities and compositions estimated from complement DNA and genomic DNA amplicons, metatranscriptomics, and morphological identifications. *ICES Journal of Marine Science*, fsab084. https://doi.org/10.1093/icesjms/fsab084

Machida, R. J., Leray, M., Ho, S., & Knowlton, N. (2017). Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. *Scientific Data*, *4*, 170027. https://doi.org/10.1038/sdata.2017.27

Machida, R. J., & Lin, Y. Y. (2017). Occurrence of mitochondrial CO1 pseudogenes in *Neocalanus plumchrus* (Crustacea: Copepoda): Hybridization indicated by recombined nuclear mitochondrial pseudogenes. *PLoS One*, *12*(2), e0172710. https://doi.org/10.1371/journal.pone.0172710

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *Embnet Journal*, *17*(1), 10. https://doi.org/10.14806/ej.17.1.200

Molik, D. C., Pfrender, M. E., & Emrich, S. J. (2020). Uncovering effects from the structure of metabarcode sequences for metagenetic and microbiome analysis. *Methods and Protocol*, *3*, 22. https://doi.org/10.3390/mps3010022.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., & Wagner, H. (2019). *Vegan: Community Ecology Package*. R package version 2.5–6. Retrieved from https://CRAN.R-project.org/package=vega

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, *28*(11), 1420–1428. https://doi.org/10.1093/bioinformatics/bts174

Pérez-Portela, R., & Riesgo, A. (2013). Optimizing preservation protocols to extract high-quality RNA from different tissues of echinoderms for next-generation sequencing. *Molecular Ecology Resources*, *13*, 884–889. https://doi.org/10.1111/1755-0998.12122

Perna, N. T., & Kocher, T. D. (1996). Mitochondrial DNA: Molecular fossils in the nucleus. *Current Biology*, *6*, 128–129. https://doi.org/10.1016/S0960-9822(02)00441-4

Piñol, J., Mir, G., Gomez-Polo, P., & Agustí, N. (2014). Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular Ecology*, *15*, 819–830. https://doi.org/10.1111/1755-0998.12355

Piñol, J., Senar, M. A., & Symondson, W. O. (2019). The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. *Molecular Ecology*, *28*, 407–419.

Piper, A. M., Batovska, J., Cogan, N. I. O., Weiss, J., Cunningham, J. P., Rodoni, B. R., & Blacket, M. J. (2019). Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. *GigaScience*, *8*(8), giz092. https://doi.org/10.1093/gigascience/giz092

Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A. H., Nieuwdorp, M., & Levin, E. (2020). Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLoS One*, *15*(1), e0227434. https://doi.org/10.1371/journal.pone.0227434

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ranwez, V., Harispe, S., Delsuc, F., & Douzery, E. J. P. (2011). MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PLoS One*, *6*(9), e22594. https://doi.org/10.1371/journal.pone.0022594

Richly, E., & Leister, D. (2004). NUMTs in sequenced eukaryotic genomes. *Molecular Biology and Evolution*, *21*(6), 1081–1084. https://doi.org/10.1093/molbev/msh110

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., & Sánchez-Gracia, A. (2017). DnaSP 6: DNA sequence polymorphism analysis of large datasets. *Molecular Biology and Evolution*, *34*, 3299–3302. https://doi.org/10.1093/molbev/msx248

Santoferrara, L. S. (2019). Current practice in plankton metabarcoding: optimization and error management. *Journal of Plankton Research*, *41*(5), 571–582. https://doi.org/10.1093/plankt/fbz041

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., & Weber, C. F. (2009). Introducing Mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*, 7537–7541. https://doi.org/10.1128/AEM.01541-09

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., Lightfoot, S., Menzel, W., Granzow, M., & Ragg, T. (2006). The RIN: An RNA integrity number for assigning integrity values to RNA measurements. *BMC Molecular Biology*, *7*, 3. https://doi.org/10.1186/1471-2199-7-3

Semmouri, I., de Schamphelaerea, K. A. C., Mees, J., Janssen, C. R., & Asselman, J. (2019). Evaluating the potential of direct RNA nanopore sequencing: Metatranscriptomics highlights possible seasonal differences in a marine pelagic crustacean zooplankton community. *Marine Environmental Research*, *153*, 104836. https://doi.org/10.1016/j.marenvres.2019.104836

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing genome assembly and annotation completeness. In M. Kollmar

(Ed.) *Gene prediction. Methods in molecular biology, 1962*, 227–245. Humana. PMID:31020564.

Shokralla, S., Gibson, J. F., Nikbakht, H., Janzen, D. H., Hallwachs, W., & Hajibabaei, M. (2014). Next-generation DNA barcoding: Using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Molecular Ecology Resources*, *14*(5), 892–901. https://doi.org/10.1111/1755-0998.12236

Song, H., Buhay, J. E., Whiting, M. F., & Crandall, K. A. (2008). Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 13486–13491. https://doi.org/10.1073/pnas.0803076105

Stefanni, S., Stanković, D., Borme, D., de Olazabal, D., Juretić, T., Pallavicini, A., & Tirelli, V. (2018). Multi-marker metabarcoding approach to study mesozooplankton at basin scale. *Scientific Reports*, *8*, 12085. https://doi.org/10.1038/s41598-018-30157-7

Triant, D. A., & Whitehead, A. (2009). Simultaneous extraction of high-quality RNA and DNA from small tissue samples. *Journal of Heredity*, *100*(2), 246–250. https://doi.org/10.1093/jhered/esn083

Valdes, C., & Capobianco, E. (2014). Methods to detect transcribed pseudogenes: RNA-Seq discovery allows learning through features. In L. Poliseno (Ed.), *Pseudogenes. Methods in Molecular Biology (Methods and Protocols)*, Vol. *1167* (pp. 157–183). Humana Press.

van der Loos, L. M., & Nijland, R. (2020). Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. *Molecular Ecology*, *30*, 3270–3288. https://doi.org/10.1111/mec.15592

Wang, Q., Garrity, G. M., Tiedje, J., & Cole, J. R. (2007). Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, *73*, 5261–5267. https://doi.org/10.1128/AEM.00062-07

Yang, J., Zhang, X., Xie, Y., Song, C., Zhang, Y., Yu, H., & Burton, G. A. (2017). Zooplankton community profiling in a eutrophic freshwater ecosystem–Lake Tai Basin by DNA metabarcoding. *Scientific Reports*, *7*, 1773. https://doi.org/10.1038/s41598-017-01808-y

Zhang, G. K., Chain, F. J. J., Abbott, C. L., & Cristescu, M. E. (2018). Metabarcoding using multiplexed markers increases species detection in complex zooplankton communities. *Evolutionary Applications*, *11*, 1901–1914. https://doi.org/10.1111/eva.12694

Zhao, F., Filker, S., Xu, K., Li, J., Zhou, T., & Huang, P. (2019). Effects of intragenomic polymorphism in the SSU rRNA gene on estimating marine microeukaryotic diversity: a test for ciliates using single-cell high-throughput DNA sequencing. *Limnology and Oceanography Methods*, *17*, 533–543. https://doi.org/10.1002/lom3.10330

Zischler, H., Geisert, H., von Haeseler, A., & Pääbo, S. (1995). A nuclear 'fossil' of the mitochondrial D-loop and the origin of modern humans. *Nature*, *378*, 489–492. https://doi.org/10.1038/378489a

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: Lopez, M. L. D., Lin, Y.-Y., Sato, M., Hsieh, C.-H., Shiah, F.-K., & Machida, R. J. (2022). Using metatranscriptomics to estimate the diversity and composition of zooplankton communities. *Molecular Ecology Resources*, 22, 638–652. https://doi.org/10.1111/1755-0998.13506