Metatranscriptomics provides closer diversity and composition estimates with morphology than PCR-based methods: a zooplankton mock community case study

Running title: Metatranscriptomics & PCR-based metabarcoding

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

[1]*Biodiversity Program, Taiwan International Graduate Program, Academia Sinica and National Taiwan Normal University, Nankang District, Taipei 11529, Taiwan*

[2]*Department of Life Science, National Taiwan Normal University, Wenshan District, Taipei 11677, Taiwan*

[3]*Biodiversity Research Center, Academia Sinica, Nankang District, Taipei 11529, Taiwan*

[4]*Department of Environment and Fisheries Resources, Nagasaki University, Nagasaki City 852-8521, Nagasaki Prefecture, Japan*

[5]*Institute of Oceanography, National Taiwan University, Da'an District, Taipei 10617, Taiwan*

[6]*Environmental Change Research Center, Academia Sinica, Nankang District, Taipei 11529, Taiwan*

*Corresponding author: Biodiversity Research Centre, Academia Sinica
128 Academia Road Sec. 2, Nankang Taipei 115, Taiwan
Tel : +886-2-2787-1585
Fax: +886-2-2785-8059
e-mail: ryujimachida@gmail.com

ABSTRACT

Studying complex metazoan communities requires taxonomic expertise and laborious work if done using the traditional morphological approach. Nowadays, the popular use of molecular-based methods accompanied by massively parallel sequencing (MPS) provides rapid and higher resolution diversity analyses. However, diversity estimates derived from the molecular-based approach can be biased by the co-detection of environmental DNA (eDNA), pseudogene contamination, and PCR amplification biases. Here, we constructed microcrustacean zooplankton mock communities to compare species diversity and composition estimates from PCR-based methods using genomic (gDNA) and complementary DNA (cDNA), metatranscriptomic transcripts, and morphology data. Mock community analyses show that gDNA mitochondrial cytochrome c oxidase I (mtCOI) amplicons inflate species richness due to environmental and nontarget species sequence contamination. Significantly higher amplicon sequence variant (ASV) and nucleotide diversity in gDNA amplicons than cDNA indicated the presence of putative pseudogenes. Last, PCR-based methods failed to detect the most abundant species in mock communities due to priming site mismatch. Overall, metatranscriptomic transcripts provided estimates of species richness and composition that closely resembled morphological data. The use of metatranscriptomic transcripts was further tested in field samples. The results showed that it could provide consistent species diversity estimates among biological and technical replicates while allowing monitoring of the zooplankton temporal species composition changes using different mitochondrial markers. These findings show that community characterization based on metatranscriptomic transcripts reflects the actual community more than PCR-based approaches.

Keywords: PCR bias, pseudogenes, mitochondrial transcripts, metatranscriptome.

1 | INTRODUCTION

Molecular-based approaches in tandem with massively parallel sequencing (MPS) are now widely used to estimate the diversity and composition of metazoan communities in marine, freshwater, and terrestrial ecosystems (Kennedy et al., 2020; Piredda et al., 2018; Sun et al., 2018; Wilson, Sing, Lee, & Wee, 2016; Yang et al., 2017). Advantages of using molecular-based methods over more traditional methods that rely on morphology include, but are not limited to, (a) the effective detection of rare species (Leasi et al., 2018), (b) the ability to identify and estimate diversity from samples that include early life stages lacking diagnostic morphological characters (Machida, Hashiguchi, Nishida, & Nishida, 2009), and (c) the high speed and low cost required to quantify alpha and beta diversity from samples that include thousands to millions of individual specimens (Yang et al., 2017). However, molecular approaches have their own biases (van der Loos & Nijland, 2020). Some of them are inherent to the polymerase chain reaction (PCR) typically used to enrich specific genes. Other biases are linked to the codetection of nonfunctional gene sequences (i.e., pseudogenes) when using a genomic DNA (gDNA) template. As a result, molecular methods that bypass PCR and target mRNA (i.e., metatranscriptomics) rather than gDNA are likely to gain popularity in the coming years. Yet, metatranscriptomics' performance has not been rigorously compared with the performance of PCR-based methods and morphology for taxonomic and ecological characterization of metazoan community samples.

DNA metabarcoding is the most commonly used method in molecular-based metazoan community studies (Braukmann et al., 2019; Cowart et al., 2015). It requires PCR amplification of a target gene region (e.g., mtCOI) from the gDNA before MPS library preparation (Cristescu, 2014; Elbrecht & Leese, 2015). With its dependence on PCR amplification, the metabarcoding approach may provide inaccurate diversity estimation due to

amplification biases. This occurs when primers fail to bind effectively to sequences of specific taxa, thus misrepresenting the composition of complex or diverse samples (Krehenwinkel, Wolf, Lim, Simison, & Gillespie, 2017). Moreover, the amplification of nuclear-encoded mitochondrial pseudogenes may lead to another set of biases when using a gDNA template. Occurrences of the pseudogene were well documented in various metazoan taxa, especially in animals with large nuclear genome sizes (Bensasson, Zhang, Hartl, & Hewitt, 2011; Machida & Lin, 2017). The presence of putative mitochondrial pseudogenes inflated species richness in previous studies (Song, Buhay, Whiting, & Crandall, 2008).

RNA-based methods that do not rely on PCR, such as metatranscriptomics, are potentially less prone to biases when characterizing metazoan communities (Semmouri, de Schamphelaerea, Mees, Janssen, & Asselmanad, 2019). Isolating mRNA transcripts rather than gDNA excludes pseudogenes because pseudogenes are not transcribed into a mature mRNA (Collura, Auerbach, & Stewart, 1996; Hlaing et al., 2009; Valdes & Capobianco, 2014). Also, metatranscriptome library preparation does not require amplification of a target gene region through PCR, thus avoiding biases related to primer binding efficiency. As such, metatranscriptomic transcripts may provide more accurate estimates of diversity in complex metazoan communities.

Here, we use freshwater microcrustacean zooplankton mock communities with known taxonomic composition to compare three molecular-based methods and morphological analysis for diversity estimation: (a) morphological analysis as a standard taxonomic approach for studying metazoan communities; (b) mtCOI amplicons from gDNA as a template to see any possible effects of pseudogene contamination and PCR amplification bias; (c) mtCOI amplicons from RT-PCR complementary DNA (cDNA) as a template, where we can avoid contamination of the mitochondrial pseudogenes (Collura et al., 1996; Hlaing et al.,

2009; Valdes & Capobianco, 2014) but we may still see effects of PCR-derived bias due to amplification of the target gene; and (d) bioinformatically selected mtCOI transcripts from metatranscriptomics to avoid pseudogene contamination and PCR amplification bias. Last, we evaluate the suitability of using metatranscriptomic transcripts in monitoring temporal changes in the composition of microcrustacean zooplankton communities from a subtropical reservoir (Fei Tsui Reservoir, Taiwan). Our results indicate that the characterization of metazoan communities can be more reliable with metatranscriptomic transcripts than with PCR-based approaches.

2 | MATERIALS AND METHODS

2.1 | Sample collection

Freshwater microcrustacean zooplankton (Arthropoda: Cladocera and Copepoda) collected from Fei Tsui Reservoir, a subtropical reservoir located in Northeastern Taiwan (24°54'34.9"N 121°34'53.0"E; altitude of 160 m) were used in this study for several reasons: (a) the taxonomy of resident species is well known and (b) long-term and ongoing monitoring data for the zooplankton community in Fei Tsui reservoir is available. A 45 cm mouth-wide conical plankton net (50 μm mesh size) with an attached flow meter was hauled vertically from 50 m to the surface to collect zooplankton samples. The sample was further filtered with a 100 μm mesh bag to remove lake water and small nontarget taxa like rotifers and phytoplankton. It was then immediately soaked in 10X sample volume of RNAlater (Invitrogen, USA) for 15 minutes to allow the remaining lake water to mix with the solution (Gorokhova, 2005). Afterward, the sample was transferred to a new container with the same volume of RNAlater to ensure the proper preservation of both RNA and DNA. The preserved sample was transported to the lab at room temperature (within ca. 2-3 hours), stored at 4 °C for 24 hours, and then transferred to -20 °C for longer storage until the DNA/RNA extraction.

Individuals used to prepare mock communities were isolated from the RNAlater preserved samples collected on August 20, 2019. For the replication test, biological replicates (three different plankton haulings within the same site) and technical replicates (three independent total RNA aliquots from the same biological sample) were used to check the metatranscriptomics' consistency (collected on December 24, 2019). Last, samples used for monitoring temporal changes in the species composition of microcrustacean zooplankton were collected from July 2 to December 24, 2019.

2.2 | Mock community preparation

A total of five mock communities were constructed using zooplankton samples collected from the field (see details of the composition in Table 1): (a) cladoceran dominated, (b) copepod dominated, (c) equal biomass: equal biomass among species, (d) natural assembly: mimicking actual community composition in the reservoir, and (e) with rare species: the presence of a rare species. Each community contained five cladoceran and two copepod species (Table 1). The number of species used in the mock community was limited to the dominant microcrustacean species documented in the Fei Tsui Reservoir (Chang, Shiah, Wu, Miki, & Hsieh, 2014). The body length of the preserved individuals used in constructing the mock communities was measured under the stereomicroscope (Nikon, Japan) to allow the calculation of dry weight biomass (in μg) based on the length-weight regression equation (Dumont, van de Velde, & Dumont, 1975). Figure 1 gives a summary of the workflow for processing the constructed mock communities.

2.3 | DNA and RNA extraction

The total RNA was extracted from each mock community using TriPure Isolation reagent (Roche, Switzerland) in conjunction with a PureLink RNA Mini Kit (Invitrogen, USA). First, sorted individuals (Table 1) preserved in RNAlater were homogenized in 1 mL

142 of TriPure isolation reagent until animals' tissues were thoroughly fragmented. Next, 200 μL
143 of chloroform was added to the tube and shaken vigorously for 15 seconds. The sample was
144 then incubated for 30 minutes at room temperature and placed in the centrifuge (12,000 g for
145 15 min at 4 °C) to separate into two phases. A total of 100 μL of the resulting upper aqueous
146 phase was transferred to a new tube containing an equal volume of 99.8% ethanol, while the
147 remaining phase was set aside for gDNA extraction. The solution was then run through the
148 PureLink Mini Kit spin column by spinning in the centrifuge at 12,000 g for 1 min at room
149 temperature. The resulting flow-through was discarded, and the column was transferred to a
150 new collection tube. It was then washed with 500 μL of Buffer II from the kit twice. After
151 washing, the column was centrifuged at 12,000 g for 1 minute at room temperature to dry the
152 membrane completely. Last, the column was transferred to a new recovery tube, and 100 μL
153 of RNase-free water was added to elute the RNA from the membrane. The quality and
154 concentration of all extracted total RNA samples were analyzed using Bioanalyzer RNA 6000
155 nano (Agilent Technologies, USA) to measure RNA integrity number (RIN), which is
156 calculated based on the areas of 18S rRNA and 28S rRNA, where 1 is the most degraded
157 profile and 10 is the most intact (Schroeder et al., 2006). All samples with RIN values greater
158 than 7 were processed and stored at -80 °C until the next part of the procedure (Table S1).
159 Genomic DNA extractions from the same mock community sample were performed
160 using a DNeasy kit (Qiagen, Netherland) in conjunction with Back Extraction Buffer (BEB: 4
161 M guanidine thiocyanate, 50 mM sodium citrate, and 1 M Tris [free base];
162 https://www.thermofisher.com/tw/en/home/references/protocols/nucleic-acid-purification-
163 and-analysis/dna-extraction-protocols/tri-reagent-dna-protein-isolation-protocol.html). A total
164 of 120 μL of BEB was added to the remaining phase that was set aside during the RNA
165 extraction and mixed vigorously by hand for 1 minute. The solution was incubated for 10

minutes at room temperature and centrifuged at 12,000 g for 15 minutes at 4 °C. Afterward, 200 µL aliquot of the aqueous phase was transferred to a new 1.5 mL tube. This was followed by the addition of 200 µL AL buffer from the Qiagen kit together with 200 µL of 99.8% ethanol. The mixture was then transferred to the Qiagen kit spin column and centrifuged at 6,000 g for 1 minute at room temperature. The column was placed into a new collection tube and washed with 500 µL of Qiagen AW1 and AW2 buffer. The washed membrane was then dried by centrifugation at 20,000 g for 3 minutes. Last, the dried column was again transferred to a new tube and eluted with 100 µL of the Qiagen AE buffer. The extracted gDNA was further purified using Agencourt AMPure XP (Beckman Coulter, USA) following the manufacture's protocol. The purified gDNA's concentration and quality were measured using NanoDrop 2000 (Thermo Fisher Scientific, USA) and Qubit Fluorometric Quantitation (Thermo Fisher Scientific, USA).

In processing the field samples for metatranscriptomic replication testing (biological and technical replicates) and monitoring temporal changes in the species composition of microcrustacean zooplankton in Fei Tsui Reservoir, the preserved samples were carefully taken off the mesh bags and weighed using a micro balance (Denver Instrument, USA) to determine the wet weight. Afterward, the weighed zooplankton samples were processed using the same protocol for extracting total RNA from the mock community samples (Tables S2 and S3). All RNA samples were stored at -80 °C until the next part of the procedure.

2.4 | PCR amplification and sequencing

The gDNA used for PCR amplification was the direct product of the DNA extraction and purification from the previous steps. In contrast, the cDNA was prepared using mRNA purified from the total RNA through the use of the Dynabeads mRNA purification kit

(Invitrogen, USA). This was then followed by reverse transcription of 150 ng of isolated mRNA using the SuperScript IV VILO Master Mix (Invitrogen, USA) standard protocol.

The amplification of the mtCOI from the gDNA and cDNA templates was done by preparing a 50 µL reaction volume containing 10 ng of 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, Japan), and nuclease-free water filled up to 50 µL. The PCR amplification was run using a Veriti Thermal Cycler (Applied Biosystems, USA) with Touchdown PCR conditions: initial denaturation at 95 °C for 10 minutes; denaturation at 95 °C for 10 seconds; annealing at 62 °C for 30 seconds; and extension at 72 °C for 60 seconds. The annealing temperature was progressively reduced with advancing cycles (-1.0 °C per cycle) from 62 to 46 °C during the first 16 cycles and kept constant at 46 °C during the subsequent 20 cycles. The mtCOI primers used in this PCR are mlCOIintF: GGWACWGGWTGAACWGTWTAYCCYCC combined with jgHCO2198: TAIACYTCIGGRTGICCRAARAAYCA to target a 313 bp fragment (Leray et al., 2013). The use of mtCOI as a marker gives an advantage due to the higher number of reference sequences present in the database (Machida, Leray, Ho, Nguyen, & Knowlton, 2017). A PCR mixture without a template was also prepared as a negative control. After the PCR, the amplicon band's expected length from each sample, together with the absence of amplicon band in the negative control, was confirmed by the gel image. Lastly, the amplicons' size selection and purification were performed using Agencourt AMPure XP (Beckman Coulter, USA).

A second PCR reaction was done for the attachment of the barcode adapter. This time, the PCR reaction was carried out using different barcoded primers for each mock community reaction (Table S4). The same amount of template (10 ng) was used for each reaction using

213 the following conditions for 20 cycles: initial denaturation at 95 °C for 10 minutes;
214 denaturation at 95 °C for 10 seconds; annealing at 62 °C for 30 seconds; and extension at 72
215 °C for 60 seconds. After the PCR, the amplicon' size selection was again performed using
216 Agencourt AMPure XP (Beckman Coulter, USA). The DNA concentration measurement was
217 done using Qubit Fluorometric Quantitation (Thermo Fisher Scientific, USA). Then, a total of
218 100 ng for each of the purified samples was pooled, purified with 0.9X Agencourt AMPure
219 XP (Beckman Coulter, USA), and eluted with 30 µl of nuclease-free water. Last, the prepared
220 libraries were sent for Illumina MiSeq 300 PE sequencing (1% PhiX spike-in and 10 pM
221 loading concentration for all libraries) at the NGS High Throughput Genomics Core at the
222 Biodiversity Research Centre, Academia Sinica, Taiwan.

223 **2.5 | Metatranscriptomic library preparation and sequencing**

224 Metatranscriptomic library was prepared using NEBNext mRNA Library Prep
225 Reagent Set for Illumina (E6110) together with NEBNext Poly(A) mRNA Magnetic Isolation
226 Module (E7490) and NEBNext® Multiplex Oligos for Illumina (New England BioLabs, USA)
227 following the manufacturer's protocol. Five µg of the total RNA was used to start the library
228 preparation. Final enrichment was performed for 15 cycles. After the purification of the
229 enriched product using 0.9X Agencourt AMPure XP, equal amounts of those products were
230 pooled together and sent for the Illumina MiSeq 300 PE sequencing (1% PhiX spike-in and
231 10 pM loading concentration for all libraries) at the NGS High Throughput Genomics Core at
232 the Biodiversity Research Center, Academia Sinica, Taiwan.

233 **2.6 | Bioinformatics**

234 All codes used for the bioinformatic procedures for this study are at
235 https://bit.ly/3lDPSfd. For both gDNA and cDNA mtCOI amplicons, sequences were
236 processed by quality filtering and adapter removal with a minimum Phred quality score of 10

using Cutadapt (ver. 2.10, Martin, 2011). The number of sequences for each community sample was normalized by the random selection of an equal number of reads using Seqtk (https://github.com/lh3/seqtk) (Table S5). The sequences were then subjected to the DADA2 pipeline for further quality filtering, merging paired reads, and removing chimeras using default commands (Callahan et al., 2016). The resulting unique amplicon sequence variants (ASVs) 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 (ver. 1.44.3; Schloss et al., 2009) using COI reference dataset from MIDORI Longest 1.1 (Machida et al., 2017). This is to remove the high number of sequences from nontarget taxa, thus leaving the target species' sequences. Both the filtered and unfiltered sequences were used in comparing the methods in terms of species richness detection; however, only the filtered sequences for the target taxa were used for species diversity estimation and species composition construction to allow a more thorough analysis of the actual mock community. The ASVs were then clustered into operational taxonomic units (OTUs) with an identity criterion of 94% similarity using the –cluster_fast command of VSEARCH (ver. 2.15; Rognes, Flouri, Nichols, Quince, & Mahé, 2016). The 94% similarity threshold was chosen based on preliminary analysis delineating the target species. Meanwhile, the VSEARCH centroid sequences were used for taxonomic assignment of the OTUs using the RDP Classifier (Wang, Garrity, Tiedje, & Cole, 2007) function in the MIDORI server (Leray, Ho, Lin, & Machida, 2018) using MIDORI Longest 1.1 (Machida et al., 2017) 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 LULU's default command to remove erroneous molecular operational taxonomic units (ver. 1.2.3; Frøslev et al., 2017). The OTU

tables from the sequence curation were used to calculate the species richness indices using the iNEXT package (Hsieh, Ma, & Chao, 2020) within the R platform (R Core Team, 2017).

The metatranscriptomic transcripts were processed as follows. The raw sequences were first prepared by quality filtering and adapter removal using Cutadapt with a minimum Phred quality score of 10 (ver. 2.10; Martin, 2011). The cleaned paired-end reads were then used for transcripts assembly using Trinity assembler (ver. 2.11.0; Grabherr et al., 2011) following the default parameters (the assembly statistics are in Table S6). Then, assembled 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 protein and two ribosomal RNA) from MIDORI_LONGEST 1.1 datasets (Machida et al., 2017). From the BLAST results, the mtCOI were pulled-out using a constructed Perl script (https://bit.ly/3lDPSfd). The resulting mtCOI contigs were filtered using the classify.seqs and get.lineage functions of Mothur (ver. 1.44.3; Schloss et al., 2009) using the COI reference dataset from MIDORI Longest 1.1 (Machida et al., 2017) to get the sequences of 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 (subsampled equal number of raw reads with Seqtk; Table S1). Then, the read-level abundance was quantified in transcripts per million (TPM) with RSEM (ver. 1.2.31; Li & Dewey 2011) within the Trinity pipeline (Haas et al., 2013) using the default commands. The read-level species richness indices were calculated in the iNEXT package

283 (Hsieh et al., 2020) within the R platform (R Core Team, 2017) using the output file from
284 RSEM.

285 The same bioinformatics workflow was used for processing field community samples
286 for both metatranscriptomic replication testing and monitoring of temporal changes in the
287 community composition of microcrustacean zooplankton in the reservoir: quality filtering,
288 transcript assembly, extraction of selected mitochondrial genes for reference construction,
289 mapping back raw reads to the assembled reference, read-level abundance quantification, and
290 calculation of species diversity indices. The details for the number of processed reads for the
291 biological replicates and technical replicates are in Table S2. Meanwhile, supplemental
292 information on the use of different mitochondrial transcripts (16S, COI, and CytB) from
293 metatranscriptomics in monitoring temporal changes in zooplankton composition from July to
294 December 2019 is in Table S3. To address the lack of reference sequences for the 16S and
295 CytB for the target species in GenBank, the taxonomic assignment was carried out using a
296 modified MIDORI Longest 1.1 (Machida et al., 2017) reference dataset with added sequences
297 for the following species: *Mongolodiaptomus birulai*; *Mesocyclops leuckartii*, *Bosmina*
298 *longirostris*, *Ceriodaphnia cornuta*; and *Moina micrura*.

299 **2.7 | Statistical analyses**

300 The similarity in the species richness detected by each method was compared using a
301 Venn diagram constructed using the VennDiagram package (Chen, 2018). Furthermore,
302 statistical differences between the species diversity indices (Shannon and Simpson's Indices)
303 provided by each method were tested using ANOVA through the ggpubr package
304 (Kassambara, 2020). Last, NMDS clustering to compare similarities in the species
305 composition from each method was performed using the vegan package (Oksanen et al.,
306 2019). All these statistical analyses were done within the R platform (R Core Team, 2017).

307 Meanwhile, comparison of the gDNA and cDNA sequences through the calculation of
308 nucleotide diversity, synonymous (π (S)) and nonsynonymous substitution (π (N)), and indel
309 (insertion/deletion) events to inspect pseudogenes' presence was carried out using DNAsp
310 (Rozas et al., 2017).

311 **3 | RESULTS**

312 **3.1 | Sequencing**

313 A total of 8,717,291 and 8,815,608 raw reads were generated from the cDNA and
314 gDNA mock community mtCOI amplicon libraries, respectively. Demultiplexed sequences
315 subjected to the DADA2 pipeline retained an average of 72% and 81% good quality reads of
316 the input sequences for downstream analyses, respectively (Table S5). Meanwhile, the mock
317 communities' metatranscriptomic sequences yielded 23,382,940 reads that were
318 demultiplexed into five different mock community libraries (Table S1). First, the reads were
319 assembled into contigs that allowed construction of the mtCOI transcript reference for each
320 community (details for the assembly report are in Table S6). Next, each mock community's
321 raw sequences were subsampled and mapped back on to the assembled reference mtCOI
322 contigs. Last, a total of 8,468,448 (Table S2) and 15,728,180 (Table S3) raw reads were
323 generated for the replication test and the zooplankton community's temporal monitoring,
324 respectively. The reads were demultiplexed and processed using the same workflow as for the
325 constructed mock communities' metatranscriptomic transcript.

326 **3.2 | Comparison of gDNA and cDNA mtCOI amplicons**

327 Both PCR-based methods using DNA and RNA (cDNA) detected nontarget species
328 sequences, including those that possibly originated from epiphytes attached to the samples,
329 zooplankton gut content, and extraorganismal environmental DNA (eDNA) and RNA
330 (eRNA). A total of 45 and 19 OTUs were detected by gDNA and cDNA, respectively (Figure

2A). The gDNA amplicons provided a higher OTU richness, including 27 OTUs not observed in cDNA amplicons. Some of these OTUs from the gDNA amplicons were identified as taxa that are unusually present in the reservoir's limnetic area like marine bryozoan and a spider (Figure S1 and Table S7). The cDNA amplicons reflected the gDNA amplicons' subset data with 18 shared and one exclusive (Rotifera: *Conochilus unicornis* with 33 sequence reads in one mock community) OTUs. To examine the pseudogene contamination, sequences of six target species in the mock communities that were detected by both the gDNA and cDNA amplicons were compared (Table 2). Overall, the extent of sequence variation of gDNA amplicon for all six species was much greater than its cDNA counterparts. A much larger number of ASV (1.3-11.2 times more) was observed among the gDNA amplicon sequences than the cDNA sequences in all species. Last, greater diversity in nucleotide diversity (π) and indel events were noted in gDNA sequences relative to cDNA. This difference is prominent in *Mesocyclops leuckartii*, where 4.5 times more synonymous than nonsynonymous substitutions were observed.

3.3 | Comparison between metatranscriptomics and PCR-based methods

To compare the methods better in estimating the species diversity of the constructed mock communities, the target species' sequences were filtered (Figure 2B) and utilized for this study's subsequent analyses. From the filtered sequences, it can be noted that only the non-PCR-based method, metatranscriptomic transcript, was able to detect all species present in the actual mock communities. Both cDNA and gDNA mtCOI amplicons failed to detect *Mongolodiaptomus birulai* (the most abundant species in the Fei Tsui Reservoir) in all mock communities. Failure to detect this species can be explained by the observed mismatches between the species' priming site sequences (Figure S2).

In terms of species diversity indices, both Shannon and Simpson's diversity indices from the three molecular-based methods (Figure 2C) failed to exhibit any significant differences with the morphological data (ANOVA: 0.05 < *p*-value). This is despite the absence of one species that was not amplified in both cDNA and gDNA amplicons. In terms of examining species composition based on read-level abundance, both cDNA and gDNA amplicons showed a highly similar species composition for all mock communities. On the other hand, the metatranscriptomic transcript-based species composition showed very high similarities to the one observed on morphological data, as shown in Figure 3A (details in Table S8 and S9). The NMDS clustering (stress value = 0.1046) further supports this, where the cDNA and gDNA amplicon data clustered together, while both metatranscriptomic and morphology data spread on the other side of the plot (Figure 3B).

3.4 | Application of metatranscriptomics to the field-collected zooplankton community

In terms of metatranscriptomic transcripts' consistency in 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 (ANOVA: 0.05 < *p*-value) observed among the replicates for both biological and technical samples in terms of species diversity indices (Figure 4A and 4B) and composition (Figure 4C and Table S10). The succession of microcrustacean species composition in temporal samples was successfully monitored using different mitochondrial transcript markers that showed similar patterns for each sampling date while detecting all known species documented in the sampling site (Figure 5 and Table S11) based on the previous literature. This reflects metatranscriptomics' versatility in providing consistent taxonomic information for community ecology studies with the convenience of using various taxonomically important markers.

4 | DISCUSSION

We compared three molecular-based methods and morphological analysis in characterizing constructed zooplankton mock communities in the present study. For molecular methods, we have used encoded mitochondrial (mt) markers for characterizing zooplankton communities. The mitochondria produce the energy currency, ATP, through cellular respiration. Therefore, it is assumed that the mt gene abundance reflects each species' energy production or respiration potential in the community. Here, we have used both DNA and RNA (cDNA and metatranscriptomics) as a starting template for the analyses, where gDNA mtCOI abundance shows the copy number of the mt genome present in each individual per species. In contrast, cDNA and metatranscriptomic mtCOI abundances reflect transcribed mt protein-coding genes at the current time point. The transcription of mt protein-coding genes requires large quantities of phosphorus, which often becomes a limiting factor for animal growth in many environments (Warner, 1999). For this reason, we assume that the RNA abundance reflects a more accurate picture of the short-term respiration potential dynamics in energy production than gDNA reads.

Furthermore, several biases that created diversity and community estimation errors were encountered using gDNA for PCR-based metabarcoding. First, it is assumed that the higher OTU richness in gDNA amplicon sequences than cDNA (Figure 2) was due to the contamination of environmental DNA, small nontarget taxa attached to the target species, and zooplankton gut content sequences (Figure S1). The use of gDNA as a PCR template tends to include nontarget sequence contaminants compared to RNA (Rees, Maddison, Middleditch, Patmore, & Gough, 2014). This observation suggests that the potential overestimation of diversity was caused by eDNA contamination and nontarget species sequences in the mock communities' extracted gDNAs.

Another bias encountered with the use of gDNA amplicons in this study is the amplification of putative pseudogene sequences. Mitochondrial pseudogenes are usually not transcribed into mature mRNA (Collura et al., 1996). Therefore, we can avoid contamination of mitochondrial pseudogenes by analyzing the prepared cDNA. The comparison between gDNA and cDNA mtCOI amplicons demonstrated much higher diversity (Table 2; the number of ASVs and nucleotide diversity) in the gDNA amplicons; however, the observed differences were not consistent between taxonomic groups. For example, more than 10 times as many ASVs were observed from gDNA than from cDNA in *Moina micrura*. In contrast, less difference was observed in *Ceriodaphnia cornuta* (1.3 times). This observation indicates difficulty in estimating the impact of pseudogene on the analyses, which are amplified from gDNA. Additionally, higher nucleotide diversity in synonymous substitution than in nonsynonymous substitution of the gDNA amplicons was observed in one species: *Mesocyclops leuckartii*. The repeated transfer and fossilization of the continuously evolving mt DNA segments inserted in the nuclear genome may create multiple haplotypes with a predominance of synonymous substitutions (Perna & Kocher, 1996; Zischler, Geisert, von Haeseler, & Pääbo, 1995). This may confuse mitochondrial pseudogenes, making them look functional, despite being nonfunctionally encoded in the nuclear genome. A similar result was observed in individual-based analyses of marine copepods (Machida & Lin, 2017). Consequently, standard methods like MACSE (Ranwez, Harispe, Delsuc, & Douzery, 2011) only scan for frameshift and/or stop codons caused by indels' presence in detecting pseudogenes, which may be insufficient (Leray & Knowlton, 2015). Overall, this study's findings demonstrate the importance of careful interpretation of amplicon sequences, especially those from gDNA.

425 Moreover, amplification bias in PCR-based methods was another source of taxonomic
426 bias in diversity estimation. PCR amplification bias commonly happens mainly due to
427 variable primer-template mismatches in selected species (Piñol, Mir, Gomez−Polo, & Agustí,
428 2014). This explains the case of *Mongolodiaptomus birulai* in our mock community samples
429 (Figures 2B and S2) that is not detected in either PCR-based method. *Mongolodiaptoms*
430 *birulai* is the most abundant species in the Fei Tsui Reservoir; thus, failure to detect dominant
431 species among the samples can lead to an altered conclusion about the zooplankton
432 community ecology in the studied system (Elbrecht & Leese, 2015; Krehenwinkel et al.,
433 2017).
434 In comparing the three molecular-based methods, metatranscriptomic transcripts
435 provided the most reliable species diversity estimates, which resembled morphological data.
436 First, the extraction of total RNA tends to remove eDNA and zooplankton gut content
437 sequence contaminants in the samples. Second, mRNA sequences' isolation avoids the effect
438 of nuclear-encoded mitochondrial pseudogenes (Collura et al., 1996). Third, its independence
439 from the marker gene's PCR amplification excludes any bias related to the target gene
440 amplification process. Last, the application of metatranscriptomic transcripts in field samples
441 demonstrated its consistency in species diversity estimation using different mitochondrial
442 markers (16S, COI, and CytB). Overall, this study shows the potential use of
443 metatranscriptomic transcript for long-term ecological monitoring of complex metazoan
444 communities like freshwater zooplankton.
445 Despite the stated advantages of using metatranscriptomics in studying complex
446 communities, it still comes with some shortfalls. First, the possible degradation of RNA if the
447 samples not preserved correctly in the field. With this, the use of RNAlater (Invitrogen, USA)
448 has been proven to prevent RNA degradation at 4 °C or even at room temperature

(Gorokhova, 2005). At the same time, checking the RIN can help to ensure the use of high-quality RNA in the study. For MPS applications, RIN values over 8 indicate nondegraded usable RNA; however, this standard is optimized for samples consisting of a single species or individual (Pérez−Portela & Riesgo, 2013). In contrast, community-based analyses of many species tend to have slightly lower RIN values without concerns over RNA degradation. For our sample, an average RIN value of 7 was observed without evidence of RNA degradation. Second, though not inherent in the metatranscriptomic approach, limited taxonomic coverage of available reference sequences in the Genbank may alter the "observed" community composition (false negative observations) (Leray, Knowlton, Ho, Nguyen, & Machida, 2019). Third, the technical limitations involved in the metatranscriptomics workflow like the use of random primers in cDNA synthesis may contribute minimal bias due to the difference in the GC contents among RNA fragments that affect the annealing and eventually its successful amplification (Frey, Bachmann, Peters, & Siffert, 2008). Last, selecting appropriate library preparations and insert sizes for sequencing must be carefully thought out to ensure a more efficient assembly of transcripts.

5 | CONCLUSIONS

Several taxonomic biases can be encountered with the use of gDNA for mtCOI metabarcoding. The presence of eDNA, amplification of putative pseudogenes, and PCR amplification bias may cause amplified errors in estimating complex metazoan communities' diversity; however, this study's results prove that these can be avoided with the use of metatranscriptomic transcripts. Aside from its capacity to provide data for documenting active biological processes using mRNA transcripts, this study shows that metatranscriptomics can also monitor community species diversity and compositional changes in a given ecological context.

6 | ACKNOWLEDGEMENTS

The first author is supported by the Taiwan International Graduate Program (TIGP) 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, 109-2611-M-001 (RJM), the Scientific Committee on Oceanic Research working group 157 (RJM), and the National Taiwan University 109L8836 (CHH). The funding agency 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, and the Fei-Tsui Reservoir Administration Bureau is deeply appreciated. The authors would also like to thank the NGS High Throughput Genomics Core at the Biodiversity Research Center, Academia Sinica, for sequencing assistance. Last, the authors would like to acknowledge Matthieu Leray for his significant contributions to improving the manuscript's early draft for this study.

7 | AUTHORS' CONTRIBUTIONS

MLDL, RJM, MS, CHH, and FKS conceived the ideas and designed the methodology; CHH and FKS provided all the means for fieldwork to collect zooplankton samples from Fei Tsui Reservoir; MLDL and YYL conducted the molecular experiments; MLDL and RJM analyzed the data; MLDL and RJM led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

8 | DATA ACCESSIBILITY

Raw sequences are accessible from the DNA Data Bank of Japan (DDBJ) under the accession number PSUB013509. The rest of the metadata for the mock community and field samples are available in the supplementary materials.

9 | 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

Bensasson, D., Zhang, D., Hartl, D. L., & Hewitt G. M. (2011). Mitochondrial pseudogenes: Evolution's misplaced witnesses. *Trends in Ecology & Evolution, 16,* 314-321.

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

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. doi: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. doi: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. doi: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. doi:10.1016/S0960-9822(02)70720-3

Cowart, D. A., Pinheiro, M., Mouchel, O., Maguer, M., Grall, J., Miné, J., & Arnaud-Haond, S. (2015). Metabarcoding is powerful yet still blind: A comparative analysis of morphological and molecular surveys of seagrass communities. *PLoS ONE, 10*(2), e0117562. doi:10.1371/journal.pone.0117562

Cristescu, M. E. (2014). From barcoding single individuals to metabarcoding biological communities: Towards an integrative approach to the study of global biodiversity. *Trends Ecology & Evolution, 29*, 566-571.

Dumont, H. J., van de Velde, I., and 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.

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. doi:10.1371/journal.pone.0130324

Frey, U., Bachmann, H., Peters, J., & Siffert, W. (2008). PCR-amplification of GC-rich regions: 'slowdown PCR.' *Nature Protocols, 3*, 1312-1317. doi:10.1038/nprot.2008.112

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. doi:10.1038/s41467-017-01312-x

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology, 29*, 644-652. doi:10.1038/nbt.1883

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.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols, 8*, 1494-1512.

Hlaing, T., Willoughby, T. L., Somboon, P., Socheat, D., Setha, T., Min, 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. doi: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

Kennedy, S. R., Prost, S., Overcast, I., Rominger, A. J., Gillespie, R. G., & Krehenwinkel, H. (2020). High-throughput sequencing for community analysis: The promise of DNA barcoding to uncover diversity, relatedness, abundances and interactions in spider communities. *Development Genes and Evolution, 230*, 185-201. doi:10.1007/s00427-020-00652-x

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

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

Leasi, F., Sevigny, J. L., Laflamme, E. M., Artois, T., Curini-Galletti, M., de Jesus Navarrete, A., …. Thomas, W. K. (2018). Biodiversity estimates and ecological interpretations of meiofaunal communities are biased by the taxonomic approach. *Communications Biology, 2018*(1), 112. doi:10.1038/s42003-018-0119-2

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, doi: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. doi: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. doi:10.1073/pnas.1911714116

Leray, M., Yang, J. Y., Meyer, C., 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. *Frontier in Zoology, 10*, 34. doi:10.1186/1742-9994-10-34

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

598 Machida, R. J., Hashiguchi, Y., Nishida, M., & Nishida, S. (2009). Zooplankton diversity
599 analysis through single-gene sequencing of a community sample. *BMC Genomics, 10*,
600 438. doi:10.1186/1471-2164-10-438
601 Machida, R. J., Leray, M., Ho, S., & Knowlton, N. (2017). Metazoan mitochondrial gene
602 sequence reference datasets for taxonomic assignment of environmental samples.
603 *Scientific Data, 4*, 170027. doi:10.1038/sdata.2017.27
604 Machida, R. J., & Lin Y. Y. (2017). Occurrence of mitochondrial CO1 pseudogenes in
605 *Neocalanus plumchrus* (Crustacea: Copepoda): Hybridization indicated by
606 recombined nuclear mitochondrial pseudogenes. *PLoS ONE, 12*(2), e0172710.
607 doi:10.1371/journal.pone.0172710
608 Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing
609 reads. *EMBnet.Journal, 17*(1). doi:10.14806/ej.17.1.200
610 Oksanen, J., Blanchet F. G, Friendly, M., Kindt, R., Legendre, P., McGlinn, D., … Wagner,
611 H. (2019). *Vegan: Community Ecology Package. R package version 2.5-6*. Retrieved
612 from https://CRAN.R-project.org/package=vega
613 Pérez−Portela, R., & Riesgo, A. (2013). Optimizing preservation protocols to extract high−
614 quality RNA from different tissues of echinoderms for next−generation sequencing.
615 *Molecular Ecology Resources, 13*, 884-889. doi:10.1111/1755-0998.12122
616 Perna, N. T., & Kocher, T. D. (1996). Mitochondrial DNA: Molecular fossils in the nucleus.
617 *Current Biology, 6*, 128-129.
618 Piñol, J., Mir, G., Gomez−Polo, P., & Agustí, N. (2014). Universal and blocking primer
619 mismatches limit the use of high−throughput DNA sequencing for the quantitative
620 metabarcoding of arthropods. *Molecular Ecology, 15*, 819-830. doi:10.1111/1755-
621 0998.12355
622 Piredda, R., Claverie, J., Decelle, J., de Vargas C., Dunthorn M., Edvardsen B., Eikrem W.,
623 … Zingone A. (2018). Diatom diversity through HTS-metabarcoding in coastal
624 European seas. *Scientific Reports, 8*, 18059. doi:10.1038/s41598-018-36345-9
625 Ranwez, V., Harispe, S., Delsuc, F., & Douzery, E. J. P. (2011). MACSE: Multiple
626 Alignment of Coding SEquences accounting for frameshifts and stop codons. *PLoS*
627 *ONE 6*(9), e22594. doi:10.1371/journal.pone.0022594
628 R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna,
629 Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-
630 project.org/.
631 Rees, H. C., Maddison, B. C., Middleditch, D. J., Patmore, J. R. M., & Gough, K. C. (2014).
632 The detection of aquatic animal species using environmental DNA – A review of
633 eDNA as a survey tool in ecology. *Journal of Applied Ecology, 51*, 1450-1459.
634 doi:10.1111/1365-2664.12306
635 Rognes. T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a Versatile
636 open source tool for metagenomics. *PeerJ, 4*, e2584 doi:10.7717/peerj.2584
637 Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-
638 Onsins, S. E., & Sánchez-Gracia, A. (2017). DnaSP 6: DNA sequence polymorphism
639 analysis of large datasets. *Molecular Biology and Evolution, 34*, 3299-3302.
640 doi:10.1093/molbev/msx248
641 Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., … Weber,
642 C.F. (2009). Introducing Mothur: Open-source, platform-independent, community-
643 supported software for describing and comparing microbial communities. *Applied and*
644 *Environmental Microbiology, 75*, 7537-7541.

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., … Ragg, T. (2006) The RIN: An RNA integrity number for assigning integrity values to RNA measurements. *BMC Molecular Biology, 7*, 3. doi: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. doi:10.1016/j.marenvres.2019.104836

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. doi:10.1073/pnas.0803076105

Sun, Y., Shi, Y. L., Wang, H., Zhang, T., Yu, L. Y., Sun, H., & Zhang, Y. Q. (2018). Diversity of bacteria and the characteristics of actinobacteria community structure in Badain Jaran Desert and Tengger Desert of China. *Frontiers in Microbiology, 9*, 1068. doi:10.3389/fmicb.2018.01068

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). New York, NY: Humana Press.

Van der Loos, L. M., & Nijland, R. (2020). Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. *Molecular Ecology, 00*, 1-19. doi: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. doi:10.1128/AEM.00062-07

Warner, J. R. (1999). The economics of ribosome biosynthesisin yeast. *Trends in Biochemistry Sciences, 24*, 437-440. doi:10.1016/S0968-0004(99)01460-7

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.

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*. doi:10.1038/s41598-017-01808-y

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. doi:10.1038/378489a

682 **TABLES**

683 **Table 1.** Summary of the mock community composition constructed in the study

Taxa	Species	Individual wet weight (μg)	Mock Communities (Number of individuals (dry weight biomass: 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)

684 *Note*. Values outside the parenthesis represent the number of individuals per species present in each mock community, while the values in parenthesis reflect the dry
685 weight biomass (μg) per species calculated using weight-length regression equation (Dumont et al., 1975).

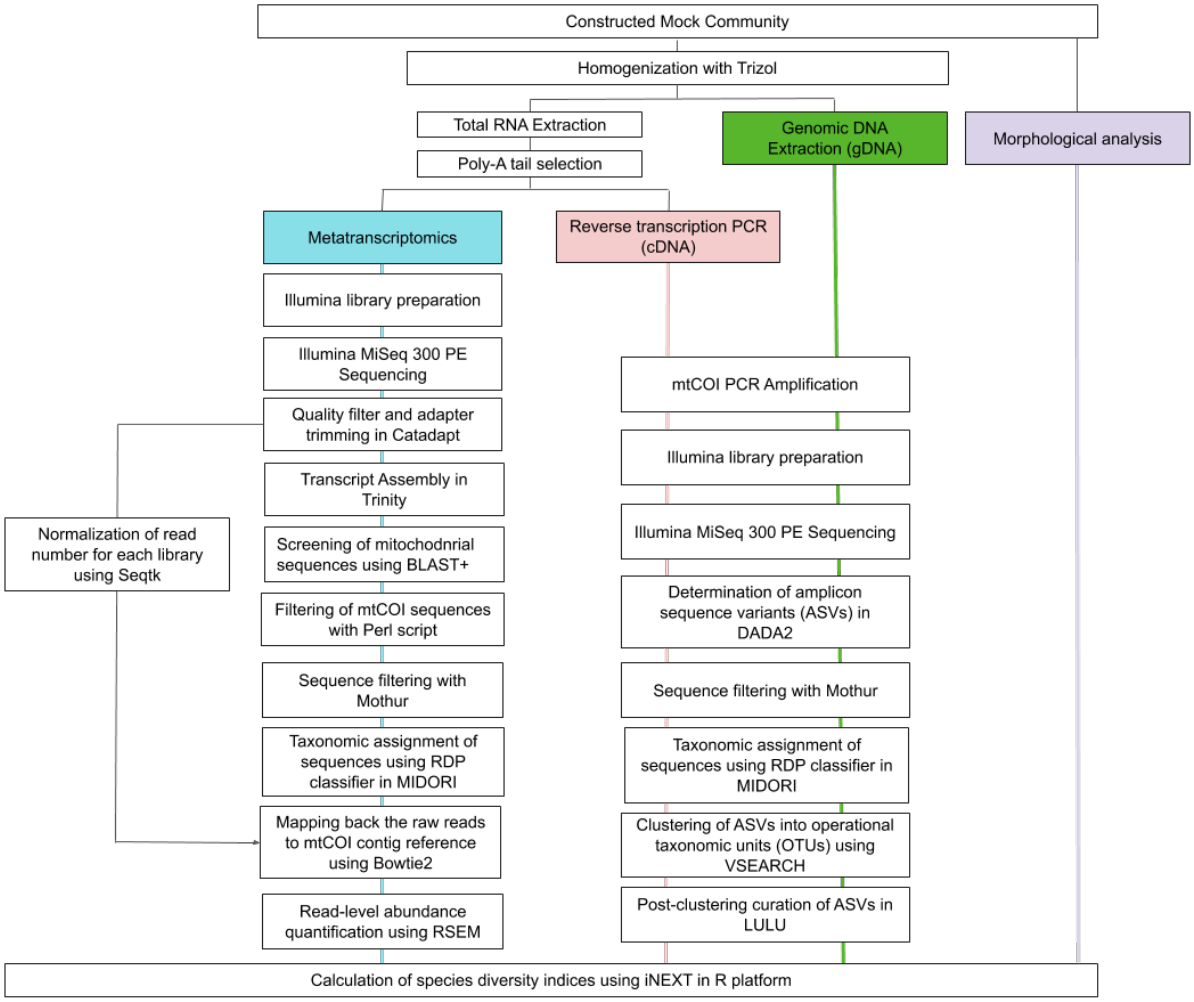
686 **Table 2.** Comparison of the number of amplicon sequence variants (ASV), nucleotide diversity, number of substitutions, and indels
687 between genomic DNA and complement DNA mtCOI amplicons for six microcrustacean species

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)	907	0.022	0.012/ 0.054	24	91	0.015	0.017/ 0.005	0
Cladocera	*Bosmina longirostris* (65)	44	0.039	0.043/ 0.022	0	19	0.011	0.014/ 0	0
	Ceriodaphnia cornuta (48)	68	0.032	0.033/ 0.024	0	52	0.012	0.012/ 0.017	0
	Daphnia galeata (68)	120	0.023	0.025/ 0.012	5	13	0.002	0/ 0.008	0
	Diaphanosoma dubium (32)	160	0.035	0.045/ 0.003	0	40	0.032	0.034/ 0.026	0
	Moina micrura (89)	459	0.027	0.022/ 0.023	6	41	0.003	0.047/ 0.008	0

688 *Note*. π: Nucleotide diversity; (N): nonsynonymous substitution; (S): synonymous substitution; and indels: insertion/deletion events. Values reflect the data from the
689 combined sequences of all five mock communities constructed in this study.

690

691 **FIGURES**



692

693 **Figure 1**. Methodology workflow of the mock community analysis.
694

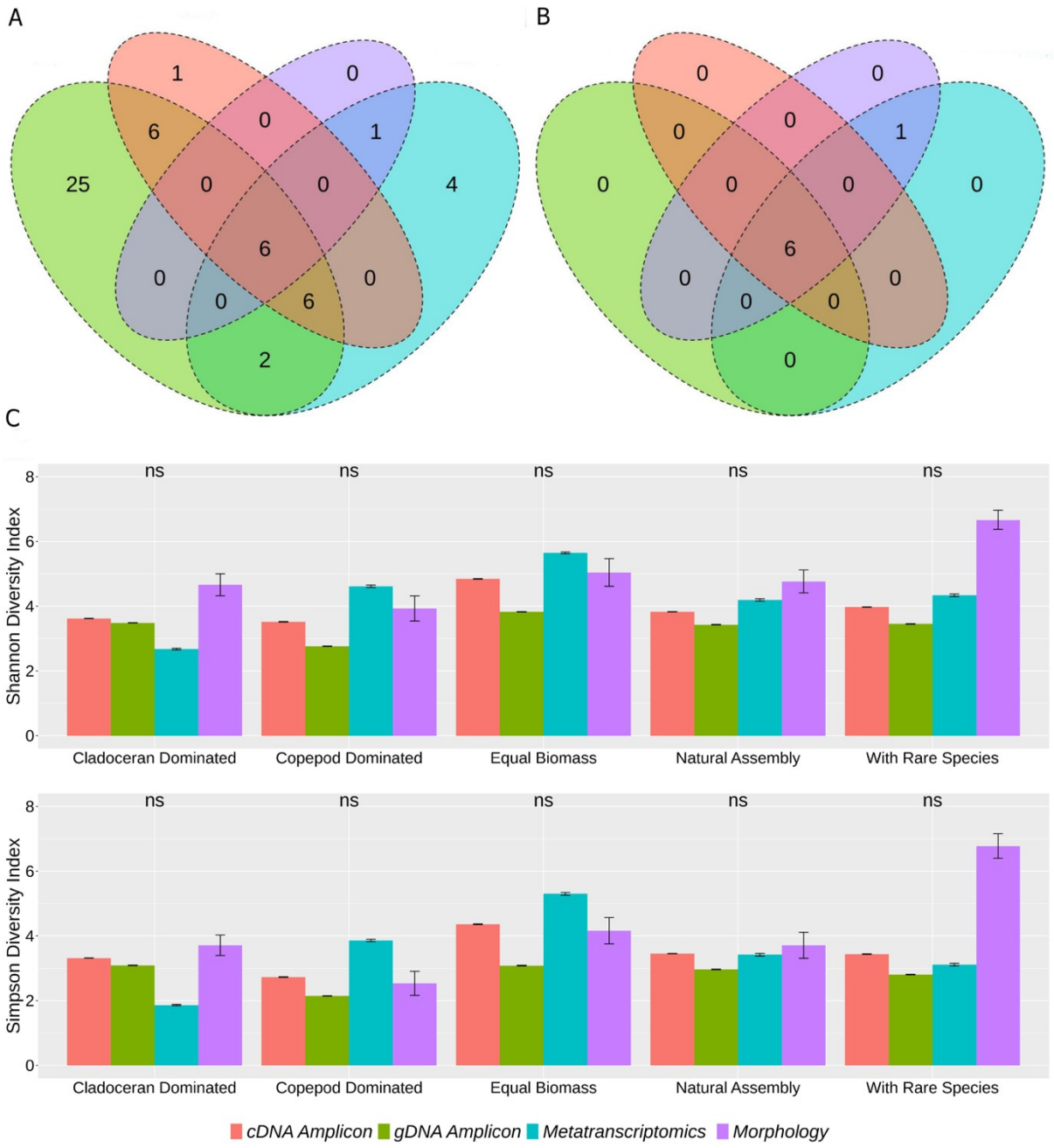


Figure 2. Comparison between the species diversity estimates from molecular-based approaches (cDNA, gDNA, and metatranscriptomic transcript) and morphological data: (A) Venn diagram showing the number of shared observed species between the methods with environmental contaminants; (B) number of shared species after extracting only target species sequences using Mothur (Schloss et al., 2009) and MIDORI dataset (Machida et al., 2017); and (C) diversity estimation using Shannon and Simpson Indices (ANOVA: 0.05 < *p*-value).

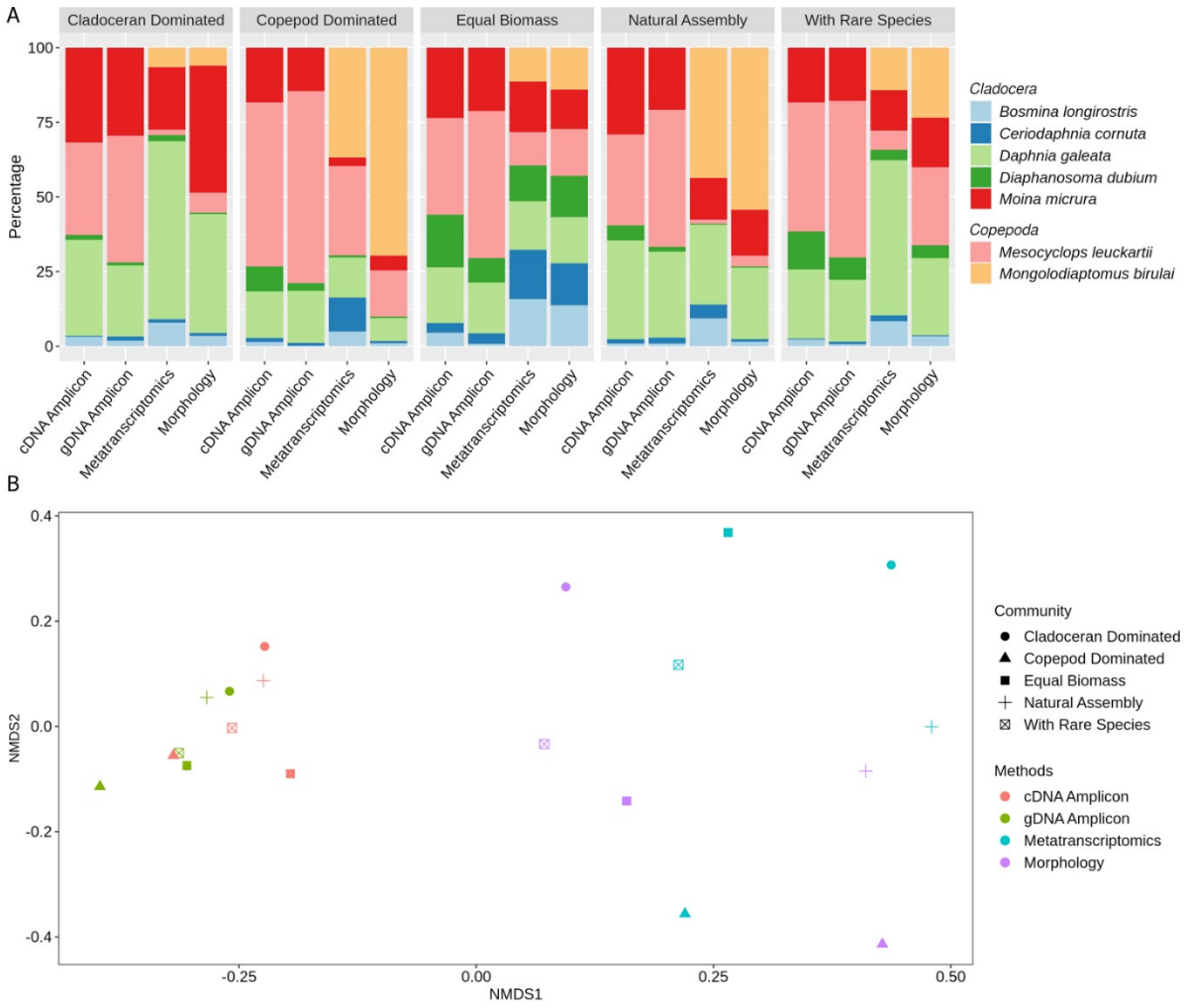


Figure 3. Comparison between the community composition of mock communities depicted by the molecular-based approaches (cDNA, gDNA, and metatranscriptomic transcript) and morphology data: (A) percentage read-level abundance (cDNA, gDNA, and metatranscriptomics) and relative dry weight biomass (morphology) of each species; and (B) NMDS plot of community composition constructed using each method (stress value = 0.1046).

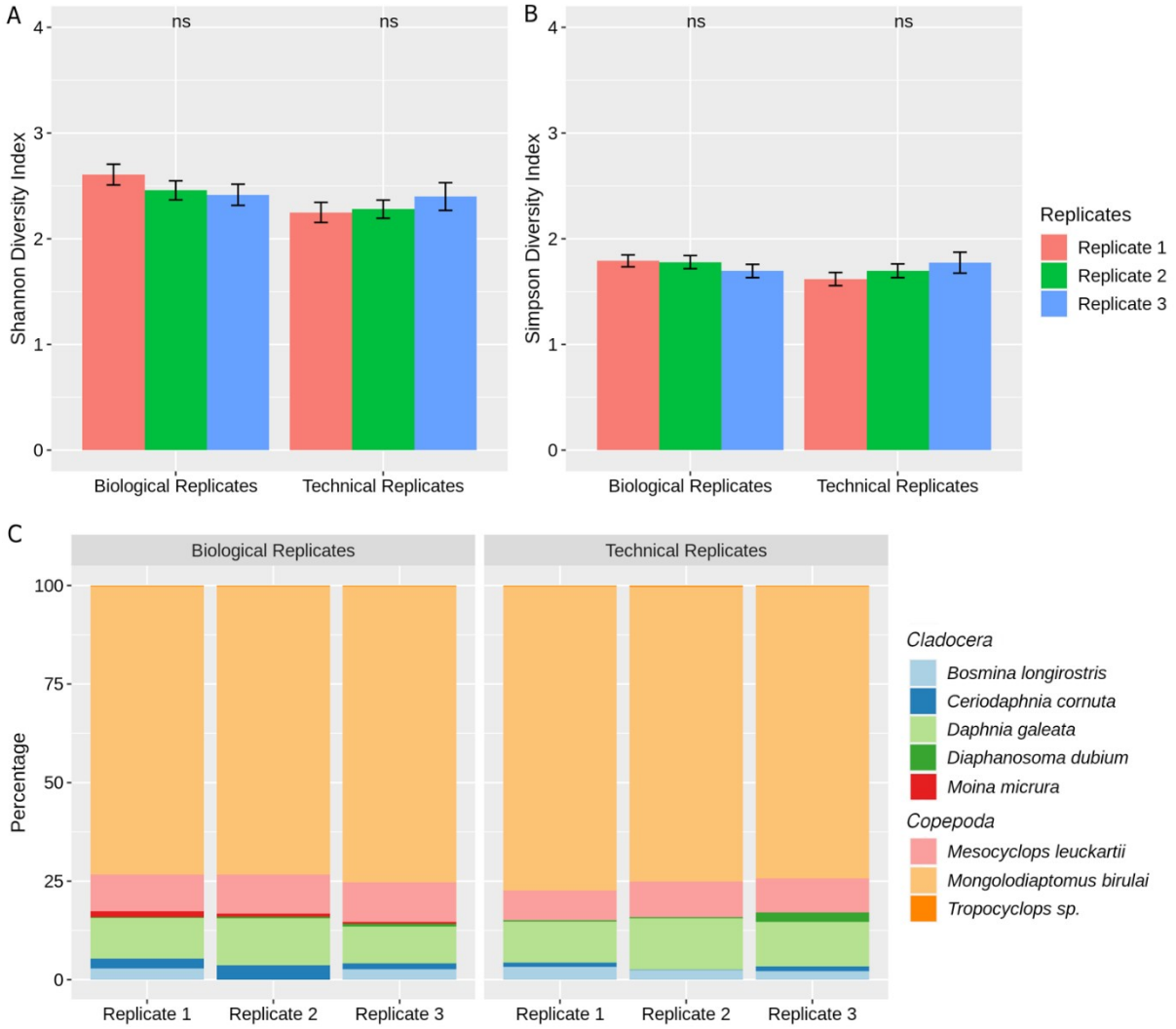
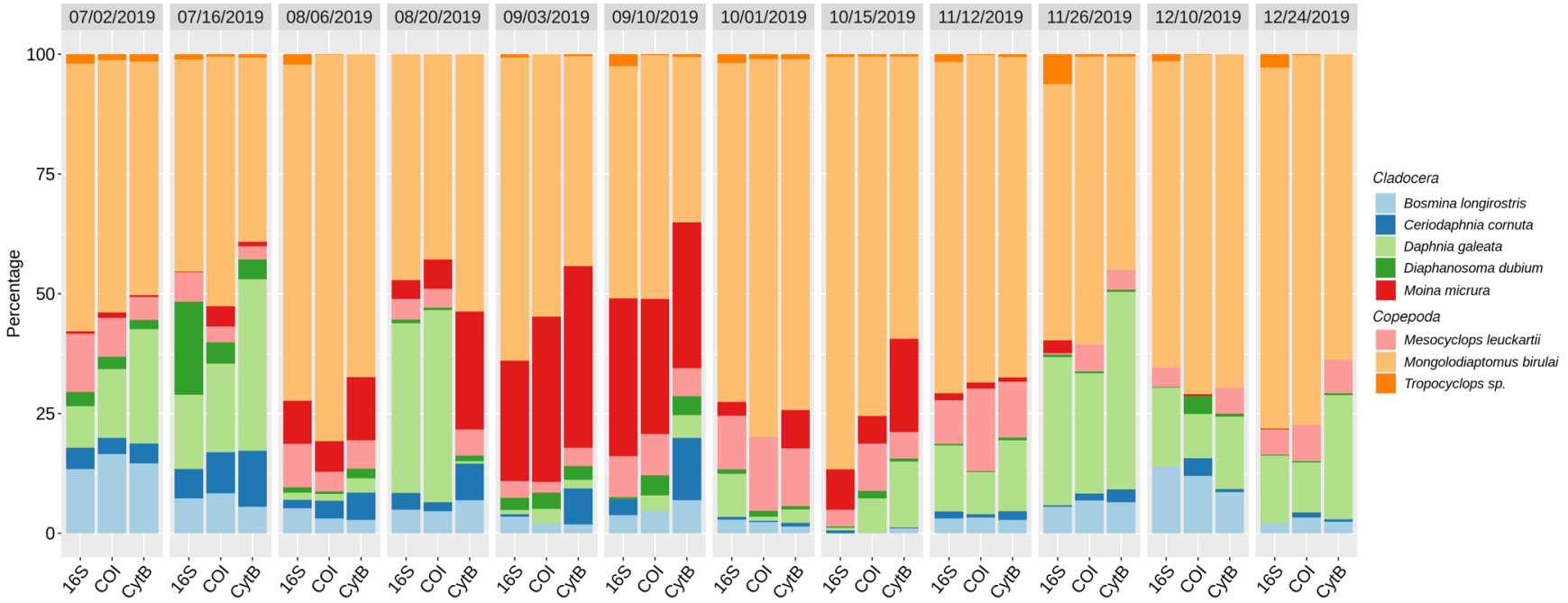


Figure 4. Comparisons of diversity indices and community composition between biological and technical replicates of microcrustacean zooplankton samples from Fei Tsui reservoir inferred from metatranscriptomic mtCOI transcripts: (A) diversity estimation using Shannon Index; (B) diversity estimation using Simpson Index (ANOVA: $0.05 < p$-value); and (C) community composition based on percentage read-level abundance per species. Biological replicate: zooplankton samples from three independent vertical plankton net tows. Technical replicate: three independent metatranscriptome sequencing libraries prepared from a single extracted zooplankton community RNA. Technical replicates were prepared from Biological Replicate 3.



722

723 **Figure 5**. Temporal community composition changes of freshwater microcrustacean zooplankton in the Fei Tsui Reservoir, constructed
724 using three mitochondrial markers (mt 16S, mtCOI, and mtCytB) from the metatranscriptomics.
725