ICES Journal of
Marine Science

ICES Journal of Marine Science (2021), 78(9), 3428–3443. https://doi.org/10.1093/icesjms/fsab084

Contribution to the Themed Section: '*Patterns of biodiversity of marine zooplankton based on molecular analysis*'

Original article

Comparative analysis of zooplankton diversities and compositions estimated from complement DNA and genomic DNA amplicons, metatranscriptomics, and morphological identifications

Ryuji J. Machida [1]*, Haruko Kurihara[2], Ryota Nakajima[3], Takashi Sakamaki[4], Ya-Ying Lin[1], and Kazushi Furusawa[5]

[1] *Biodiversity Research Center, Academia Sinica, Nankang, Taipei 11529, Taiwan*
[2] *Faculty of Science, University of the Ryukyus, Senbaru 1, Nishihara, Nakagami, Okinawa 903-0213, Japan*
[3] *Marine Plastic Research Group, Japan Agency for Marine-Earth Science and Technology, Natsushima, Yokosuka, Kanagawa 237-0061, Japan*
[4] *Graduate School of Engineering, Tohoku University, Aoba 6-6, Sendai, Miyagi 980-8579, Japan*
[5] *Marine Biological Research Institute of Japan Co., Ltd., Yutakamachi 4-3-16, Shinagawa-ku, Tokyo 142-0042, Japan*

**Corresponding author: e-mail: ryujimachida@gmail.com.*

Machida, R. J., Kurihara, H., Nakajima, R., Sakamaki, T., Lin, Y.-Y., and Furusawa, K. Comparative analysis of zooplankton diversities and compositions estimated from complement DNA and genomic DNA amplicons, metatranscriptomics, and morphological identifications. – ICES Journal of Marine Science, 78: 3428–3443.

Received 1 August 2020; revised 14 March 2021; accepted 7 April 2021; advance access publication 28 June 2021.

Community-based diversity analyses, such as metabarcoding, are increasingly popular in the field of metazoan zooplankton community ecology. However, some of the methodological uncertainties remain, such as the potential inflation of diversity estimates resulting from contamination by pseudogene sequences. Furthermore, primer affinity to specific taxonomic groups might skew community composition and structure during PCR. In this study, we estimated OTU (operational taxonomic unit) richness, Shannon's *H'*, and the phylum-level community composition of samples from a coastal zooplankton community using four approaches: complement DNA (cDNA) and genomic DNA (gDNA) mitochondrial COI (Cytochrome oxidase subunit I) gene amplicon, metatranscriptome sequencing, and morphological identification. Results of mismatch distribution demonstrated that 90% is good threshold percentage to differentiate intra- and inter-species. Moderate level of correlations appeared upon comparing the species/OTU richness estimated from the different methods. Results strongly indicated that diversity inflation occurred in the samples amplified from gDNA because of mitochondrial pseudogene contamination (overall, gDNA produced two times more richness compared with cDNA amplicons). The unique community compositions observed in the PCR-based methods indicated that taxonomic amplification bias had occurred during the PCR. Therefore, it is recommended that PCR-free approaches be used whenever resolving community structure represents an essential aspect of the analysis.

Keywords: metabarcoding, PCR amplification bias, pseudogene diversity inflation, zooplankton

Introduction

Community-based genetic analyses, such as metabarcoding, are increasingly popular as analytical methods for studying the diversity of metazoan zooplankton communities (e.g. Machida *et al.*, 2009; Lindeque *et al.*, 2013; Pearman *et al.*, 2014; Hirai *et al.*, 2015; Sommer *et al.*, 2017). Zooplankton plays an important ecological role in the marine ecosystem, transferring energy and materials to higher trophic levels, such as fishes and

Table 1. Location, date, time, depth, and filtered water volume of the zooplankton samples used in this study.

Location	GPS	Sampling date	Time	Depth (m)	Filtered water volume (m^3; (80% filtering rate assumed)	Sample weight (g)	RIN
Nishino	30°48′28.6″N 131°01′28.7″E	7 August 2012	19:50	6.7	0.65 (1.00)	0.30	7.5
Sumiyoshi	30°39′57.2″N 130°56′20.1″E	8 August 2012	20:34	8.9	0.64 (1.34)	0.44	9.4
Kurio	30°16′21.7″N 130°24′47.4″E	10 August 2012	19:50	9.5	0.90 (1.44)	0.35	9.2
Chinen	26°07′19.8″N 127°46′19.2″E	4 July 2011	22:13	10.0	1.80 (1.53)	0.45	9.6
Ginoza	26°29′18.6″N 128°00′40.1″E	6 July. 2011	21:51	7.0	1.56 (1.02)	0.40	8.1
Nakijin	26°42′27.2″N 128°01′54.7″E	12 August 2011	20:45	9.4	NA (1.42)	0.48	9.3
Haemi	24°14′47.5″N 123°53′57.5″E	15 September 2011	19:55	10.5	0.58 (1.61)	0.37	9.0
Itoman	26°09′08.1″N 127°38′09.6″E	22 August 2011	21:55	12.8	NA (2.00)	0.43	8.2
Uehara	24°28′52.0″N 123°48′12.0″E	25 September 2011	20:05	10.1	1.38 (1.54)	0.35	9.1
Funauki	24°19′53.8″N 123°44′44.9″E	26 September 2011	20:30	8.9	0.94 (1.34)	0.36	8.4

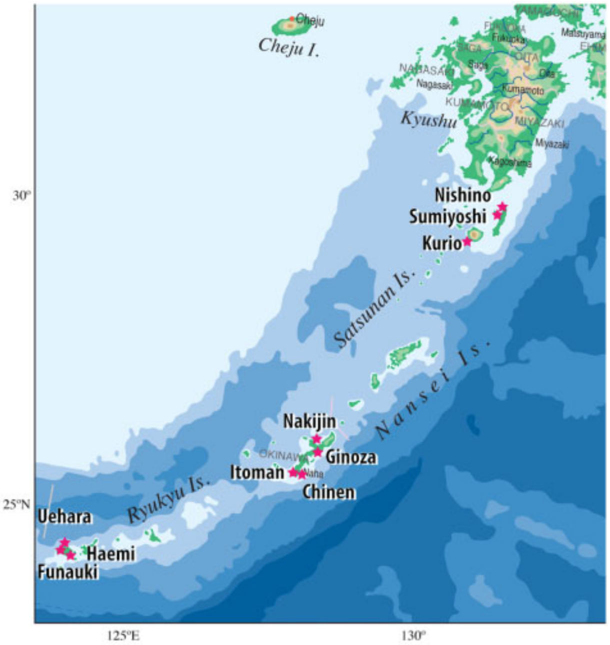

Figure 1. Sampling locations along the Ryukyu Islands.

whales (Lalli and Parsons, 1997). Therefore, careful estimation of its diversity and community composition is critically important for a better understanding of the role. There are three main reasons for the popularity of this approach for assessing diversity. First, community-based genetic analyses do not require sorting and identifying individual specimens, a laborious stem that requires a great deal of training and expert taxonomic knowledge, especially given that many of the zooplankton species are very small. Second, it is possible to assign larval forms of marine animals to taxonomic groups using genetic analyses based on the similarity between the nucleotide sequences of the subject and a reference (Machida *et al.*, 2017; Leray *et al.*, 2018), while their morphological identification is not feasible, in most cases. Third, the reason for the popularity of community-based analysis is the potential to deploy a massive parallel DNA sequencer to estimate diversity in samples that are not amenable to individual-based genetic analyses, which are for therefore more expensive.

However, there remain some uncertainties regarding the application of community-based genetic analyses to zooplankton communities. First, accidental contamination of datasets with pseudogene sequences might overestimate the diversity (Song *et al.*, 2008; Machida and Lin, 2017), the inaccurately estimated diversity figures potentially conclude wrong results in the studies comparing diversity of different communities. Nuclear-encoded mitochondrial pseudogenes vary in length, at times reaching nearly 8 000 bp (Lopez *et al.*, 1994; Richly and Leister, 2004), and have been reported in a wide variety of metazoan animals (Bensasson *et al.*, 2001; Zhang and Hewitt, 1996). Moreover, the sequence similarity of these pseudogenes to the genuine mitochondrial DNA varies widely (Zhang and Hewitt, 1996). It is in any case clear that animals with small nuclear genomes tend to possess relatively few nuclear-encoded mitochondrial pseudogenes; by contrast, the frequency of nuclear-encoded mitochondrial pseudogene sequences in animals with large genomes appears to vary considerably (Richley and Leister, 2004). Though concerns about this issue have been raised for over a quarter-century (Zhang and Hewitt, 1996), pseudogenes' specific effects on community-based genetic analyses have not yet been estimated.

Second, uncertainty regarding the application of community-based genetic analyses to zooplankton concerns amplification bias in the PCR analysis (Elbrecht and Leese, 2015). PCR-amplification bias potential skew community composition and underestimate diversity, if some species are not amplified. If the method is used to address the questions, which try to elucidate mechanisms controlling changes in community composition and diversities, then the results might not reflect actual community. The occurrence of PCR-amplification bias can be expected from individual-based analyses too. For example, we could amplify only 13 species plus one form of the mitochondrial COI (Cytochrome oxidase subunit I) gene sequences using gDNA extracted from a total of 25 species and 3 forms of oncaeid copepod individuals. In contrast, we observed a much higher amplification success rate when mitochondrial 12S rRNA (ribosomal RNA) gene primer was used (Böttger-Schnack and Machida, 2011). Previous researchers have reported numerous metazoan PCR primers, including those designed by these authors, (Machida and Knowlton, 2012; Machida *et al.*, 2012; Leray *et al.*, 2013). However, no primer set amplifies the genomes of all metazoan species, especially mitochondrial-encoded genes, due to the rarity of regions that are conserved across all metazoan groups (Machida *et al.*, 2012; Leray *et al.*, 2013). In cases in which the primers have low affinity for specific taxa, taxonomic bias develops exponentially during PCR amplifications.

For this study, we compared diversity and community composition of coastal metazoan zooplankton obtained from four

Table 2. Matrix of zooplankton species observed using morphological identification.

Phylum	Class/Subclass	Order/ Suborder	Family	Species/larvae/egg	Richness counting class	Nishino	Sumiyoshi	Kurio	Chinen	Ginoza	Nakijin	Haemi	Itoman	Uehara	Funauki
Cnidaria	Hydroidolina	Siphonophorae		Siphonophorae spp.	3		3				1				2
	Hydrozoa			Hydrozoa spp.	3			4	1		8			1	31
Nemertea	Anopla (pilidium)			Anopla (pilidium) spp.	3										1
Mollusca	Gastropoda	Pteropoda	Creseidae	*Creseis acicula*	1		6	1	1		1				
				Gastropoda (larva)	3	3	16	67	7	6	31	12	11	5	2
	Bivalvia			Bivalvia (D-shaped larva)	3			18							
				Bivalvia (umbo larva)	3	1	1	4	1	1	3	1			
Annelida	Polychaeta			Polychaeta (larva)	3	2	13		2		8	2	8	4	5
Arthropoda	Ostracoda	Halocyprida	Halocyprididae	*Conchoecia* sp.	1			12							
				Ostracoda sp.	3	23	3		5		3		4	1	
	Copepoda	Calanoida	Acartiidae	*A. bispinosa*	1										9
				Acartia erythraea	1				1	3	3				
				Acartia fossae	1				11	11	9	7	14	4	14
				Acartia sinjiensis	1					2					
				Acartia spp. (copepodite)	0			1	28	27	7	16	33	5	43
			Calanidae	*Cosmocalanus darwini*	1		3								
				Mesocalanus sp. (copepodite)	1				2						
				Nannocalanus minor	1			3							
				Neocalanus gracilis	1			1							
				Undinula vulgaris	1			4							
				Calanidae (copepodite)	0			4	2				2	1	
			Candaciidae	*Candacia catula*	1			2							
				Candacia spp. (copepodite)	0		2	4	1						
				Paracandacia truncata	1		1								
			Centropagidae	*Centropages calaninus*	1			1							
				Centropages orsinii	1				17		1		2		1
				Centropages spp. (copepodite)	0		1	4	4		6		4	1	
			Clausocalanidae	*Clausocalanus furcatus*	1		4	4		1	13	6	1	1	1
				Clausocalanus mastigophorus	1		1	3			1				
				Clausocalanus minor	1				1						
				Clausocalanus pergens	1		3	3		1	6	2		1	
				Clausocalanus spp. (copepodite)	0		7	8	2		8		2		
			Eucalanidae	*Eucalanus subtenuis*	1			2							
				Eucalanus sp. (copepodite)	0		1	1	1						
			Euchaetidae	Euchaetidae (copepodite)	0		3	4							
			Paracalanidae	*Acrocalanus gibber*	1			1							
				Acrocalanus gracilis	1			2			1				
				Acrocalanus longicornis	1			1							
				Acrocalanus spp. (copepodite)	0		1	3	2	1	5				1
				Bestiolina similis	1		2	2	134	3	12	164	51	22	117
				Calocalanus pavo	1		3	3	2			1			
				Calocalanus pavoninus	1		5			1				1	
				Calocalanus plumulosus	1		4		1				2		1
				Calocalanus spp. (copepodite)	0		6	1	1	4	3	1	1	1	2
				Delius nudus	1		6	4	4	21	1		2	3	
				Mecynocera clausi	1	1	2	1			1				

Continued

Table 2. continued

Phylum	Class/Subclass	Order/ Suborder	Family	Species/larvae/egg	Richness counting class	Nishino	Sumiyoshi	Kurio	Chinen	Ginoza	Nakijin	Haemi	Itoman	Uehara	Funauki
				Mecynocera clausi (copepodite)	0		1								
				Paracalanus parvus s.l.	1		1	9	9	3			2		
				Paracalanidae (copepodite)	0		12	11	37	1	13	32	26	8	88
			Metridinidae	*Pleuromamma* sp. (copepodite)	1									1	
			Pontellidae	*Calanopia minor*	1			1							
				Labidocera pavo	1									5	
				Labidocera spp. (copepodite)	0						27	5	1	2	
				Pontellidae (copepodite)	0					1					7
			Pseudocyclopiidae	*Pseudocyclopia* sp.	1				1						
			Pseudodiaptomidae	*Pseudodiaptomus* sp.	1								5		
				Pseudodiaptomus spp. (copepodite)	0								4		
			Temoridae	*Temora discaudata*	1		1	1							
				Temora turbinata	1			1							
				Temora sp. (copepodite)	0			2	2						
			Tortanidae	*Tortanus gracilis*	1										1
				Tortanus sp. (copepodite)	0										2
		Harpacticoida	Ectinosomatidae	*Microsetella norvegica*	1		1								
				Microsetella rosea	1				1				1		
				Microsetella sp. (copepodite)	0		1								
			Miraciidae	*Macrosetella gracilis*	1			1			1				
			Peltidiidae	*Peltidium* sp.	1	2									
			Tachidiiae	*Euterpina acutifrons*	1						1				
			Tegastidae	Tegastidae spp.	3	2				2			1		
				Harpacticoida spp.	3	29	2	9		2	13		6		1
		Cyclopoida	Corycaeidae	*Corycaeus affinis*	1										1
				Corycaeus agilis	1			47							
				Corycaeus sp. (cf.pumilus)	1			4							
				Corycaeus speciosus	1		1						1		
				Farranula gibbula	1		6	15	3		8	3	8	6	4
				Corycaeidae (copepodite)	0		1	53	3	2	9	1		1	1
			Oithonidae	*Oithona dissimilis*	1			3	7	9	4	1	37	15	2
				Oithona nana	1					9	1				
				Oithona oculata	1	2				2			3		
				Oithona plumifera	1	2	3	8	1	1		2		1	1
				Oithona simplex	1					2					
				Oithona spp. (copepodite)	0		7	4	4	7	8	4	9	4	2
			Oncaeidae	*Oncaea media*	1		39	16	4		2	1	7	1	
				Oncaea venusta	1		21	12			1	5			
				Triconia conifera	1		1	1							
				Oncaeidae (copepodite)	0		1	5	4		3	4		1	
			Clausidiidae	*Hemicyclops* sp.	1							1			
		Monstrilloida	Monstrillidae	*Monstrilloidae* sp.	3						1				
				Copepoda (nauplius)	0	2	4	1	3		13	2	8	2	
	Thecostraca			Cirripedia (cypris)	0	2							2		
				Cirripedia (nauplius)	3		1	5	13	27	3	2	1	2	
	Malacostraca	Stomatopoda		Stomatopoda (alima)	3					1	1				
		Mysida	Mysidae	*Pseudanchialina inermis*	1									1	

Continued

Table 2. continued

Phylum	Class/Subclass	Order/ Suborder	Family	Species/larvae/egg	Richness counting class	Nishino	Sumiyoshi	Kurio	Chinen	Ginoza	Nakijin	Haemi	Itoman	Uehara	Funauki
				Neomysis sp.	1	3									
				Anisomysis sp.	1	1									
				Mysidacea	3	46	9		2				1	6	1
		Cumacea		*Cumacea* sp.	3	37	3					1	1		
		Tanaidacea		Tanaidacea	3	1		1							
		Amphipoda	Synopiidae	*Synopia* sp.	1	14		3	1				1		
			Gammaridea	Gammaridea sp.	3	9	1		1					1	
		Amphipoda	Caprellidae	Caprellidae sp.	3	1									
		Euphausiacea		Euphausiacea (juvenile)	0				1						
				Euphausiacea (calyptopis)	3				5	1			1		
		Decapoda	Luciferidae	*Lucifer* sp. (mysis)	0						2				
				Lucifer sp. (zoea)	1						2		1		
				Macrura (mysis)	0	1									
				Macrura (zoea)	3	7	5	33	17	1	2	1	6	1	4
				Anomura (glaucothoe)	0								4	1	
				Anomura (zoea)	3	2	4	17	12		2	1		3	5
				Brachyura (megalopa)	0								9		1
				Brachyura (zoea)	3	8	2	26	1	1	4	2	7		2
Chaetognatha	Sagittoidea	Phragmophora	Spadellidae	*Spadella cephaloptera*	1						1				
			Sagittidae	*Sagitta enflata*	1			3							1
				Sagitta spp. (juvenile)	0			15	2	1	2	2			11
Echinodermata	Asteroidea			Asteroidea (bipinnaria)	3		1		2						
				Echinodermata (pluteus)	3				2					1	1
Chordata	Thaliacea	Doliolida	Doliolidae	*Doliolum* sp.	1		3	1							
		Salpida	Salpidae	*Thalia* sp.	1				1						
	Appendicularia	Copelata	Oikopleuridae	*Oikopleura* spp.	3		3	1		19	15	12	4	6	111
			Fritillariidae	*Fritillaria* spp.	3			2			1			2	
	Actinopterygii			Actinopterygii (larva)	0			5	1		6	2			1
				Actinopterygii (egg)	3		1	6	7	6	4	1	2	2	
				Total number of identified individuals	201	233	495	378	180	282	297	296	124	478	
				Estimated species richness		49	70	73	64	47	68	41	55	52	48

Refer the Material and methods for the definition of richness counting class.

Table 3. Estimated species and OTU richness/Shannon *H′* with four methods at each station.

Station	Morphology		cDNA COI		gDNA COI		Metatranscriptome COI	
	Species richness	Shannon *H′*	OTU richness	Shannon *H′*	OTU richness	Shannon *H′*	OTU richness	Shannon *H′*
Nishino	49	2.40	118.0	10.87	228.6	16.37	171.6	17.16
Sumiyoshi	70	3.40	186.5	16.05	260.0	38.76	338.4	41.68
Kurio	73	3.42	162.3	20.07	362.5	31.42	482.7	108.68
Chinen	64	2.82	175.3	29.01	336.5	33.61	361.4	36.55
Ginoza	47	2.88	119.0	16.43	237.5	22.83	308.5	26.31
Nakijin	68	3.44	104.8	9.65	184.1	23.40	243.5	37.26
Haemi	41	2.14	68.0	5.14	212.2	11.23	205.4	20.26
Itoman	55	3.09	62.9	10.37	329.2	25.74	369.3	45.53
Uehara	52	3.11	33.0	6.03	176.9	14.49	183.8	47.74
Funauki	48	2.31	58.3	6.31	98.61	15.85	175.4	12.39

methods, morphological identification, amplicon sequencing of the mitochondrial COI genes amplified from both cDNA (complement DNA) and gDNA (genomic DNA), and metatranscriptomic. The goal of the study is to discuss advantages and disadvantages of the methods when those methods are applied to zooplankton communities.

Material and methods

Zooplankton samples

Zooplankton community samples were collected from coastal areas around the Ryukyu Islands, Japan. The specific locations and sampling times are presented in Table 1 and Figure 1. A plankton net with a mouth opening of 30 cm in diameter and mesh size of 180 µm (Rigosha & Co., Ltd, Saitama, Japan) was used for sampling, with a pre-calibrated flow metre mounted within the mouth of the net (Rigosha & Co., Ltd., Saitama, Japan). All sampling was performed after sunset. The same sampling strategy was employed at each of the collection stations, and, therefore, the same effort was expended. Four vertical net samples were taken, proceeding from near the seafloor to the surface at each station, of which three were pooled for morphological identification and the fourth reserved for genetic analysis. All samples were brought to the laboratory within 2 h of collection. In the laboratory, the samples for RNA/DNA extraction were passed through Millipore filters (SO-Pak Filters 0.8 µm 47 mm, Merck Millipore Corp., MA, USA), transferred into Nunc CryoTubes (Thermo Fisher Scientific, Carlsbad, CA, USA), and then kept in liquid nitrogen until nucleotides extraction. The samples used for morphological identification were fixed in a buffered 5% formalin.

Morphological identification

The samples were divided into aliquots measuring one-eighth of the original volume using a plankton splitter. A taxonomic expert (K.F.) performed the morphological identification. Identification was performed based on Smith (1941), Dan *et al.* (1983, 1988), Koga (1984), Okiyama (1988), Nishimura (1995), Chihara and Murano (1997), Boltovskoy (1999), and Böttger-Schnack (1999). Upon counting of the morphological species richness, we created a richness counting class (Table 2). Assigning this class was necessary to compensate for species' richness, which was difficult to identify at the species level. Based on the counting class, we assigned one species, when we identified individuals at the species-level (e.g. *Acartia bispinosa*). We assigned three species, when we had difficulty identifying individuals at the species-level, of which more than two species were likely to be included (e.g. Siphonophorae spp.). We assigned one species, when we had difficulty identifying individuals at the species-level, but the specimen likely represented only one species (e.g. *Conchoecia* sp.). We assigned three species to larval individuals, which is difficult to identify to species-level (e.g. Gastropoda larvae). In contrast, we assigned zero species to larval individuals, which is difficult to identify at the species-level, but adult individuals of the taxa were already if counted [e.g. *Acartia* spp. (copepodite)].

RNA and DNA extraction

Differences in cDNA and gDNA mitochondrial COI amplicon results indicate the effect of pseudogene contamination (pseudogene sequences will be contaminated into gDNA but not cDNA because pseudogene sequences are not to be transcribed into mRNA). Differences in the results of PCR-based method (cDNA and gDNA mitochondrial COI gene amplicon) and PCR-free method (metatranscriptome sequencing) indicate the effect of PCR process on the community composition, especially PCR primer affinity difference between the taxa.

RNA extraction

RNA was extracted using TriPure Isolation reagent (Roche, Basel, Switzerland) in conjunction with the PureLink RNA Mini Kit (Thermo Fisher Scientific, Carlsbad, CA, USA). First, each frozen zooplankton sample together with the filter was carefully removed from the tube and homogenized using a mortar and pestle until reduced to a fine powder. During homogenization, the filter was removed, and liquid nitrogen was continuously poured over the sample. In total 5 ml of TriPure Isolation Reagent was prepared in a 50 ml Falcon tube along with a stir bar, and the powdered sample was then poured into the tube. After the transfer, each tube was incubated for 30 min at room temperature while stirring. The rest of the extraction procedure followed the standard manufacturer's protocol with the following modifications. All the chemicals were proportionally increased following the protocol. Two millilitre of the colourless upper aqueous phase was used for the extraction. One hundred microlitre of RNase-Free water were used for the final elution. The quality and concentration of the extracted RNA was assessed using a Bioanalyzer (Agilent Technology, CA, USA). Integrity of the extracted RNA was assessed using a Bioanalyzer (Agilent Technology, CA, USA). High integrity indexes (RIN: RNA Integrity Number) were observed for all samples used in this study (Table 1, RIN 7.5–9.6).

DNA extraction

gDNA extractions from the same community sample were performed using the DNeasy kit (Qiagen, Venlo, The Netherlands) in conjunction with Back Extraction Buffer (BEB; Thermo Fisher Scientific, 2020). After removal of the upper aqueous phase in the RNA extraction procedure, 1 250 µl of BEB were added to the lower non-aqueous phase and mixed vigorously by hand for 1 min and then incubated for 10 min at room temperature. Next, the tube was centrifuged at 12 000*g* for 15 min at 4°C after which 200 µl of the aqueous phase were transferred to a fresh 1.5-ml tube. Two hundred microlitre of buffer AL (Qiagen) were then added to the aqueous phase and mixed well, after which 200 µl of ethanol (99.8%) were added to the mixture and mixed well. The rest of the extraction procedure followed the standard manufacture's protocol. For the final extraction, 200 µl of Buffer AE were used. The extracted DNA was further purified using Agencourt AMPure XP (Beckman Coulter, CA, USA) following the manufacturer's protocol. The gDNA concentration and quality were measured using a NanoDrop 2000 (Thermo Fisher Scientific, Carlsbad, CA, USA).

PCR and library preparation for Illumina sequencing

Genomic DNA PCR

All PCR reactions were performed in triplicate and combined after the reactions. The reactions were carried out in a 50 µl reaction volume containing 10 ng of template, 5 µl of 2 SA PCR buffer, 4.0 µl of dNTP (Deoxyribonucleotide), 1.0 µl of each primer (5 µM), 1.0 µl of Advantage 2 Polymerase Mix (Takara Bio, Kyoto, Japan) and made up to a volume of 50 µl with nuclease-free water. A Veriti Thermal Cycler (Thermo Fisher Scientific) was used for the reaction. The primers used in this round of PCR were mlCOIintF: GGWACWGGWTGAA CWGTWTAYCCYCC and jgHCO2198: TAIACYTCIGGRTGI CCRAARAAYCA (Leray *et al.*, 2013). A PCR mixture without a template was also prepared as a negative control. The initial denaturation was carried out at 95°C for 10 min. To reduce misannealing of the primers, touchdown PCR was applied to the reaction, which involved denaturation at 95°C for 10 s, annealing at 62°C for 30 s, and extension at 72°C for 60 s. The temperature for the annealing was progressively decreased with each successive cycle (in increments of −1.0°C per cycle) from 62°C to 46°C for the first 16 cycles and kept constant at 46°C for the subsequent 20 cycles. After the PCR, the sample and negative amplification were confirmed by gel electrophoresis, and the size selection of the products was performed using Agencourt AMPure XP (Beckman Coulter), 0.4× (supernatant retained), and 0.5× (standard procedure) solutions. The second PCR mixture was prepared in the same manner as the first apart from different primers that incorporated sample-specific barcoding sequences for use in the reaction (Supplementary Material SA). The same amount of template (10 ng) was again used in each reaction. Once more, a PCR mixture without a template was also prepared as a negative control. The initial denaturation was carried out at 95°C for 10 min. This time, 20 cycles of standard PCR were applied for the reaction, which involved denaturation at 95°C for 10 s, annealing at 62°C for 30 s, and extension at 72°C for 60 s. After the PCR, the sample and negative amplification were confirmed by gel electrophoresis, and the size selection of the products was again performed using an Agencourt AMPure XP (Beckman Coulter) as described above. After measuring the concentrations using Qubit (Thermo Fisher Scientific), 500 ng of each of the purified samples were pooled, purified with 0.9× Agencourt AMPure XP (Beckman Coulter), eluted with 30 µl of nuclease-free water, and sequenced using an Illumina MiSeq 300 PE.

Complement DNA PCR

First, the mRNA was purified from the total RNA using a Dynabeads mRNA purification kit (Thermo Fisher Scientific) following the manufacturer's protocol. In total 10 µl of 10 mM Tris-HCL were used for the final elution, and 8 µl of the purified mRNA were recovered from the process. Next, the gDNA was digested using ezDNase Enzyme (Thermo Fisher Scientific) following the manufacturer's protocol. The reverse transcription was performed using SuperScript IV VILO Master Mix (Thermo Fisher Scientific) following the manufacture's protocol, and then the PCR amplification from the cDNA library was performed using 10 ng of the cDNA libraries prepared as described earlier. The PCR amplification, the second PCR for barcode adapter's attachment, the pooling of the samples, and the sequencing were all performed in the same manner as described for preparing the gDNA PCR amplicon sequencing library.

Metatranscriptomic library preparation

The metatranscriptomic library was prepared using a NEBNext mRNA Library Prep Reagent Set for Illumina (E6110) together with a NEBNext Poly(A) mRNA Magnetic Isolation Module (E7490) and a NEBNext® Multiplex Oligos for Illumina (New England BioLabs, MA, USA) following the manufacturer's protocol. Five microgram of the total RNA were used to start the library preparation. The final enrichment was performed for 12 cycles. After the enriched product's purification using 0.9× Agencourt AMPure XP, equal amounts of those products were pooled from all libraries and sequenced on an Illumina MiSeq with 300 cycles and paired-end reads.

Bioinformatics

Estimation of mismatch distribution from the amplicon libraries

First, the mismatch distribution was estimated for both the gDNA and cDNA amplicon library sequences. Sequences beginning with the primer and each of the barcode adapters were culled from the original FASTQ file using Unix grep command together with the "-B 1 -A 2" options and separated into each sample. Next, the primer and barcode adapter were removed, and the total length was trimmed to 200 nt using Cutadapt (Martin, 2011) with "-u 33 -u -68" options. We did not perform quality filtering at this stage to avoid creating length discrepancies between the sequences. Instead, we trimmed the sequences to 200 nt so as to use only high quality regions. One thousand sequences were selected randomly from each sequence dataset using seqtk (https://github.com/lh3/seqtk, last accessed April 23, 2021) with "sample -s100" options. Alignment of the sequences was then performed using MAFFT version 7.310 (Katoh and Standley, 2016) with the options "–globalpair –maxiterate 1000". After the alignment, the output files were imported into a Geneious R8 (Biomatters, Auckland, New Zealand), and the pairwise % identity of each sequence pair was calculated. At this point, the estimated frequencies for forward and reverse were combined into a single dataset. We further combined all of the frequency data that had been calculated independently at each station into a single

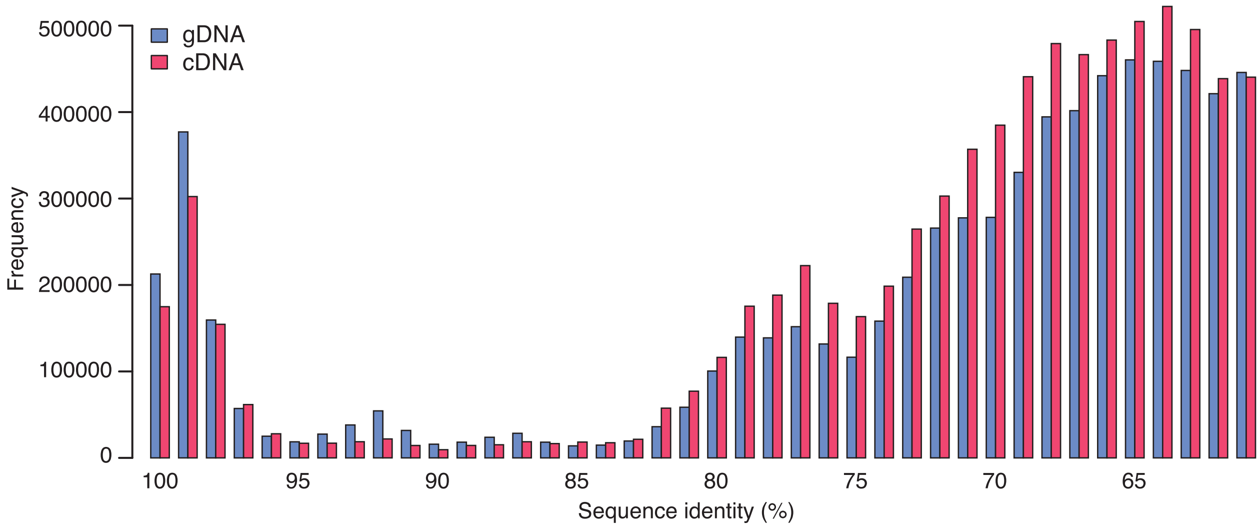

Figure 2. Mismatch distribution estimated from the gDNA and cDNA mitochondrial COI amplicon sequencings. The denoted frequencies represent the sum of the results from all of the stations. Percentage similarity, rather than genetic distance, was used for the estimation because the estimated value (90%) was used for the clustering analyses.

dataset, rather than by comparing the sequences collected from various locations to make sure sympatric mating incompatibility, by summing up values from each station to have a representative figure from all stations combined (Figure 2 and Supplementary Material SB). The mismatch distribution result indicated that 90% represent the most optimum species delineation percentage because of the lowest frequency observed at this value.

Diversity and composition estimation from the amplicon libraries

First, the original FASTQ files were separated into each sample using Cutadapt (Martin, 2011) with an option “-g file: Adapter.fasta”. Next, the primer sequences were removed, again using Cutadapt, with options “–cut 26 –minimum-length 100”. Next, DADA2 was further used for quality filtering to quality filter, merge paired reads (Callahan *et al.* 2016), and remove chimaeras (Supplementary Material SC). After the quality profile had been checked, the sequences were truncated to 250- and 200-nt lengths for read1 and read2, respectively. The sequences from forward and reverse were then merged, and chimaeras removed. Next, the sequences that passed the DADA2 quality filter were exported for further use in downstream analyses. The clustering of the sequences was performed with an identity criterion of 90% similarity (estimated from the mismatch distribution explained above) using VSEARCH (Rognes *et al.*, 2016) with the options “–cluster_fast in.fasta –strand both -id 0.90 –uc out.uc –centroids out.fasta”. After clustering, the total number of reads for each cluster was counted. The asymptotic richness was then calculated from the read counts obtained as above as input abundance data using iNEXT (Hsieh *et al.*, 2016) with the following command “iNEXT (list_data, q = 0, datatype = “abundance”)” within R Core Team (2017).

Diversity and community composition estimation from the metatranscriptome libraries

First, sequence quality filtratering and adapter removal were performed using Cutadapt (Martin, 2011: cutadapt -q 10 -a AGATCGGAAGAGC –minimum-length 100). Next, the FASTQ files were converted to FASTA files in preparation for the next BLAST step using FASTX-Toolkit (fastq_to_fasta -n -Q33). All converted FASTA files served as the query sequence for the BLAST search (Camacho *et al.*, 2009: blastn -db MIDORI_Longest -num_alignments 100 -word_size 11 -outfmt 7 -dust ‘no’ -soft_masking ‘false‘). The reference dataset for the BLAST search was created by combining all 13 protein and 2 ribosomal RNA datasets of MIDORI_LONGEST 1.1 (Machida *et al.*, 2017). Before the MIDORI datasets were combined, abbreviated names of each of the genes, such as COI, CytB (cytochrome b) and 16S (small subunit rRNA), were inserted into all of the fasta format sequences in preparation for the procedure involving the extraction of COI, CytB, and 16S gene from the BLAST results. From those results, we extracted the list of sequences that showed a high degree of similarity to mitochondrial COI, CytB, and 16S gen (Supplementary Material SD). Four criteria were used for the extraction: (i) listed results from BLAST analyses were removed when the e-value exceeded 1e-4; (ii) queries were removed when the top-listed hit was not the target gene; (iii) queries were removed when three or more genes were listed as hits; and (iv) queries were removed when only one subject remained in the BLAST result. After the lists of mitochondrial COI, CytB, and 16S sequences were created, the corresponding sequences were retrieved from the quality-filtered FASTQ datasets. In order to rarefy the sequence number between the samples, we subsampled 9 000, 3 300, and 18 000 mitochondrial COI, CytB, and 16S sequences, respectively, from datasets using Seqtk (https://github.com/lh3/seqtk: sample –s100). The target genes (mitochondrial COI, CytB, and 16S) sequences were culled with above procedure, and the contigs of each gene were then created using Mira 4 (Chevreux *et al.*, 1999). A manifest file of the Mira 4 is available as Supplementary Material SE. After the construction of the contigs, the sequence read numbers used for each contig were added to the FASTA header of contigs using the command (awk ‘FNR==NR{a[“>”$1]=“_”$4; next}{print $0a[$0]}’ info_contigstats.txt contig_fasta). Next, clustering of the contigs was performed with an identity criterion of 90% similarity (estimated from the mismatch distribution explained above) using

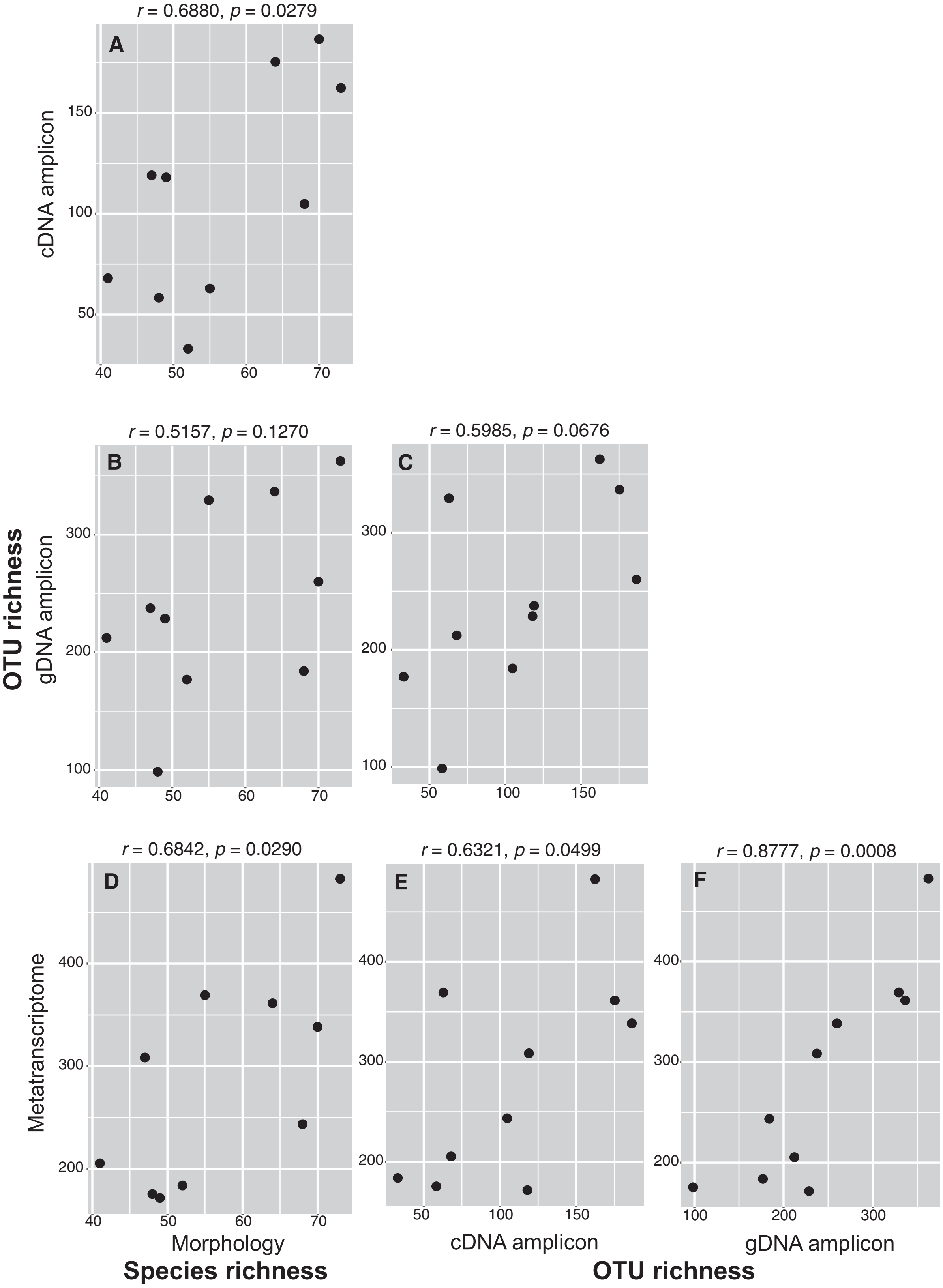

Figure 3. Relationship of the species/OTU richness estimates for the various taxonomic methods of morphology, cDNA and gDNA mitochondrial COI gene amplicon, and metatranscriptome sequencing (mitochondrial COI). The estimated correlation coefficient and associated *p* values appear above each figure.

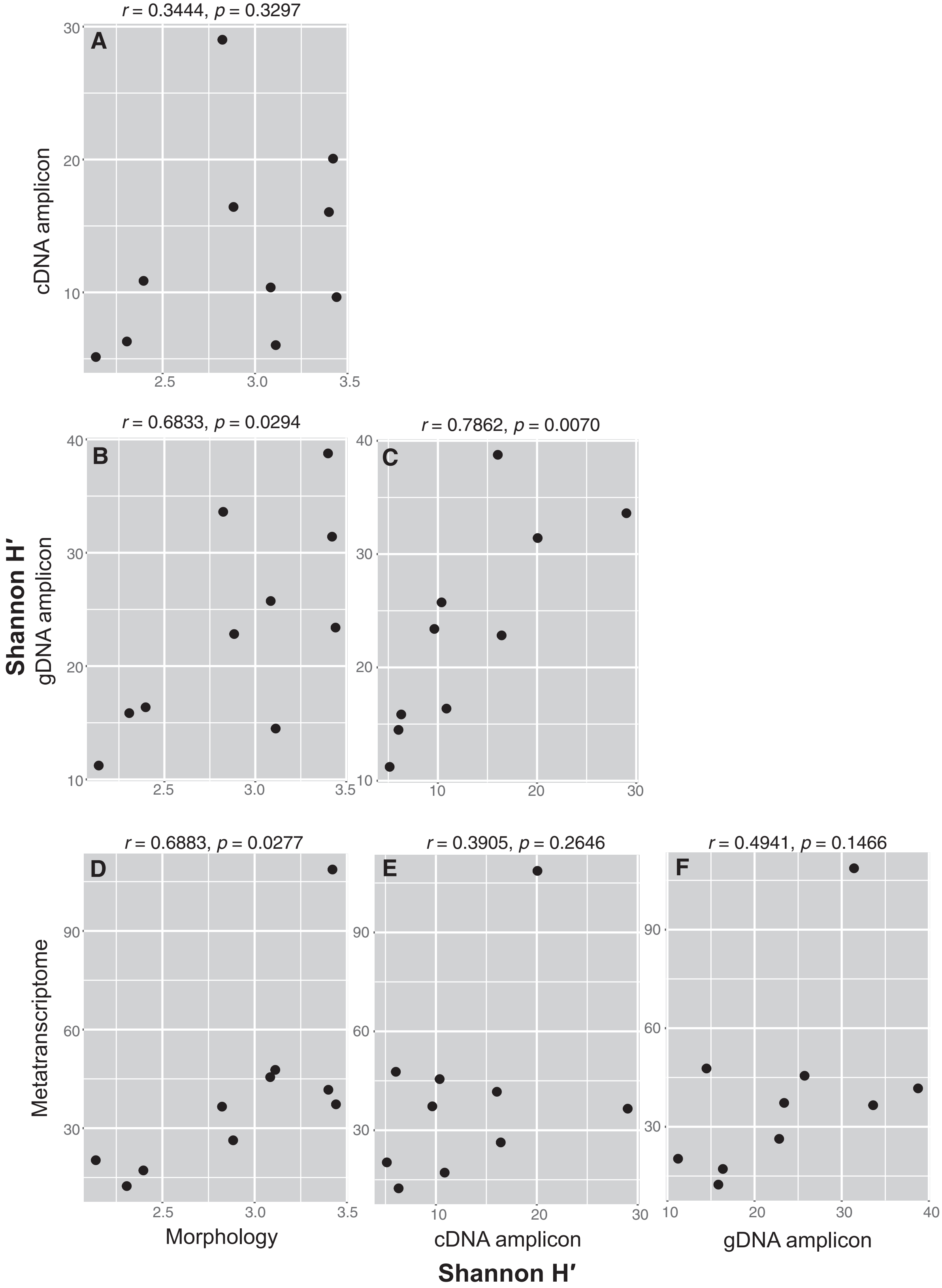

Figure 4. Relationship of the Shannon's *H′* estimates of the taxonomic methods of morphology, cDNA and gDNA mitochondrial COI gene amplification, and metatranscriptome sequencing (mitochondrial COI). The estimated correlation coefficient and associated *p* values appear above each figure.

VSEARCH (Rognes *et al.*, 2016) with the options "–cluster_fast in.fasta –strand both -id 0.90 –uc out.uc –centroids out.fasta". After clustering, the read number associated with each cluster was counted. Next, the asymptotic richness was calculated from the read counts obtained as above as input abundance data. For the calculation, we used the program iNEXT (Hsieh *et al.*, 2016) with the following command "iNEXT (list_data, q = 0, datatype = 'abundance')" within R Core Team (2017).

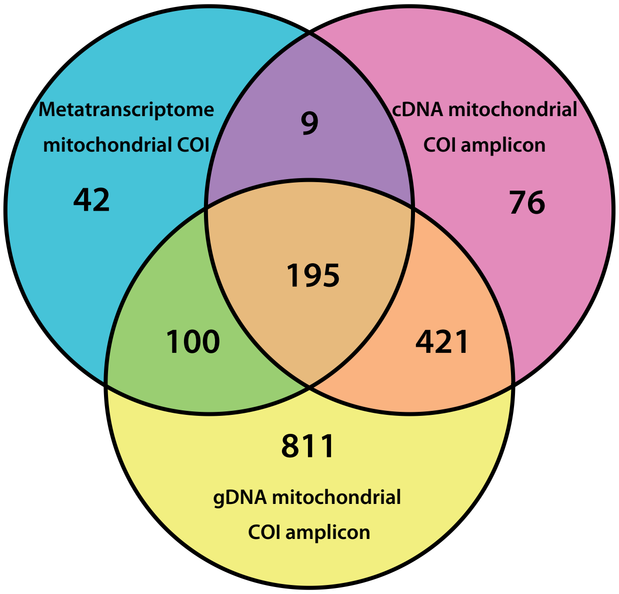
Metatranscriptome mitochondrial COI
cDNA mitochondrial COI amplicon
gDNA mitochondrial COI amplicon
42
9
76
195
100
421
811

Figure 5. Venn diagram illustrating the unique and shared mitochondrial COI clusters after the clustering analyses of centroid sequences obtained from all three molecular methods: cDNA and gDNA mitochondrial COI, and metatranscriptome analyses. In the case of metatranscriptome, only the centroids longer than 1 000 bp were used in this clustering.

Sequence taxonomic assignment

The taxonomic assignments of the contigs were performed using RDP Classifier (Wang *et al.*, 2007) and MIDORI server (Leray *et al.* 2018), with MIDORI Longest 1.1 as the reference dataset (Machida *et al.*, 2017). In this study, a confidence threshold of 50% or more at the phylum-level served as the significant cut-off.

Diversity and composition estimation from the morphological analysis

For the morphological taxonomy, identification was performed on only a portion of the total zooplankton sample because of the large sample volumes. Therefore, observed indexes were used instead of asymptotic indexes (species richness and Shannon's *H*') in the following analyses.

Statistical analyses

Analyses of correlation coefficients were performed for operational taxonomic unit (OTU) richness, and Shannon's *H*', which was estimated from the four methods (Figures 3 and 4). Clustering analyses of the mitochondrial COI gene sequences obtained from all methods (cDNA and gDNA mitochondrial COI amplicon and metatranscriptome) were performed using VSEARCH (Rognes *et al.*, 2016) with the options "–cluster_fast in.fasta –strand both -id 0.90 –uc out.uc". From the .uc files, we have depicted Venn Diagrams using the venn.diagram function in R (Figure 5; R Core Team, 2017). Centroid sequences obtained from 90% clustering analyses at each station (all three methods) were used for the input sequence in this clustering. In the case of metatranscriptome, only the centroids longer than 1 000 bp were used in clustering. This threshold makes sure that at least some portion of the contigs created from metatranscriptome overlap with the PCR amplicon region. Phylum-level community composition was estimated using all four methods (Figure 6; R Core Team, 2017). Similarity/dissimilarity among samples was evaluated using multivariate UPGMA clustering analyses, after square

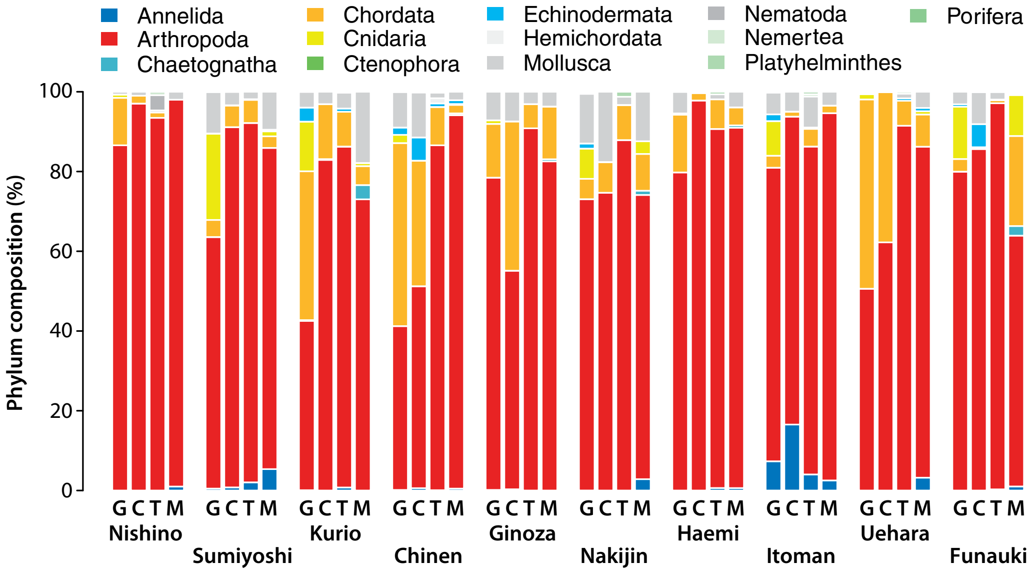

Figure 6. Phylum-level community composition estimates from the four methods, the gDNA and cDNA mitochondrial COI gene amplicon reads, metatranscriptome, and morphological identification. Abbreviations of the methods used in the figure are G, genomic DNA mitochondrial COI amplicon sequencing; C, complement DNA mitochondrial COI amplicon sequencing; T, metatranscriptome (mitochondrial COI); M, morphological identification.

root transformation of relative read abundance, using the Bray–Curtis transformation. Vegan (Oksanen *et al.*, 2019) within the R was used for the analyses (R Core Team, 2017). Analyses of the correlation coefficients between the three mitochondrial gene sequences (COI, Cyt B, and 16S rRNA) culled from the metatranscriptome data were also performed to see if the metatranscriptome analyses provide consistent diversity estimation or not (Supplementary Material SF). For these tests, data from a total of 82 samples were used.

Results

The number of identified individuals using the morphological method ranged from 124 to 495 individuals depending on the stations. The observed species richness's total number ranged from 41 at Haemi to 73 at Kurio (Table 2).

In order to estimate the boundary of within and between species sequence percent similarity, which will be used in the clustering analyses, mismatch distributions were estimated from the sequences generated from the gDNA and cDNA mitochondrial COI amplicon libraries (Figure 2, Supplementary Material SB). If we can observe non-overlapping characters between sympatric individuals, this condition most likely fits satisfactory biological species (Machida and Tsuda, 2010). We did not estimate the mismatch distribution from the metatranscriptome data because the sequences obtained were not always from homologous regions. It is also important to note that the estimated mismatch distributions were calculated independently for each sample (rather than by comparing the sequences collected from various locations) because the assumption of random mating population is required for this estimation. Further, while genetic distance is commonly used for standard mismatch distributions, we relied on percentage similarity because distributions of the estimated percentages informed the downstream analyses, with clustering defined based on percent similarity. In the mismatch distribution, the first peak was observed in the range from 100% to 97%, with the highest peak at 99% identity (Figure 2). The low frequencies continued from 96% to 83% and gradually increased after that. Although some differences were observed in a few samples (e.g. there was a large second peak in the range from 94% to 90% in the Haemi sample; Supplementary Figure SB7), the overall patterns of the results obtained from the gDNA and cDNA were very similar across all of the samples (Supplementary Material SB). The lowest frequency was found at 90% identity in both the gDNA and cDNA amplicon libraries (Figure 2). Therefore, 90% identity was used for the clustering analyses. This observation was consistent with our previous analyses using oncaeid and *Neocalanus* copepod individual (Machida and Tsuda, 2010; Böttger-Schnack and Machida, 2011).

The rarefaction curve (Supplementary Materials SI–SK) and the small values of the standard errors in the results for asymptotic OTU richness and Shannon's *H*' estimator indicated that the sampling depth was sufficient (Table 3; Supplementary Materials SG and SH). The degree of correlation observed in all of the species/OTU richness comparisons estimated from the different methods was moderate, with the correlation coefficient ranging from 0.5157 to 0.8777 (Figure 3). The correlations among the methods were weaker for Shannon's *H*' compared with species/OTU richness, with the correlation coefficients ranging from 0.3444 to 0.7862 (Figure 4). The lowest species/OTU richness was observed in morphology identification, followed by the cDNA and gDNA COI amplicons, with the metatranscriptome showing the greatest richness. The regression analyses indicated that roughly twice as much OTU richness was estimated from the gDNA COI amplicon than the cDNA COI amplicon (Table 4; Figure 3C, $Y = 1.9970X$).

In general, very high correlations were observed in all of the comparisons of OTU richness and Shannon's *H*' was estimated from the three genes (mitochondrial COI, Cyt B, and 16S rRNA) culled from the metatranscriptome analyses. The correlation coefficients ranged from 0.8238 to 0.9020 in OTU richness and from 0.70933 to 0.8785 in Shannon's *H*'. This observation demonstrated the consistency of the metatranscriptome analyses (Supplementary Material SF).

Figure 5 is a Venn diagram depicting results from the clustering of centroid sequences obtained from all three-sequencing methods: cDNA and gDNA mitochondrial COI amplicon and metatranscriptomic sequencing (in case of metatranscriptome, only centroids longer than 1 000 bp were used for the analyses). A very small proportion of non-overlapping centroids were observed from metatranscriptome (42) and cDNA mitochondrial COI amplicon (76). Those are 12% and 11% of the centroids created in each method (Figure 5). In contrast, the proportion of the non-overlapping centroids (811; 53%) for gDNA mitochondrial COI amplicon was much higher.

Community compositions at the phylum level were estimated using all methods (Figure 6). As expected, arthropods dominated in most of the samples. In other cases, non-arthropods dominated, especially when using methods involving PCR amplification, such as the gDNA COI amplicon results from Kurio, Chinen, and Uehara and the cDNA amplicon results from Chinen, Ginoza, and Uehara. In these cases, either Chordata or Cnidaria represented the major taxonomic group rather than Arthropoda. Chaetognatha was observed in the morphological identification of some samples, including Kurio, Nakijin, and Funauki. In contrast, sequences assigned to Chaetognatha were much less evident when any of the molecular methods were used (the proportion is too small to be visible in Figure 6).

Similarity-based clustering (UPGMA clustering) demonstrated the bifurcation of communities based on the methods, one of which involved PCR amplification and the other of which did not (except for one cDNA COI sample collected from Itoman; Figure 7).

Discussion

The regression analyses of the estimated species/OTU richness between the methods demonstrated that the highest estimates of species/OTU richness appeared when using the metatranscriptome sequencing, followed by gDNA and cDNA mitochondrial COI amplicons, with morphological taxonomy yielding the lowest estimates (Table 2). When compared with the metatranscriptome estimates, the gDNA and cDNA mitochondrial COI amplicons and morphological taxonomy estimated species/OTU richness to be less by 1/1.1166, 1/2.3664, and 1/5.056, respectively (Table 2). In contrast with these large discrepancies in the estimates of species/OTU richness estimates, moderate correlation levels were observed among methods (though the correlation was weaker for the Shannon's *H*' results). The results make clear, though, that the values obtained using these methods are not comparable—furthermore, each method either under- or overestimated species/OTU richness, as discussed below.

The morphological identification, performed by a trained specialist in morphological taxonomy (K.F.), yielded the lowest

Table 4. Summary of the regression analyses of the estimated species and OUT richness and Shannon's H' from the cDNA and gDNA mitochondrial COI gene amplicons, metatranscriptome sequencing, and morphological identifications.

	Independent component	Dependent component	Slope	SE	p	Adjusted R^2
Species and OTU richness						
	Morphology	Metatranscriptome	5.0560	0.4269	<0.001	0.933
	cDNA COI amplification	Metatranscriptome	2.3664	0.2759	<0.001	0.8789
	gDNA COI amplification	Metatranscriptome	1.1166	0.0625	<0.001	0.972
	cDNA COI amplification	gDNA COI amplification	1.9970	0.2405	<0.001	0.8717
Shannon H'						
	Morphology	Metatranscriptome	14.181	2.4950	<0.001	0.7578
	cDNA COI amplification	Metatranscriptome	2.5687	0.5815	0.001	0.6656
	gDNA COI amplification	Metatranscriptome	1.6598	0.3016	<0.001	0.7454
	cDNA COI amplification	gDNA COI amplification	1.6042	0.1687	<0.001	0.8994

Estimated results from metatranscriptome were used as dependent components because of the consistent results observed in comparisons of various mitochondrial genes (Supplementary Material SF). Regression analyses were also performed on the results obtained from the cDNA and gDNA mitochondrial COI gene amplicons in order to observe the effect of the pseudogenes.

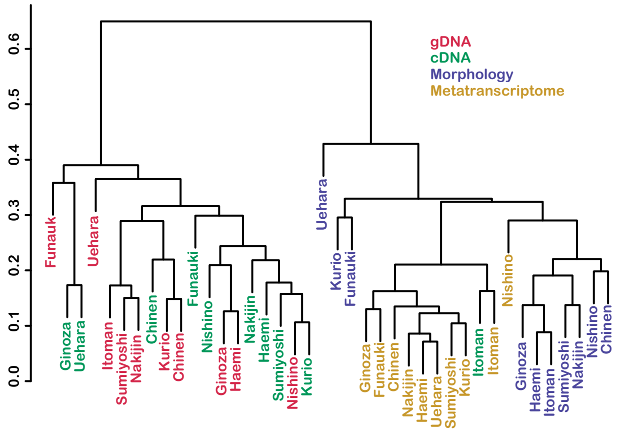

Figure 7. Dendrogram showing a relationship of hierarchical similarity among the samples. The colours used for the various methods were red: gDNA mitochondrial COI amplicon sequencing; green, cDNA mitochondrial COI amplicon sequencing; yellow, metatranscriptome (mitochondrial COI); and blue, morphological identification. The average linkage with the Bray–Curtis distance was used as the clustering algorithm.

species richness estimates among the methods tested, with a total of 89 species belonging to ten phyla in the samples. The species richness and Shannon's *H'* estimates using this method were consistently much lower than the estimates obtained using any of the molecular methods. Two factors appear to be relevant to this discrepancy. First, morphological methods cannot identify the larval forms at the species level, a fact that is especially significant here given that we collected the samples from coral reefs, where large numbers of benthic planktonic larvae are common. The second consideration is the difference in the effort required to perform the identification. Since the general pattern of species abundance follows a Whittaker plot, the number of rare species identified is proportionate to the amount of work dedicated to the identification process. If each sequence's processing can be considered a single identification, we clearly spent more effort on the molecular-based methods identification effort than on the morphology-based method. These factors appear to explain, for the most part, the discrepancies in the diversity index estimates obtained using the molecular methods compared with those obtained through morphological identification. Morphologically cryptic species may have been present in the samples, but they would only account for a small proportion of the discrepancies.

PCR-based community analyses are emerging as one standard approach for metazoan zooplankton community analyses. However, the findings presented here indicate the potential for taxonomic skew in the form of either under- or overestimation of diversity. The results of the clustering analyses revealed differences in the phylum composition of PCR-based analyses compared with other methods (Figure 7). These observations strongly indicate that PCR-based analyses suffer from both taxonomic bias and underestimation of diversity, as, indeed, previous studies have suggested (Elbrecht and Leese, 2015; Krehenwinkel *et al.*, 2017). In this study, we demonstrated PCR-free based analysis (metatranscriptomic analysis) as a method to avoid the bias during the amplification process. However, many researches are also taking multiple gene markers approach, which might abate the bias too (Stefanni *et al.*, 2018).

On the other hand, contamination with mitochondrial pseudogene sequences in gDNA COI amplicon sequencing is the most likely reason for observed diversity inflation. Researchers have reported mitochondrial pseudogenes have been reported in a large number of metazoan groups, and many pseudogenes are found in animals with large genomes (Bensasson *et al.*, 2001; Richly and Leister, 2004). The genomes of metazoans vary greatly in size; to be precise, over more than a 3 300-fold range (Gregory, 2004). Furthermore, since a complex set of factors influences genome size, it is difficult to predict and must be measured (Biémont, 2008). Given the context, it is difficult to estimate the extent to which the contamination of sequence datasets from metabarcoding analyses with mitochondrial pseudogene sequences inflates diversity. Mitochondrial pseudogenes are not to be transcribed into mRNA owing to the difference between nuclear and mitochondrial transcription promoters and genetic codes. Exploiting this feature in this study, we prepared mitochondrial pseudogene-free libraries using poly(A)+ mature mRNA as a template (cDNA) and further compared the estimated diversity indexes for the two results, one amplified from cDNA and the other amplified from gDNA. We observed roughly twice the OTU richness for gDNA compared with the cDNA mitochondrial COI amplicon libraries (Table 2). Since we extracted both the gDNA and poly(A)+ mature mRNA from the same samples, we conclude that the greater richness estimated from the gDNA resulted from the co-amplification of paralogous mitochondrial pseudogenes. Because of the small pore size filter (0.8 µm) used to trap the zooplankton samples, environmental DNA contamination might also explain the richness discrepancy between the methods. However, major proportions of biomass processed in this study are actual zooplankton samples (300–480 mg of zooplankton samples were processed for the extractions in this study; Table 1). Therefore, proportion of richness, which can be explained with environmental DNA are expected to be small. It is largely believed that mitochondrial pseudogenes have stop codons because of the insertion or deletion in the sequence. However, this condition (presence of stop codon, insertion, and deletion) is sufficient but not necessary to be pseudogenes (Perna and Kocher, 1996). Therefore, the identification of mitochondrial pseudogenes based on indels' presence has limited meaning in our study.

It is important to note that it is very likely we would observe the same or even greater diversity inflation if we applied the same cDNA and gDNA comparison to nuclear ribosomal RNA gene regions. In the case of nuclear-encoded mitochondrial pseudogenes, trans-genomic relocation is required to create the pseudogenes. However, within chromosome or within genome translocation is sufficient to create additional nuclear ribosomal RNA gene copies, and they will become pseudogenes once they do not involve in concerted evolution. This topic was not a focus of this study: however, it is important to note because many metazoan zooplankton metabarcoding studies are using nuclear ribosomal RNA gene regions as markers (e.g. 28S, 18S primarily; Lindeque *et al.*, 2013; Pearman *et al.*, 2014; Hirai *et al.*, 2015; Sommer *et al.*, 2017).

Krehenwinkel *et al.* (2017) performed arthropod community analyses using both PCR-based and PCR-free (metagenome) methods. They concluded that the latter method failed to recover a reliable target quantity, possibly due to variation in the target loci copy numbers. We also employed the PCR-free method for this study, but we used polyA(+) mRNA as the template and culled the mitochondrial transcripts informatically from the total transcripts. Therefore, copy number variation did not influence our analysis. Rather, the read numbers should reflect the active respiration potential of the species, since these are the transcribed genes responsible for mitochondria activities. Mitochondria is an organelle in cells responsible for energy (adenosine triphosphate) production through respiration. Because of the function, we expected that mitochondrial mRNA abundance in the zooplankton communities follows Kleiber's law (Kleiber, 1932; Ikeda, 1985) with higher mass-specific abundance in smaller animals.

In contrast with PCR-based sequencing methods, metatranscriptome analysis does not rely on PCR to amplify a specific gene, so no taxonomic amplification bias or diversity underestimation due to primer affinities is expected. Moreover, contamination with pseudogene sequences is also avoided because poly(A)+ mRNA serves as the starting template for the metatranscriptome method. However, it is very likely that OTU richness and Shannon's *H*' are overestimated in metatranscriptome analyses. In preparing the metatranscriptome library, the mRNA is randomly fragmented, so the sequences obtained with this method do not always derived from homologous gene regions. If the sequence coverage of a species-specific gene is low and the sequences obtained do not overlap during the assembly, these sequences are counted separately as coming from distinct OTU even when they come from the same gene of the same species. Thus, this technical issue is being responsible for the overestimating OTU richness when the metatranscriptome method is employed. We could avoid this technical issue by using long-read sequencers; however, we could not afford this approach at the time of our experiments. Sequence quality and the cost of long-read parallel sequencing are becoming more reliable and reasonable (e.g. systems from Pacific Biosciences, CA, USA, and Nanopore sequencing, Oxford, UK). Therefore, metatranscriptome methods will be promising soon if the estimation of energetically active species in the community is the focus of study.

Molecular methods also may suffer some taxonomic bias due to the limited availability of the reference sequences for specific taxa in the database. For example, Chaetognatha were observed in the morphological identification but very rarely detected using any of the molecular methods, which might be because of fewer Chaetognatha sequences in the reference dataset (29 mitochondrial COI sequences in this study). Furthermore, Chaetognatha mitochondrial genes are known to display considerable genetic diversity (Miyamoto *et al.*, 2010; Marlétaz *et al.*, 2017). For these reasons, it appears that the number of sequences identified as Chaetognatha is an underestimate. If the study's questions using the method require precise estimation of community composition, this taxonomic bias might negatively affect findings (especially if the assignment of lower taxonomic ranking was the goal). It is, then, essential to add more sequences to the database by performing individual-based analyses following reliable morphological identification.

Conclusions

In this study, we compared four methods for assessing zooplankton community diversity and composition, including the cDNA and gDNA mitochondrial COI gene amplicon sequencing, metatranscriptome sequencing, and morphological identification. We found a moderate level of correlation among the diversity index estimates using all methods. The results demonstrated that PCR-based analyses suffer from taxonomic bias. Therefore, it is recommended that the PCR-based approach not be used if the community structure is an important aspect of the study. Furthermore, roughly twice the OTU richness was observed using gDNA compared with cDNA mitochondrial amplicon libraries, demonstrating that pseudogenes likely contribute significantly to estimated diversity inflation if gDNA is used as starting template for the community-based analyses.

Availability of data and materials

All raw sequence datasets generated during this study are available from the International Nucleotide Sequence Database Collaboration with the BioProject accession number PRJDB9097.

Supplementary data

Supplementary material is available at the *ICESJMS* online version of the manuscript.

Acknowledgements

We thank Shouji Touyama, Takahiro Kobuchi, Tomoki Watanabe, Asami Takahashi, and Haruko Yamazaki for their help with the field sampling and NGS High-Throughput Genomics Core at Biodiversity Research Center, Academia Sinica, for the sequencing. We are also grateful to Matt Leray for helpful suggestions and discussion of an early draft of this article.

Funding

This project was supported by Academia Sinica, Taiwan (RM), the Ministry of Science and Technology, Taiwan 108-2611-M-001 (R.M.); 109-2611-M-001 (R.M.), and the NEXT Program (no. GR083) of the Japan Society for the Promotion of Science (T.S.). This publication resulted in part from support to SCOR WG 157 from the US National Science Foundation (Grant OCE-1840868) to the Scientific Committee on Oceanic Research (SCOR) and from funds contributed by national SCOR committees.

Authors' contributions

H.K., R.N., and T.S. organized the sampling program. R.J.M., H.K., R.N., and T.S. designed the experiments. YYL performed the molecular experiment. K.F. performed morphological identification of the zooplankton. R.J.M. analysed the data. R.J.M., H.K., R.N., K.F., and T.S. wrote the article. All of the authors read and approved the final draft of the article.

References

Bensasson, D., Zhang, D. X., Hartl, D. L., and Hewitt, G. M. 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends in Ecology and Evolution, 16: 314–321.

Biémont, C. 2008. Genome size evolution: within-species variation in genome size. Heredity, 101: 297–298.

Boltovskoy, D. 1999. South Atlantic Zooplankton. Backhuys, Leiden.

Böttger-Schnack, R. 1999. Taxonomy of Oncaeidae (Copepoda, Poecilostomatoida) from the Red Sea. I. 11 species of *Triconia* gen. nov. and a redescription of *T. similis* (Sars) comb. Nov. from Norwegian waters. Mitteilungen aus den Hamburgischen Zoologischen Museum und Institute, 96: 37–128.

Böttger-Schnack, R., and Machida, R. J. 2011. Comparison of morphological and molecular traits for species identification and taxonomic grouping of oncaeid copepods. Hydrobiologia, 666: 111–125.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. Nature Methods, 13: 581–587.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. 2009. BLAST+: architecture and applications. BMC Bioinformatics, 10: 421.

Chevreux, B., Wetter, T., and Suhai, S. 1999. Genome sequence assembly using trace signals and additional sequence information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics, 99: 45–56.

Chihara, M., and Murano, M. 1997. An Illustrated Guide to Marine Plankton in Japan. Tokai University Press, Tokyo.

Dan, K., Sekiguchi, Y., Ando, H., and Watanabe, H. 1983. Development of Invertebrates I. Baifukan, Tokyo.

Dan, K., Sekiguchi, Y., Ando, H., and Watanabe, H. 1988. Development of Invertebrates II. Baifukan, Tokyo.

Elbrecht, V., and 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.

Gregory, T. 2004. The Evolution of the Genome. Elsevier Academic Press, Oxford.

Hirai, J., Kuriyama, M., Ichikawa, T., Hidaka, K., and Tsuda, A. 2015. A metagenetic approach for revealing community structure of marine planktonic copepods. Molecular Ecology Resources, 15: 68–80.

Hsieh, T. C., Ma, K. H., and Chao, A. 2016. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). Methods in Ecology and Evolution, 7: 1451–1456.

Ikeda, T. 1985. Metabolic rates of epipelagic marine zooplankton as a function of body mass and temperature. Marine Biology, 85: 1–11.

Katoh, K., and Standley, D. M. 2016. A simple method to control over-alignment in the MAFFT multiple sequence alignment program. Bioinformatics, 32: 1933–1942.

Kleiber, M. 1932. Body size and metabolism. Hilgardia, 6: 315–351.

Koga, F. 1984. Morphology, ecology, classification, and specialization of copepods nauplius. Bulletin of the Nansei Regional Fisheries Research Laboratory, 16: 95–229.

Krehenwinkel, H., Wolf, M., Lim, J. Y., Rominger, A. J., Simison, W. B., and Gillespie, R. G. 2017. Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. Scientific Reports, 7: 17668.

Lalli, C. M., and Parsons, T. R. 1997. Biological oceanography: An Introduction, 2nd edn. Elsevier Academic Press, Oxford.

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., and Boehm, J. T. 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.

Leray, M., Ho, S. L., Lin, I. J., and Machida, R. J. 2018. MIDORI server: a webserver for taxonomic assignments of unknown metazoan mitochondrial-encoded sequences using a curated database. Bioinformatics, 34; 3753–3754.

Lindeque, P. K., Parry, H. E., Harmer, R. A., Somerfield, P. J., and Atkinson, A. 2013. Next generation sequencing reveals the hidden diversity of zooplankton assemblages. PLoS One, 8: e81327.

Lopez, V. J., Yuhki, N., Masuda, R., Modi, W., and O'Brien, S. J. 1994. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. Journal of Molecular Evolution, 39; 174–190.

Machida, R. J., and Knowlton, N. 2012. PCR primers for metazoan nuclear 18S and 28S ribosomal DNA sequences. PLoS One, 7; e46180.

Machida, R. J., and Lin, Y. Y. 2017. Occurrence of mitochondrial CO1 pseudogenes in *Neocalanus plumchrus* (Crustacea: Copepoda): hybridization indicated by recombined nuclear mitochondrial pseudogenes. PLoS One, 12; e0172710.

Machida, R. J., and Tsuda, A. 2010. Dissimilarity of species and forms of planktonic *Neocalanus* copepods using mitochondrial COI, 12S, nuclear ITS, and 28S gene sequences. PLoS One, 5; e10278.

Machida, R. J., Hashiguchi, Y., Nishida, M., and Nishida, S. 2009. Zooplankton diversity analysis through single-gene sequencing of a community sample. BMC Genomics, 10; 438.

Machida, R. J., Kweskin, M., and Knowlton, N. 2012. PCR primers for metazoan mitochondrial 12S ribosomal DNA sequences. PLoS One, 7; e35887.

Machida, R. J., Leray, M., Ho, S. L., and Knowlton, N. 2017. Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. Scientific Data, 4; 170027.

Marlétaz, F., Parco, Y. L., Liu, S., and Peijnenburg, K. T. C. A. 2017. Extreme mitogenomic variation in natural populations of Chaetognaths. Genome Biology and Evolution, 9;1374–1384.

Martin, M. 2011. Cutadapt removes adapter sequences from high--throughput sequencing reads. EMBnet.journal, 17; 10–12.

Miyamoto, H., Machida, R. J., and Nishida, S. 2010. Genetic diversity and cryptic speciation of the deep sea chaetognath *Caecosagitta macrocephala* (Fowler, 1904). Deep-Sea Research Part II: Topical Studies in Oceanography, 57: 2211–2219.

Nishimura, S. 1995. Guide to Seashore Animals of Japan With Color Picture and Keys, II. Hoikusha, Osaka.

Okiyama, M. 1988. An Atlas of the Early Stage Fishes in Japan. Tokai University Press, Tokyo.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D. *et al.* 2019. vegan: Community ecology package. R package version 2.5-5.

Pearman, J. K., El-Sherbiny, M. M., Lanzén, A., Al-Aidaroos, A. M., and Irigoien, X. 2014. Zooplankton diversity across three Red Sea reefs using pyrosequencing. Frontiers in Marine Science, 1: 27.

Perna, N. T., and Kocher, T. D. 1996. Mitochondrial DNA: molecular fossils in the nucleus. Current Biology, 6; 128–129.

R Core Team. 2017. R: A Language and Environment for Statistical Computing. R Core Team, Vienna.

Richly, E., and Leister, D. 2004. NUMTs in sequenced eukaryotic genomes. Molecular Biology and Evolution, 21: 1081–1084.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ, 4: e2584.

Smith, L. V. 1941. *Labidoceara glauca* sp. nov., a blue copepod of Puerto Galera Bay, Mindoro. The Philippine Journal of Science, 75: 307–322.

Sommer, S. A., Van Woudenberg, L., Lenz, P. H., Cepeda, G., and Goetze, E. 2017. Vertical gradients in species richness and community composition across the twilight zone in the North Pacific Subtropical Gyre. Molecular Ecology, 26: 6136–6156.

Song, H., Buhay, J. E., Whiting, M. F., and 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.

Stefanni, S., Stanković, D., Borme, D., de Olazabal, A., Juretić, T., Pallavicini, A., and Tirelli, V. 2018. Multi-marker metabarcoding approach to study mesozooplankton at basin scale. Scientific Reports, 8: 12085.

Thermo Fisher Scientific. 2020. *TRI Reagent DNA/Protein Isolation Protocol.* https://www.thermofisher.com/tw/zt/home/references/protocols/nucleic-acid-purification-and-analysis/dna-extraction-protocols/tri-reagent-dna-protein-isolation-protocol.html (last accessed 27 November 2020).

Wang, Q., Garrity, G. M., Tiedje, J. M., and 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.

Zhang, D. X., and Hewitt, G. M. 1996. Nuclear integrations: challenges for mitochondrial DNA markers. Trends in Ecology and Evolution, 11: 247–251.

Handling editor: David Fields