Improving taxonomic classification of marine zooplankton by molecular approach: registration of taxonomically verified 18S and 28S rRNA gene sequences

Tsuyoshi Watanabe [Corresp., 1], **Junya Hirai** [2], **Sirje Sildever** [3, 4], **Kazuaki Tadokoro** [5], **Kiyotaka Hidaka** [3], **Iwao Tanita** [6], **Koh Nishiuchi** [7], **Naoki Iguchi** [8], **Hiromi Kasai** [1], **Noriko Nishi** [9], **Seiji Katakura** [10], **Yukiko Taniuchi** [1], **Taketoshi Kodama** [3], **Satokuni Tashiro** [6], **Misato Nakae** [8], **Yuji Okazaki** [5], **Satoshi Kitajima** [7], **Sayaka Sogawa** [3], **Toru Hasegawa** [7], **Tomonori Azumaya** [1], **Yutaka Hiroe** [7], **Daisuke Ambe** [3], **Takashi Setou** [3], **Daiki Ito** [3], **Akira Kusaka** [3], **Takeshi Okunishi** [5], **Takahiro Tanaka** [5], **Akira Kuwata** [5], **Daisuke Hasegawa** [5], **Shigeho Kakehi** [5], **Yugo Shimizu** [3], **Satoshi Nagai** [11]

[1] Kushiro Field Station, Fisheries Resources Institute, Japan Fisheries Research and Education Agency, Kushiro, Japan

[2] Atmosphere and Ocean Research Institute, The University of Tokyo, Kashiwa, Japan

[3] Yokohama Field Station, Fisheries Resources Institute, Japan Fisheries Research and Education Agency, Yokohama, Japan

[4] Department of Marine Systems, Tallinn University of Technology, Tallinn, Estonia

[5] Shiogama Field Station, Fisheries Resources Institute, Japan Fisheries Research and Education Agency, Shiogama, Japan

[6] Yaeyama Field Station, Fisheries Technology Institute, Fisheries Research and Education Agency, Ishigaki, Japan

[7] Nagasaki Field Station, Fisheries Resources Institute, Japan Fisheries Research and Education Agenc, Nagasaki, Japan

[8] Niigata Field Station, Fisheries Resources Institute, Japan Fisheries Research and Education Agency, Niigata, Japan

[9] AXIOHELIX Co. Ltd., Tokyo, Japan

[10] City of Mombetsu, Mombetsu, Japan

[11] Yokohama Field Station, Fisheries Technology Institute, Fisheries Research and Education Agency, Yokohama, Japan

Corresponding Author: Tsuyoshi Watanabe
Email address: watanabe_tsuyoshi25@fra.go.jp

Background. Zooplankton plays an important role in the marine ecosystem. A high level of taxonomic expertise is necessary for accurate species identification based on morphological characteristics. As an alternative method to morphological classification, we focused on a molecular approach using 18S and 28S ribosomal RNA (rRNA) gene sequences. This study investigates how the accuracy of species identification by metabarcoding improves when taxonomically verified sequences of dominant zooplankton species are added to the public database. The improvement was tested by using natural zooplankton samples. **Methods.** rRNA gene sequences were obtained from dominant zooplankton species from six sea areas around Japan and registered in the public database for improving the accuracy of taxonomic classifications. Two reference databases with and without newly registered sequences were created. Comparison of detected OTUs associated with single species between the two references was done using field-collected zooplankton samples from the Sea of Okhotsk for metabarcoding analysis to verify whether or not the newly registered sequences improved the accuracy of taxonomic

classifications. **Results.** A total of 166 sequences in 96 species based on the 18S marker and 165 sequences in 95 species based on the 28S marker belonging to Arthropoda (mostly Copepoda) and Chaetognatha were registered in the public database. The newly registered sequences were mainly composed of small non-calanoid copepods, such as species belonging to *Oithona* and *Oncaea*. Based on the metabarcoding analysis of field samples, a total of 18 out of 92 OTUs were identified at the species level based on newly registered sequences in the data obtained by the 18S marker. Based on the 28S marker, 42 out of 89 OTUs were classified at the species level based on taxonomically verified sequences. Thanks to the newly registered sequences, the number of OTUs associated with a single species based on the 18S marker increased by 16% in total and by 10% per sample. Based on the 28S marker, the number of OTUs associated with a single species increased by 39% in total and by 15% per sample. The improved accuracy of species identification was confirmed by comparing different sequences obtained from the same species. The newly registered sequences had higher similarity values (mean >0.003) than the pre-existing sequences based on both rRNA genes. These OTUs were identified at the species level based on sequences not only present in the Sea of Okhotsk but also in other areas. **Discussion.** The results of the registration of new taxonomically verified sequences and the subsequent comparison of databases based on metabarcoding data of natural zooplankton samples clearly showed an increase in accuracy in species identification. Continuous registration of sequence data covering various environmental conditions is necessary for further improvement of metabarcoding analysis of zooplankton for monitoring marine ecosystems.

1 **Improving taxonomic classification of marine**
2 **zooplankton by molecular approach: registration of**
3 **taxonomically verified 18S and 28S rRNA gene**
4 **sequences**

6 Tsuyoshi Watanabe[1]*, Junya Hirai[2], Sirje Sildever[3,4], Kazuaki Tadokoro[5], Kiyotaka Hidaka[4],
7 Iwao Tanita[6], Koh Nishiuchi[7], Naoki Iguchi[8], Hiromi Kasai[1], Noriko Nishi[9], Seiji Katakura[10],
8 Yukiko Taniuchi[1], Taketoshi Kodama[4], Satokuni Tashiro[6], Misato Nakae[8], Yuji Okazaki[5],
9 Satoshi Kitajima[7], Sayaka Sogawa[4], Toru Hasegawa[7], Tomonori Azumaya[1], Yutaka Hiroe[7],
10 Daisuke Ambe[4], Takashi Setou[4], Daiki Ito[4], Akira Kusaka[4], Takeshi Okunishi[5], Takahiro
11 Tanaka[5], Akira Kuwata[5], Daisuke Hasegawa[5], Shigeho Kakehi[5], Yugo Shimizu[4], Satoshi Nagai[4]

13 [1]Kushiro Field Station, Fisheries Resources Institute, Japan Fisheries Research and Education
14 Agency, Kushiro, Hokkaido, Japan
15 [2]Atmosphere and Ocean Research Institute, The University of Tokyo, Kashiwa, Chiba, Japan
16 [3]Department of Marine Systems, Tallinn University of Technology, Tallinn, Estonia
17 [4]Yokohama Field Station, Fisheries Resources Institute, Japan Fisheries Research and Education
18 Agency, Yokohama, Kanagawa, Japan
19 [5]Shiogama Field Station, Fisheries Resources Institute, Japan Fisheries Research and Education
20 Agency, Shiogama, Miyagi, Japan
21 [6]Yaeyama Field Station, Fisheries Technology Institute, Fisheries Research and Education
22 Agency, Ishigaki, Okinawa, Japan
23 [7]Nagasaki Field Station, Fisheries Resources Institute, Japan Fisheries Research and Education
24 Agency, Nagasaki, Nagasaki, Japan
25 [8]Niigata Field Station, Fisheries Resources Institute, Japan Fisheries Research and Education
26 Agency, Niigata, Niigata, Japan
27 [9]AXIOHELIX Co. Ltd., Chiyoda-ku, Tokyo, Japan
28 [10]City of Mombetsu, Mombetsu, Hokkaido, Japan
29

31 *Corresponding Author:
32 Tsuyoshi Watanabe[1]
33 116 Katsurakoi, Kushiro, Hokkaido 085-0802, Japan
34 Email address: watanabe_tsuyoshi25@fra.go.jp

36

Abstract

Background. Zooplankton plays an important role in the marine ecosystem. A high level of taxonomic expertise is necessary for accurate species identification based on morphological characteristics. As an alternative method to morphological classification, we focused on a molecular approach using 18S and 28S ribosomal RNA (rRNA) gene sequences. This study investigates how the accuracy of species identification by metabarcoding improves when taxonomically verified sequences of dominant zooplankton species are added to the public database. The improvement was tested by using natural zooplankton samples.

Methods. rRNA gene sequences were obtained from dominant zooplankton species from six sea areas around Japan and registered in the public database for improving the accuracy of taxonomic classifications. Two reference databases with and without newly registered sequences were created. Comparison of detected OTUs associated with single species between the two references was done using field-collected zooplankton samples from the Sea of Okhotsk for metabarcoding analysis to verify whether or not the newly registered sequences improved the accuracy of taxonomic classifications.

Results. A total of 166 sequences in 96 species based on the 18S marker and 165 sequences in 95 species based on the 28S marker belonging to Arthropoda (mostly Copepoda) and Chaetognatha were registered in the public database. The newly registered sequences were mainly composed of small non-calanoid copepods, such as species belonging to *Oithona* and *Oncaea*. Based on the metabarcoding analysis of field samples, a total of 18 out of 92 OTUs were identified at the species level based on newly registered sequences in the data obtained by the 18S marker. Based on the 28S marker, 42 out of 89 OTUs were classified at the species level based on taxonomically verified sequences. Thanks to the newly registered sequences, the number of OTUs associated with a single species based on the 18S marker increased by 16% in total and by 10% per sample. Based on the 28S marker, the number of OTUs associated with a single species increased by 39% in total and by 15% per sample. The improved accuracy of species identification was confirmed by comparing different sequences obtained from the same species. The newly registered sequences had higher similarity values (mean >0.003) than the pre-existing sequences based on both rRNA genes. These OTUs were identified at the species level based on sequences not only present in the Sea of Okhotsk but also in other areas.

Discussion. The results of the registration of new taxonomically verified sequences and the subsequent comparison of databases based on metabarcoding data of natural zooplankton samples clearly showed an increase in accuracy in species identification. Continuous registration of sequence data covering various environmental conditions is necessary for further improvement of metabarcoding analysis of zooplankton for monitoring marine ecosystems.

Introduction

77 Zooplankton are diverse and dominant organisms with high abundance and biomass in the oceans,
78 linking primary producers to higher trophic levels including commercially important fishes (Ward
79 et al., 2012). Zooplankton plays an important role in biogeochemical cycles through their
80 behaviors such as grazing, respiration, excretion, and vertical migration (Steinberg et al., 2008).
81 Approximately 7,000 species belonging to 15 phyla have been described in marine metazoan
82 zooplankton (Boltovskoy, Correa & Boltovskoy, 2002), and copepods are particularly dominant
83 and diverse with >2,700 described species (Razouls et al., 2020). Zooplankton has been
84 conventionally identified based on morphological characteristics; however, species identification
85 of zooplankton needs sophisticated expertise and technique, which is labor-intensive and time-
86 consuming. In addition to the immature stages of zooplankton, there are still many additional new
87 and cryptic species in marine zooplankton, making morphological classification difficult (Bucklin
88 et al., 2010b).

89   Molecular techniques can identify zooplankton species without depending on
90 morphological characteristics, which enables an understanding of the community structure of
91 zooplankton with high taxonomic resolution (Fuller et al., 2001; Machida et al., 2009). A
92 metabarcoding approach using massively parallel next-generation sequencing (NGS) platforms is
93 an especially promising method, which can rapidly reveal large-scale patterns of community
94 structure in zooplankton based on massive sequence data (Lindeque et al., 2013; Hirai, Tachibana
95 & Tsuda, 2020). Metabarcoding is also a powerful method to recover taxonomic information on
96 damaged organisms with no clear morphological characteristics, for example, this technique has
97 been used for diet analysis of zooplanktivorous fish (e.g. Hirai et al., 2017a). Metabarcoding
98 analysis of zooplankton is thus useful to monitor marine ecosystems and understand food webs;
99 however, a reference library is required for taxonomic classifications. According to Lindeque and
100 others (2013), the most valuable reference library consists of well-populated sequences based on
101 correctly identified specimens. Continuous registration of sequence data is thus indispensable for
102 the accurate identification of thousands of zooplankton species by molecular methods.

103   The mitochondrial cytochrome *c* oxidase subunit I (COI) gene is a common genetic marker
104 for molecular analysis of marine zooplankton (Bucklin, Steinke & Blanco-Bercial, 2011).
105 However, the COI gene with high evolutionary rates is not suitable for designing universal primer
106 pairs to cover a broad range of species in zooplankton (Machida & Knowlton, 2012; Hirai,
107 Shimode & Tsuda, 2013; Clarke et al., 2017). The COI gene is commonly used for metabarcoding
108 together with other molecular markers such as nuclear small and large subunit rRNA genes (Clarke
109 et al., 2017; Carroll et al., 2019; Berry et al., 2019). Although 18S and 28S rRNA genes are
110 conserved and used for resolving deep phylogenetic relationships (Kiesling et al., 2002;
111 Sonnenberg, Nolte & Tautz, 2007; Raupach et al., 2010; Bucklin, Steinke & Blanco-Bercial, 2011;
112 Blanco-Bercial, Bradford-Grieve & Bucklin, 2011), these genes are common for metabarcoding
113 analyses due to existences of both variable and conserved regions (Bucklin et al., 2016). The 18S
114 rRNA gene is the most commonly targeted for metabarcoding analysis in eukaryotes (de Vargas
115 et al., 2015), and the number of 18S rRNA gene sequences is larger than that of 28S rRNA gene
116 sequences in the public databases (Yilmaz et al., 2014; SILVA https://www.arb-silva.de/). The

117 28S rRNA gene is more variable than the 18S rRNA gene, and it is used for species identification
118 in metazoans (Kiesling et al., 2002; Sonnenberg, Nolte & Tautz, 2007; Raupach et al., 2010;
119 Bucklin, Steinke & Blanco-Bercial, 2011; Blanco-Bercial, Bradford-Grieve & Bucklin, 2011), as
120 well as for metabarcoding of zooplankton (Harvey et al., 2018; Hirai, Tachibana & Tsuda, 2020).
121   Sequence differences within zooplankton species are commonly observed among different
122 geographical regions even in the open ocean without physical barriers (Goetze, 2003), and a
123 reference library specific to geographic regions of interest improves the accuracy of taxonomic
124 identification by metabarcoding (Questel et al., 2021). The western North Pacific around Japan is
125 adjacent to several water masses, and a high species diversity of zooplankton has been reported
126 (Tittensor et al., 2010). In the previous study, 28S rRNA gene sequences have been registered for
127 > 100 species of copepods in the Kuroshio area off Japan (Hirai et al., 2015). However, an
128 improved database with the 18S and 28S rRNA gene sequences is also needed for other areas in
129 Japanese waters, e.g. the western North Pacific and several marginal seas such as the Sea of
130 Okhotsk, the Sea of Japan and the East China Sea (Fig. 1). To cover the potential changes in
131 biodiversity, it would also be beneficial to obtain sequences in different seasons. For example,
132 based on weekly-collected zooplankton samples from the port of Mombetsu at the Sea of Okhotsk
133 in Hokkaido (northern part of Japan) analyzed using the 18S marker by metabarcoding,
134 unclassified sequences were commonly observed throughout the year (Hirai et al., 2017b). This
135 was explained by seasonal changes in zooplankton community composition, which is strongly
136 influenced by the transport by ocean currents such as the Kuroshio and Oyashio.
137   This study investigated how the accuracy of species identification by metabarcoding
138 improves by adding taxonomically verified sequences of target rRNA genes of zooplankton from
139 six sea areas around Japan to the public database. This was tested by using natural zooplankton
140 samples. The effects of seasonality and ocean currents as well as the importance of including
141 sequence data from different sea areas for improving the accuracy of zooplankton species
142 identification is also discussed.
143
144
145 **Materials & Methods**

146 ## Zooplankton samples for obtaining taxonomically verified 18S and
147 ## 28S rRNA gene sequences

148 Zooplankton samples were collected at 32 sampling stations in six areas surrounding Japan in 2018
149 (**Fig. 1; Table 1; Supplemental Table S1**). These areas included the Sea of Okhotsk, the Sea of
150 Japan, the Cold Temperate Western North Pacific (CTWNP), the Warm Temperate Western North
151 Pacific (WTWNP), the East China Sea, and the Nansei Islands (**Table 1**). The bulk zooplankton
152 samples were collected by a North Pacific Standard Plankton (NORPAC) net with 100 μm mesh,
153 to recover epipelagic samples by vertical tows from 150 m. When the water depth was <150 m,
154 sampling was carried out from the seafloor to the surface (**Table 1**). Zooplankton was preserved

155   in 99.5% ethanol immediately for genetic analysis. The ethanol was replaced after 24 h of initial
156   preservation, and samples were kept at -20°C. In each sampling area, zooplankton communities
157   were analyzed based on morphological characteristics under light microscopy by the taxonomic
158   experts, and we selected a total of 20 to 30 numerically dominant species in each sampling area
159   for updating the public sequence database (**Table 1; Supplemental Table S1**). The taxonomic
160   category (kingdom, phylum, class, order, family, genus, and species) of each zooplankton species
161   followed the taxonomy in WoRMS (WoRMS Editorial Board, 2021).

162 Primer selection for targeting the 18S and 28S rRNA genes

163   The pair of universal primers to amplify the V7–9 hypervariable regions (approximately 500 bp)
164   was used for the 18S rRNA gene (Sildever et al., 2019): 18S-V7F (5′-
165   TGGAGYGATHTGTCTGGTTDATTCCG-3') and 18S-V9R (5'-
166   TCACCTACGGAWACCTTGTTACG-3′). On the other hand, we designed a new forward primer
167   for the 28S rRNA gene. For the primer design, sequence identification number information (GI)
168   of Amphipoda, Chaetognatha, Cnidaria, Copepoda, Decapoda, Euphausiacea, Mollusca,
169   Mysidacea and Ostracoda for 28S rRNA gene was downloaded from the National Center for
170   Biotechnology Information nucleotide database (NCBI nt: https://www.ncbi.nlm.nih.gov/) (n =
171   46,217). The GI data was converted to fasta format using blastdbcmd (version 2.6.0) based on the
172   NCBI nt downloaded on August 20, 2020. Overlapping GI-s were removed (n = 46,210). The
173   sequences containing D1-2 regions were extracted using a primer pair of F63 (5′-
174   GCATATCAATAAGCGGAGGAAAAG-3′) and R635 (5′-GGTCCGTGTTTCAAGACGG-3′)
175   (Kiesling et al., 2002) by *in silico* PCR with the aid of ecoPCR v0.8.0 (Bellemain et al., 2010). In
176   this analysis, three bp mismatches in each primer sequence were allowed, but no mismatch was
177   allowed in the three bases at the 3′-end. Sequences that contained poly-N longer than 4 bp in the
178   amplified region were excluded from the data (n = 1,339). The alignment was done by MAFFT
179   (version: 7.402-with-extensions) (Katoh & Standley, 2013). The nucleotide homology was
180   checked by GENETYX version 15 (Genetyx, Tokyo, Japan) to find conservative regions to design
181   a new forward primer. The forward primer of Zoop_28S_D2F (5′-
182   GAGAGTTCAAVAGTACGTGAA-3′) was determined by considering the sequence variability
183   and the length of the amplicon (< 500 bp), in combination with the reverse primer of R635 to
184   amplify the D2 region in 28S rRNA gene (Kiesling et al., 2002).

185 Sequencing and registration of the 18S and 28S rRNA genes

186   Genomic DNAs were extracted from an individual of morphologically identified species
187   (**Supplemental Table S1**) using the QuickGene-810 (Fujifilm, Tokyo, Japan) according to the
188   manufacturer's protocol. PCR amplification was carried out on a thermal cycler (PC-808, ASTEC,
189   Fukuoka, Japan) with a reaction mixture consisting of 1 μL template DNA, 1 μM each of 18S
190   (18S-V7F and 18S-V9R) and 28S (Zoop_28S_D2F and R635) rDNA primer sets, 0.2 mM of each
191   dNTP, 1× PCR buffer, 1.5 mM Mg^{2+}, 1U KOD-Plus-Ver.2 (TOYOBO, Osaka, Japan), and

192 DNA/RNA-free dH$_2$O bringing up the volume to 25 μL. The PCR for both markers was carried
193 out separately. The PCR cycling conditions for the 18S marker were as follows: 2 min at 94 °C,
194 30 cycles at 94 °C for 15 sec, 56 °C for 30 sec, and 68 °C for 40 sec. The same PCR conditions
195 were used for amplifying the 28S marker, except for the annealing temperature (58 °C). Sequences
196 of the target regions were obtained by the direct Sanger sequencing method using the Dynamic ET
197 terminator cycle sequencing kit (GE Healthcare, Little Chalfont, UK) and a DNA sequencer
198 (ABI3730, Applied Biosystems).
199   When sequences were not successfully obtained by the direct sequencing method, we used
200 the sub-cloning method (**Supplemental Table S1**). Because KOD-Plus-Ver.2 is a thermostable
201 polymerase containing extensive 3′ to 5′ exonuclease activity and results in PCR products with a
202 blunt end, the amplicons were further treated for adenine addition to the 3′ end per A-tailing
203 procedure following the pGEM-T® Easy Vector System technical manual. Ajani and others (2022)
204 previously described the operation. Specifically, these 3′-adenine overhang products were
205 immediately ligated to the pGEM-T® Easy Vector (Promega, Madison, WI, USA) and
206 transformed into DH5α cells (Promega, Madison, WI, USA) following the manufacturer's protocol
207 (Promega, 2010). After incubating the plates at 37 °C for colony growth, six white colonies were
208 randomly chosen from each plate. The insertion of the target region was confirmed by colony PCR,
209 and three positive PCR products in each sample were processed for the Sanger sequencing using
210 the universal primers: U19 (F): (5'-GGTTTTCCCAGTCACGACG-3') and M13 (R) (5'-
211 CAGGAAACA GCTATGAC-3'; Vassart et al., 1987) .
212   The sequences were aligned using MEGA version 10 (Kumar et al., 2018) and the
213 consensus sequences were obtained for each species. The BLAST search was performed to
214 confirm the availability of sequences of the same species on the GenBank. All newly obtained
215 sequences were deposited into the DDBJ databank (accession numbers: LC581890-LC582220).

Sampling for metabarcoding analysis based on field samples

217 The metabarcoding samples of zooplankton were collected weekly at the Okhotsk Tower in the
218 Port of Mombetsu (**Fig. 1**) facing the northeastern coastal Sea of Okhotsk. Zooplankton samples
219 were collected from the bottom (10 m depth) to the surface using NORPAC net with 335 μm mesh
220 size, preserved at -20 °C until the following experiments. We used a total of 118 samples for
221 obtaining the 18S rRNA gene sequences from 10 April 2012 to 19 February 2016 and a total of
222 194 samples for obtaining the 28S rRNA gene sequences from 10 April 2012 to 16 March 2016
223 (**Fig. 1; Table 1; Supplemental Table S2**).

DNA extraction, library preparation, and sequencing for metabarcoding

226 Genomic DNAs were extracted from bulk frozen samples using the QuickGene-810 (Fujifilm)
227 according to the manufacturer's protocol. Samples were split for DNA extraction in the case of a

228   large number of specimens. For metabarcoding analysis using the MiSeq 300PE platform
229   (Illumina, USA), we followed the workflow of "16S metagenomic sequencing library preparation:
230   preparing 16S ribosomal gene amplicons for the Illumina MiSeq system" distributed by Illumina
231   (part no. 15044223 Rev. B). A two-step PCR approach was used to construct the paired-end
232   libraries of PCR amplicons, which were flanked by primer-binding sites for sequencing, dual-
233   index (i.e., barcode) sequences, and adapter sequences for binding to the flow cells of the MiSeq
234   platform.
235   The first PCR amplified the target regions using the following primer pairs: 5′-
236   ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′ + 18S or 28S marker (forward) and 5′-
237   GTGACTGGAGTTCAGACGTGTGCTCTTCCGA TCT-3′ + 18S or 28S marker (reverse). We
238   used the same primer sequences as for obtaining the taxonomically verified sequences for 18S
239   (18S-V7F and 18S-V9R) and 28S (Zoop_28S_D2F and R635) rRNA genes. The first PCR was
240   performed using a thermal cycler (PC-808; ASTEC, Fukuoka, Japan) in a 25-μL reaction mixture
241   containing 1.0 μL template DNA (<1 ng), 0.2 mM of each dNTP, 1× PCR buffer, 1.5 mM Mg^{2+},
242   1.0 U KOD-Plus-ver.2 (Toyobo, Osaka, Japan), and 1.0 μM of each primer. The PCR cycling
243   conditions for the 18S marker were as follows: initial denaturation at 94 °C for 3 min, followed by
244   30 cycles at 94 °C for 15 sec, 56 °C for 30 sec, and 68 °C for 40 sec. PCR amplification was
245   verified by 1.5% agarose gel electrophoresis. The same PCR conditions were used for amplifying
246   the 28S marker, except for the annealing temperature (58 °C). The PCR products were purified
247   using an Agencourt AMPure XP (Beckman Coulter, Brea, CA, USA) and eluted in 25 μL TE
248   buffer according to the manufacturer's protocol.
249   After purification, the PCR products were diluted 1:5 in Milli-Q water and used as a
250   template for the second PCR. The following primers: 5′-
251   AATGATACGGCGACCACCGAGATCTACAC-8   bp   index-
252   ACACTCTTTCCCTACACGACGC-3′   (forward)   and   5′-CAAGCA
253   GAAGACGGCATACGAGAT-8 bp index-GTGACTGGAGTTCAGACGTGTG-3′ (reverse)
254   were used in the second PCR. The 8-bp segments represent dual-index sequences for recognizing
255   each sample, the 5′end-sequences are adapters that allow the final product to bind or hybridize to
256   short oligonucleotides on the surface of the Illumina flow cell, and the 3′ end-sequences are
257   priming sites for MiSeq sequencing. The second PCR was performed in the reaction mixture
258   volume of 50 μL including 2.0 μL of purified and diluted PCR product. The PCR cycling
259   conditions were as follows: initial denaturation at 94 °C for 3 min, followed by 10–12 cycles at 94
260   °C for 15 sec, 59 °C for 30 sec, and 68 °C for 40 sec. PCR amplification was again verified by
261   agarose gel electrophoresis, and the PCR products were purified using an Agencourt AMPure XP
262   (Beckman Coulter). The amplified PCR products were quantified by Qubit (Invitrogen, USA), and
263   the indexed PCR products from the second PCR were pooled in equal concentrations and stored
264   at −30 °C until used for sequencing. Sequencing runs were carried out using MiSeq Reagent Kit
265   v3 on the Illumina MiSeq platform.
266

Bioinformatics analyses and operational taxonomic unit picking

Nucleotide sequences were demultiplexed depending on the 5'-multiplex identifier (MID) tag and primer sequences using the default format in MiSeq. Trimming was performed as previously described in Nagai and ohers (2022). Specifically, the sequences containing palindrome clips longer than 30 bp and homopolymer longer than 9 bp were trimmed from the sequences at both ends. The 3' tails with an average quality score of less than 30 at the end of the last 25-bp window were also trimmed from each sequence. The 5' and 3' tails with an average quality score of less than 20 at the end of the last window were also trimmed from each sequence. Sequences longer than 300 bp were truncated to 300 bp by trimming the 3' tails. The trimmed sequences shorter than 250 bp were filtered out. The demultiplexing and trimming were performed using Trimmomatic version 0.35 (http://www.usadellab.org/cms/?page=trimmomatic). The remaining sequences were merged into paired reads using Usearch version 8.0.1517 (http://www.drive5.com/usearch/). The clustering of merged sequences was performed using the cluster-features-de-novo method included in the vsearch plugin of Qiime2 version 2020.8 (https://docs.qiime2.org/2022.11/plugins/available/). The Identity option was specified at 0.990. Erroneous and chimeric sequences were detected and removed using the cluster-features-de-novo method included in the vsearch plugin of Qiime2 version 2020.8 (https://docs.qiime2.org/2022.11/plugins/available/). An OTU-count table was made using the feature-table and filter-features methods in Qiime2 with an option of removal of singleton by "–p-m-frequency 2". Demultiplexed and filtered but untrimmed sequence data were deposited into the DDBJ Sequence Read Archive (access no. DRA010320).

Taxonomic identification of the OTUs

A subset of nucleotide databases was selected from the NCBI database based on the following conditions for taxonomic identifications of OTUs using a BLAST search. The taxonomic information containing "ribosomal", "rrna", and "rdna", were extracted, but those containing "protein", "metagenome", "uncultured" and "environmental" was excluded. The sequences of retrieved GenBank IDs from the Nucleotide database downloaded from the NCBI FTP server were extracted on 13 June 2022 and used to construct a template sequence database. The taxonomic identification of each OTU was subsequently performed by BLAST search (Cheung et al., 2010) with NCBI BLAST+ 2.10.1+ (Camacho et al., 2009) using the default parameters. The nucleotide subset described above was used for the reference database and all OTU-representative sequences as the query. The taxonomic information shown as tophit_name was obtained from the BLAST hit with the top bitscores for each query sequence, and the taxonomic information showing the same top hit similarity were then merged. Namely, when an OTU had the same bitscore and the same top hit similarity, the top hit names were enumerated accordingly. The removal of sequences containing errors was imperfect after the successive processes of metabarcoding data treatment; sequences containing different types of errors derived from the original ones remained in the

305 following analytical steps. Therefore, we recognized these sequences as artificially formed OTUs
306 showing the same top hit name but slightly different sequences, resulting in showing the different
307 top hit identities, as demonstrated in the result of pyrosequencing using only *Escherichia coli*
308 MG1655 as a reference template (Kunin et al., 2010). To avoid overestimation of the OTUs, these
309 artificial OTUs were merged into a single OTU with the greatest similarity score. Species were
310 identified based on the criterion that only when the top hit similarity was >0.98 for 18S and 28S
311 rRNA gene sequences in Blast search and the OTU showed a single top hit with a full species
312 name. When a single species was clustered together with other records identified as sp. from the
313 same genus in an OTU (multi-top hit), these OTUs were regarded as the OTUs identified at species
314 levels.

Comparison of improvement in species identification between reference databases with and without new sequences

317 The improvement in the detection power of zooplankton species was investigated based on the
318 metabarcoding data from field samples from the Sea of Okhotsk (Okhotsk Tower) (**Fig. 1; Table**
319 **1; Supplemental Table S2**). Two reference databases were used for confirming the improved
320 accuracy of species identification. The conventional (Old) database was the above template
321 sequence database from GenBank, which did not include newly registered sequences. The new
322 database (New) contained the taxonomically verified sequences and the template database.
323 BLAST searches were performed on the same metabarcoding sequences using both reference
324 databases with the BLAST search settings as described above (the top hit similarity >0.98) for
325 both target genes. To verify the intraspecific variation between the sequences, the OTUs were not
326 grouped into species. To confirm the improved accuracy of species identification, the number of
327 OTUs detected per sample was tested for a statistically significant difference between the old and
328 new databases by the Wilcoxon rank sum test. Furthermore, to clarify the difference in detection
329 power for each sample, the number of OTUs detected in the new database minus the number of
330 OTUs detected in the old database (delta OTUs) was calculated: a positive value indicates
331 improved accuracy, a negative number indicates no improvement and zero indicates no change.
332 We also tested for a statistically significant difference by the Wilcoxon test in the number of
333 species detected between seasons using the Delta OTUs. Statistical analyses were conducted using
334 R (R Core Team, 2022).
335
336 **Results**

Dominant zooplankton species used for obtaining taxonomically verified sequences

339 In total, 96 species, including two unidentified species (represented by four individuals and four
340 sequences), were selected as dominant zooplankton species from six areas for obtaining

taxonomically verified sequences of 18S and 28S rRNA genes **(Fig 2)**. From the dominant species, 94 belonged to Arthropoda and 2 to Chaetognatha. In Arthropoda, 91 species belonged to Copepoda, followed by Amphipoda (2 species), Euphausiacea (2 species), and Onychopoda (1 species) **(Fig. 2)**. Copepoda species belonged to the orders of Calanoida (16 families and 31 genera), Cyclopoida (3 families and 7 genera), and Harpacticoida (3 families and 3 genera) **(Fig. 2)**. More than half (57%) of those were registered from a single sea area, whereas 25% were registered from two sea areas (**Supplemental Table S1**). At the same time, 18% were registered from three or more sea areas. Four species were also registered from five sea areas: *Calanus sinicus*, *Microsetella norvegica*, *Oithona similis*, and *Oncaea venusta* (**Supplemental Table S1**). Detailed information on species and sampling locations for taxonomically verified 18S and 28S rRNA gene sequences is available in **Supplemental Table S1**.

Registration of the taxonomically verified sequences for the target genes

A total of 166 18S rRNA gene sequences were obtained from all 96 species, including four sequences from two unidentified species associated with *Paracalanus* sp. from the WTWNP, CTWNP, and the Nansei Islands and *Metridia* sp. from the Sea of Okhotsk (**Supplemental Table S1**). From the 162 sequences belonging to 94 morphologically identified species, 30% were already registered in the public database, *i.e.* the sequences matched with a single species by BLAST ("Pre-exist" of Total in **Fig. 3**, **Table 2**). The remaining 70 % of the sequences (associated with 66 species from 38 genera) were not available in the public database ("New" of Total in **Fig. 3**, **Table 2**). Of these, 62% had a match at genus level ("New: match in genus" of Total in **Fig. 3**, **Table 2**), and 8% had a match with the higher rank (family or more) ("New: match in higher rank" of Total in **Fig. 3**, **Table 2**). In addition, 39 species had multiple sequences from different individuals and 36 of those were common sequences within a species (except for *Mesocalanus tenuicornis*, *Oithona nana*, and *Oncaea venusta*).

Based on the 18S rRNA genes, the newly registered sequences changed regionally from 12 to 25 in each area (**Fig. 3**, **Table 2**). The number of species with unregistered sequences in the public database was higher in the southern waters, including the WTWNP (83%), the East China Sea (79%), the Sea of Japan (74%), and the Nansei Islands (69%). Relatively low proportions of species with unregistered sequences were observed in the cold waters of the Sea of Okhotsk (50%) and the CTWNT (62%).

A large number of newly registered species were observed in the genus *Oithona* (for 7 out of 13 species), *Oncaea* (for 7 out of 7 species), and *Clausocalanus* (for 8 out of 8 species) in Copepoda (**Fig. 2**). These three genera occupied 56% and 69 % of newly registered species in the WTWNP and the East China Sea, respectively (**Supplemental Table S1**). In addition to *Clausocalanus* species, one species in the genera *Ctenocalanus* ("New: match in genus") and *Microcalanus* ("New: match in higher rank"), and two species in the genus *Pseudocalanus* ("New: match in genus") in the Clausocalanidae were also previously unregistered (**Fig.2**). A relatively

379 large number of newly registered species were observed in the family Paracalanidae (6 species)
380 and Corycaeidae (3 species) (**Fig. 2**).

381 For the 28S rRNA gene, a total of 165 taxonomically verified sequences of 95 were
382 registered in the public database (**Supplemental Table S1**). The sequence of one species (*Acartia*
383 *omorii* from the CTWNP) was not obtained successfully. Therefore, 161 sequences belonging to
384 93 species, except for two unidentified species associated with *Paracalanus* sp. from the WTWNP,
385 CTWNP, and the Nansei Islands and with *Metridia* sp. from the Sea of Okhotsk with four
386 sequences, were used for further analysis (**Supplemental Table S1**). 60% of the sequences
387 associated with a single species were already in the public database, *i.e.* the sequences matched in
388 species by BLAST ("Pre-exist" of Total in **Fig. 3**, **Table 2**). Further 40% of the sequences
389 associated with a single species were not available in the public database ("New" in **Fig. 3**, **Table**
390 **2**). Of those, 30% had a match at the genus level ("New: match in genus" of Total in **Fig. 3**, **Table**
391 **2**), and 10% had a match with the higher rank (family or more) ("New: match in higher rank" of
392 Total in **Fig. 3**, **Table 2**). Furthermore, 39 species had multiple sequences from different
393 individuals, and 34 of those were common sequences within a species (except for *Clausocalanus*
394 *parapergens*, *Macrosetella gracilis*, *Neocalanus plumchrus*, *Oncaea waldemari*, and *Themisto*
395 *japonica*).

396 Similarly to the species registered by the 18S rRNA gene sequences, the highest proportion
397 of newly registered sequences for 28S was associated with a single species originated from the
398 Nansei Islands (59%), followed by the East China Sea (41%) and the WTWNP (40%) (**Fig. 3**,
399 **Table 2**). Unlike the sequences obtained by the 18S marker, the sequences obtained from the
400 WTWNP based on the 28S marker showed a relatively high proportion of species already on the
401 public database (60%). The newly registered sequences for a single species were the lowest for the
402 CTWNP (25%), whereas for the Sea of Okhotsk, 35 % of the sequences were associated with a
403 single species and for the Sea of Japan 32 % of the sequences were newly registered.

404 The genera with the highest number of newly registered sequences associated with a single
405 species belonged to *Oithona* (54%) and *Oncaea* (71%) from the order Cyclopoida (**Fig. 2**). These
406 species were mainly collected in the southern part of Japan including the Nansei Islands (three
407 *Oithona* and one *Oncaea* species), the East China Sea (one *Oithona* and three *Oncaea* species),
408 and the WTWNP (three *Oithona* and five *Oncaea* species) (**Supplemental Table S1**). These two
409 genera occupied 66% and 57 % of the newly registered sequences associated with a single species
410 for the WTWNP and the East China Sea, respectively (**Supplemental Table S1**). Most of the
411 species in the genus *Clausocalanus*, which showed high proportions of unregistered sequences
412 based on the 18S rRNA gene, had a low proportion of unregistered sequences number (12%) based
413 on the 28S rRNA gene (**Fig. 2**).

Testing the improvement of taxonomic classification based on metabarcoding data from a field sample

In the metabarcoding analysis based on the 18S marker from the Sea of Okhotsk (Okhotsk Tower), 269 OTUs of Metazoa were detected with >0.980 as a single top-hit species based on the database containing newly registered sequences. The sequence identity by BLAST hit ranged from 0.980 to 0.990 in 79 OTUs, 0.991–0.999 in 172 OTUs, and 1.000 in 18 OTUs. Because the sequences registered for the dominant zooplankton taxa were associated with Arthropoda and Chaetognatha, we focused on the 97 OTUs associated with those phyla in the metabarcoding analysis.

Based on the 18S marker, from the OTUs detected in all 117 samples 79 were detected based on the conventional (old) reference database and 92 OTUs based on the new reference database including newly registered sequences (**Table 3**). 73 OTUs were common to the old and new databases, 19 OTUs were unique OTUs for the new database, and 6 OTUs for the old database (**Table 3**). In the case of 4 out of 6 OTUs in the old database, different OTUs of the same species were detected from the new database, and 3 OTUs were replaced by newly registered sequences with improved similarity (*Acartia negligens*, *Centropages abdominalis*, and *Ditrichocorycaeus anglicus*) (see below, **Supplemental Table S3**). The new database improved the zooplankton species detection by 18%. Based on the old database, the number of OTUs per sample was on average 14.2±8.9 with a minimum of 2 and a maximum of 50 OTUs (**Fig. 4; Table 3**). Based on the new database, the average number of OTUs per sample was 15.6±10.2 with a minimum of and a maximum of 53 OTUs. Per sample, the new database increased species detection by 10%, however, there was no statistically significant difference between the identification success based on the old and new databases (**Fig. 4**). For the majority of samples, the identification success improved, when using the new database (**Fig. 5**, **Supplemental Table S2**). The delta OTUs, OTUs from the new database minus the number of OTUs from the old database, was on average 1.4±1.6 with a minimum of -2 and maximum of 8, which were above zero for the majority of samples with the exception of four samples (**Fig. 5**).

Of the 92 OTUs detected based on the new database, 19 were detected only based on the newly registered sequences (**Table 3**). Four species were detected as different OTUs based on the new and old databases, of which *Acartia negligens* and *Centropages abdominalis* were detected by both databases (**Supplemental Table S3**). The newly registered sequences had higher similarity (≥0.002) among the sequences associated with a single species compared to the existing sequences. This increased the number of occurrences and detection for some species (*C. abdominalis*), whereas for some species it remained the same (*A. negligens*). For 19 species detected in metabarcoding samples from the Sea of Okhotsk (Okhotsk Tower) based on the newly registered sequences, only the sequences for five species (*Calanus pacificus*, *Corycaeus affinis*, *Eucalanus bungii*, *Pseudocalanus newmani*, *M. norvegica*) were registered from the Sea of Okhotsk (**Supplemental Table S3**). Sequences of the 18S rRNA gene for the remaining 14 species were registered from other areas, including 13 species associated with Copepoda and one species associated with Branchiopoda.

453   In total, 160 OTUs associated with Metazoa were detected based on the 28S rRNA gene
454 metabarcoding, with >0.980 similarity to a single top-hit species in the BLAST search including
455 newly registered sequences. The sequence identities were 0.981–0.990 in 32 OTUs, 0.991–0.999
456 in 98 OTUs, and 1.000 in 30 OTUs. For further analysis, 88 OTUs in the phyla Arthropoda and
457 Chaetognatha were used.

458   Based on the 28S marker, 64 OTUs were detected using the old database, and 89 OTUs
459 using the new database (**Table 3**). The identification improved by 42% based on the new database
460 compared to the old database. 44 OTUs were common to the old and new databases, 45 OTUs
461 were unique for the new database, and 20 OTUs for the old database (**Table 3**). For 19 out of 20
462 unique OTUs in the old database, different OTUs of the same species were also detected based on
463 the new database, and all OTUs were replaced by newly registered sequences with improved
464 similarity (see below, **Supplemental Table S3**). For example, an OTU associated with
465 *Ditrichocorycaeus anglicus* was only detected based on the old database, whereas, a closely related
466 species, *Ditrichocorycaeus affinis*, was detected based on the new database (a newly registered
467 sequence with higher similarity). Based on the old database, the average number of OTUs per
468 sample was 10.5±4.0 with a minimum of one OTU and a maximum of 21 OTUs (**Fig. 4; Table 3**).
469 Based on the new database, the average number of OTUs for each sample was 12.1±4.7 with a
470 minimum of 3 OTUs and a maximum of 26 OTUs. Per sample, the improvement was 20%, when
471 using the new database. The difference in identification success between the two databases was
472 statistically significant ($p<0.01$) (**Fig. 4; Table 3**). Three samples contained species that were
473 detected only based on the new database. The delta OTUs was on average 1.7±1.5 with a minimum
474 of -1, and a maximum of 8, which were above zero for the majority of samples with the exception
475 of five samples (**Fig. 5**).

476   From the 92 species detected using the new database, 45 were detected only based on the
477 newly registered sequences (**Table 3**). 18 species were detected based on several OTUs, when
478 using the new and old databases and of those, 15 species were detected by both new and existing
479 sequences (**Supplemental Table S3**). In the newly registered sequences, 13 species had higher
480 (mean 0.005±0.002) or equal similarity between the sequences of a singles species, except for two
481 species, *Eucalanus bungii*, and *Metridia lucens*. By the addition of new sequences, the occurrence
482 frequency increased or was the same as for existing sequences (14 species, mean 1.6±0.8 samples),
483 except for *Metridia pacifica* in the number of occurrences. The detection of also increased for 13
484 species (mean 215±266 sequences) with the exception of *Metridia lucens* and *Metridia pacifica*.

485   For the Sea of Okhotsk (Okhotsk Tower), 45 species associated with the newly registered
486 sequences were detected. From those, 16 species were detected based on the sequences registered
487 from the Sea of Okhotsk (**Supplemental Table S3**). Eight species (*Pseudocalanus newmani*
488 present in 175 samples, *Oithona atlantica* 127 samples, *Oithona similis* 105 samples,
489 *Pseudocalanus mimus* 103 samples, *Mesocalanus tenuicornis* 100 samples, *Neocalanus flemingeri*
490 71 samples, *Eucalanus bungii* 39 samples, *Microsetella norvegica* 39 samples) displayed high
491 occurrence frequencies (>20% in >35 samples out of 194 samples). The remaining 29 detected
492 species were registered from other areas, including 27 species of Copeoda, one species of

493 Branchiopoda, and one species of Sagittoidea. Six species (*Clausocalanus arcuicornis* present in
494 80 samples, *Clausocalanus pergens* 74 samples, *Centropages abdominalis* 69 samples,
495 *Ditrichocorycaeus affinis* 53 samples, *Clausocalanus farrani* 47 samples, *Oncaea venusta* 43
496 samples) that displayed high occurrence frequencies >20% in >35 samples, identified based on
497 sequences registered from the CTWNP and/or the Sea of Japan. Among these, 13 species of
498 Copepoda (*Acartia danae*, *Acartia negligens*, *Calanus agulhensis*, *Scolecithricella dentata*,
499 *Oithona longispina*, *Oithona nana*, *Oithona pulla*, *Oithona setigera*, *Oncaea bispinosa*, *Oncaea*
500 *venusta venella*, *Oncaea zernovi*, *Macrosetella gracilis*, *Euterpina acutifrons*) were detected based
501 on the sequences registered from warm waters, such as the WTWNP, East China Sea, and the
502 Nansei Islands.

503
504

Relationship between seasonality and inclusion of sequences from several areas

507 At the sampling location (Okhotsk Tower) in the Sea of Okhotsk, the sea surface temperature
508 (SST) showed clear seasonality with low temperatures in winter (December–January)/spring
509 (March–May) and high temperatures in summer (June–August)/autumn (September–November),
510 and the average temperatures in each season were 3.8±4.0°C for spring, 16.4±3.4°C for summer,
511 13.8±5.4°C for autumn, and 0.8±2.6°C for winter (**Fig. 5**, **Supplemental Table S2**). Based on the
512 18S marker, the delta OTUs (OTUs of the new database minus the old database) increased from
513 the summer to autumn of 2015 (**Fig. 5**). The mean of delta OTUs in each season was: 0.6±0.7 in
514 spring, 1.3±1.4 in summer, 1.9±2.0 in autumn, 1.3±1.8 in winter (**Fig. 6**). There was a significant
515 difference in Delta OTUs between spring and autumn ($p<0.01$). Based on the 28S marker, the delta
516 OTUs showed an increasing trend in the autumn (**Fig.5**). The mean of delta OTUs detected in each
517 season was 1.2±1.4 in spring, 1.7±1.3 in summer, 2.1±1.9 in autumn, and 1.8±1.5 in winter (**Fig.
518 6**). There was a significant difference in the Delta OTU between spring and summer, and spring
519 and autumn ($p<0.01$).

520
521

Discussion

523 The variable ocean environment is one of the main factors supporting marine diversity including
524 zooplankton in Japanese waters, and the molecular approach is a useful tool to solve difficulties in
525 species identification (Fujikura et al., 2010; Ohtsuka & Nishida, 2017). In this study, we focused
526 on dominant zooplankton species collected from six areas around Japan covering subarctic to
527 subtropical regions. Taxonomically verified sequences of both 18S and 28S rRNA genes were
528 obtained for nearly 100 species belonging to Arthropoda and Chaetognatha. Among those species
529 about 66 and 41 species were previously not represented in the public databases based on the 18S
530 and 28S rRNA gene sequences. This indicates that there were many unregistered rRNA gene

531   sequences even for the dominant zooplankton species in the western North Pacific Ocean. Based
532   on a test with metabarcoding data from field samples, the taxonomically verified and newly
533   registered sequences contributed to the improved taxonomic classification. More than 19 species
534   detected based on both the 18S and 28S markers were taxonomically identified based on the newly
535   registered sequences, including the dominant zooplankton species (e.g. *P. newmani*). The accuracy
536   of metabarcoding analysis largely depends on the sequence data using specimens classified by
537   taxonomic specialists (Lindeque et al., 2013). Based on the results of this study we further
538   demonstrated the importance of registration of taxonomically verified sequences for zooplankton
539   to improve identification in metabarcoding analysis from the marine waters around Japan.

540   In this study, a large number of the taxonomically verified and newly registered sequences
541   were from the small copepods belonging to *Oithona* and *Oncaea* based on both the 18S and 28S
542   markers, and *Clausocalanus* based on the 18S marker. The sequences associated with those genera
543   were mainly registered from warm sea areas of the WTWNP and the East China Sea. The sequence
544   data registration has previously mainly focused on marine zooplankton from high latitudes with
545   low water temperatures (Bucklin et al., 2010a). At the same time, high zooplankton diversity is
546   observed in the low latitudes with warm water temperatures, where the community is mostly
547   composed of small zooplankton species (Turner, 2004). In addition to species in *Oithona*, *Oncaea*,
548   and *Clausocalanus*, other small copepod species were detected based on newly registered
549   sequences including species from the family Acartiidae, Clausocalanidae, Paracalanidae, and
550   Corycaeidae (Gallienne & Robins, 2001; Turner, 2004). The importance of small zooplankton, in
551   particular non-calanoid copepods including *Oithona* and *Oncaea*, has been previously overlooked
552   (Böttger-Schnack, 1995; Turner, 2004), and we found that sequence data from these small
553   copepods were also limited in the public database. The metabarcoding approach is efficient for
554   detecting diversity, including small copepods, and in this study morphologically verified sequence
555   registration aided successful identification of small species in metabarcoding analyses targeting
556   the 18S and 28S rRNA genes.

557   Although the number of available 18S rRNA gene sequences is high for eukaryotic taxa,
558   our study showed higher proportions of species with newly registered sequences for the 18S than
559   for the 28S rRNA gene. Although both the 18S and 28S rRNA genes have been used for
560   phylogenetic analysis of marine copepods (Blanco-Bercial, Bradford-Grieve & Bucklin, 2011),
561   the 28S rRNA gene has been commonly used for species identification (Kiesling et al., 2002;
562   Blanco-Bercial, Bradford-Grieve & Bucklin, 2011). For example in the Kuroshio region in the
563   western North Pacific, sequences for more than 100 calanoid copepod species have been registered
564   based on the 28S rRNA gene (Hirai et al., 2015). The metabarcoding analysis also showed larger
565   numbers of OTUs with BLAST-hit to newly-registered sequences based on the 18S marker than
566   by the 28S marker, supporting the importance of adding taxonomically verified sequences,
567   especially for the 18S rRNA gene, genes to public databases. The species-specific homology of
568   zooplankton sequences showed higher values for newly registered sequences than for the existing
569   sequences for many species, and the frequency of occurrence and the number of sequences also
570   increased. This suggests that the registration of reference sequences of zooplankton species found

571 from the field, which may differ between individuals from different sampling location, has
572 improved identification accuracy and detection sensitivity.
573   In the metabarcoding analysis based on the filed samples from the Sea of Okhotsk, the
574 number of detected OTUs increased based on the new database containing the newly registered
575 sequences. The different OTUs of the same species were detected based on the old and new
576 databases. It was suggested that the same reads in metabarcoding samples were detected as a
577 different OTU due to registering new reference sequences with higher similarity. This result shows
578 improved species identification not only at species level but also at OTU level by the new sequence
579 registration. The seasonal differences of OTUs representing a single species with BLAST-hit to
580 newly registered sequences differed between the 18S and 28S rRNA genes. The number of OTUs
581 identified based on the newly registered sequences was relatively high based on the 18S and 28S
582 rRNA genes in autumn. This can be explained by the strong seasonal variability in the dominant
583 currents present in the sampling location. The Soya Warm Current (SWC) is especially dominant
584 in autumn on the east coast of Hokkaido in the Okhotsk Sea, and warm water zooplankton species
585 are transported by the SWC from the Sea of Japan (Hamaoka et al., 2010). In the metabarcoding
586 analysis, we also detected warm water species such as *C. elliptica*, *T. gracilis*, *O. aruensis*, for
587 which the taxonomically verified sequences originated only from the Nansei Islands. Because
588 about half of the species detected from the Sea of Okhotsk based on the newly registered sequences
589 originating from other sea areas, the accuracy of species identification can be improved by
590 including zooplankton sequences not only from the survey area but also from other sea areas. The
591 increase in the number of zooplankton species that could be detected by the taxonomically verified
592 sequences added to the database in this study shows that further registration of taxonomically
593 verified sequences is important.
594   There is also a need for the registration of rare (non-dominant) zooplankton species for
595 further improvement of taxonomic identification based on molecular techniques. For example, the
596 18S rRNA gene is relatively conserved, and the same sequences are observed between closely
597 related species (Tang et al., 2012). When only one species was registered between two closely
598 related species, the other unregistered species can be misidentified by metabarcoding analysis. In
599 addition, combinations with other molecular markers with high mutation rates (e.g. mitochondrial
600 COI) would be useful for solving problems of taxonomic resolution, although these markers have
601 difficulty in amplifying various zooplankton species with high diversity. We focused on the rRNA
602 genes, which are common for eukaryotic metabarcoding; however, registration of other markers is
603 also helpful to improve the accuracy of metabarcoding using multiple markers (Neigel, Domingo
604 & Stake, 2007). Because molecular tools are useful for the taxonomic characterization of
605 zooplankton communities, we hope that future efforts of sequence registration will support further
606 understanding of the roles of zooplankton in the marine ecosystem and facilitate the detection of
607 long-term changes in their community structure.
608
609
610 **Conclusions**

611 This study aimed to improve zooplankton species identification based on metabarcoding by
612 registering taxonomically verified sequences from six environmentally different areas in Japanese
613 waters. A total of 331 18S and 28S rRNA gene sequences from 100 zooplankton species were
614 registered to a public database. In the field samples used for investigating the improvement of
615 taxonomic identification, even the dominant species lacked the sequences available in the public
616 database. Those species with unregistered sequences were mainly composed of small non-calanoid
617 copepods (i.e. *Oithona* and *Oncaea*). Verified improvement of taxonomic identification was
618 demonstrated by more than one-fifth of OTUs from the field samples that were identified to species
619 level based on the newly registered sequences of the 18S and 28S rRNA genes. The reference
620 database including the taxonomically verified sequences improved the accuracy of zooplankton
621 detection both overall and in individual samples, and the accuracy of species identification also
622 improved at the population level due to higher similarity values among sequences of the same
623 species. Continuous registration of sequence data covering various environmental conditions is
624 necessary for further improvement of taxonomic identification of zooplankton using
625 metabarcoding analysis and it also facilitates monitoring of marine ecosystems by molecular tools.
626
627

Acknowledgments

629 K. Kamiya and S. Ori are thanked for the technical assistance. We thank H. Yoshida and K. Murai
630 for the plankton sampling and CTD observation around the Okhotsk Tower. We thank the captains,
631 officers, crews, and scientists of R/V Hokko-Maru, R/V Wakataka-Maru, R/V Shunyo-Maru, R/V
632 Tenyo-Maru, R/V Soyo-Maru, R/V Yoko-Maru, and the Fisheries Agency of Japan, for their
633 invaluable support with sampling. We are grateful to Dr. Y. Igeta for the management of sampling
634 in the Sea of Japan. We extend our gratitude to Drs. H. Kuroda and T. Nakanowatari for collecting
635 the essential samples.
636
637

References

639 Ajani PA, Henriquez-Nunez HF, Verma A, Nagai S, Uchida H, Tesoriero MJ, Farrell H, Zammit
640   A, Brett S, Murray SA. 2022. Mapping the development of a Dinophysis bloom in a
641   shellfish aquaculture area using a novel molecular qPCR assay. *Harmful Algae* 116:102253.
642   DOI: 10.1016/J.HAL.2022.102253.
643 Bellemain E, Carlsen T, Brochmann C, Coissac E, Taberlet P, Kauserud H. 2010. ITS as an
644   environmental DNA barcode for fungi: An in silico approach reveals potential PCR biases.
645   *BMC Microbiology* 10:189. DOI: 10.1186/1471-2180-10-189.
646 Berry TE, Saunders BJ, Coghlan ML, Stat M, Jarman S, Richardson AJ, Davies CH, Berry O,
647   Harvey ES, Bunce M. 2019. Marine environmental DNA biomonitoring reveals seasonal
648   patterns in biodiversity and identifies ecosystem responses to anomalous climatic events.
649   *PLOS Genetics* 15:e1007943. DOI: 10.1371/journal.pgen.1007943.

650 Blanco-Bercial L, Bradford-Grieve J, Bucklin A. 2011. Molecular phylogeny of the Calanoida
651   (Crustacea: Copepoda). *Molecular Phylogenetics and Evolution* 59:103–113. DOI:
652   10.1016/j.ympev.2011.01.008.

653 Boltovskoy D, Correa N, Boltovskoy A. 2002. Marine zooplanktonic diversity: a view from the
654   South Atlantic. *Oceanologica Acta* 25:271–278. DOI: https://doi.org/10.1016/S0399-
655   1784(02)01199-4.

656 Böttger-Schnack R. 1995. Summer distribution of micro- and small mesozooplankton in the Red
657   Sea and Gulf of Aden, with special reference to non-calanoid copepods. *Marine Ecology*
658   *Progress Series* 118:81–102. DOI: 10.3354/meps118081.

659 Bucklin A, Hopcroft RR, Kosobokova KN, Nigro LM, Ortman BD, Jennings RM, Sweetman CJ.
660   2010a. DNA barcoding of Arctic Ocean holozooplankton for species identification and
661   recognition. *Deep-Sea Research Part II: Topical Studies in Oceanography* 57:40–48. DOI:
662   10.1016/j.dsr2.2009.08.005.

663 Bucklin A, Lindeque PKP, Rodriguez-Ezpeleta N, Albaina A, Lehtiniemi M. 2016.
664   Metabarcoding of marine zooplankton: prospects, progress and pitfalls. *Journal of Plankton*
665   *Research* 38:393–400. DOI: 10.1093/plankt/fbw023.

666 Bucklin A, Nishida S, Schnack-Schiel S, Wiebe PH, Lindsay D, Machida RJ, Copley NJ. 2010b.
667   A Census of Zooplankton of the Global Ocean. In: *Life in the World's Oceans: Diversity,*
668   *Distribution, and Abundance*. Wiley-Blackwell, 247–265. DOI:
669   10.1002/9781444325508.ch13.

670 Bucklin A, Steinke D, Blanco-Bercial L. 2011. DNA Barcoding of Marine Metazoa. *Annual*
671   *Review of Marine Science* 3:471–508. DOI: 10.1146/annurev-marine-120308-080950.

672 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009.
673   BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. DOI: 10.1186/1471-
674   2105-10-421.

675 Carroll EL, Gallego R, Sewell MA, Zeldis J, Ranjard L, Ross HA, Tooman LK, O'Rorke R,
676   Newcomb RD, Constantine R. 2019. Multi-locus DNA metabarcoding of zooplankton
677   communities and scat reveal trophic interactions of a generalist predator. *Scientific Reports*
678   9:281. DOI: 10.1038/s41598-018-36478-x.

679 Cheung KLY, Huen J, Houry WA, Ortega J. 2010. Comparison of the multiple oligomeric
680   structures observed for the Rvb1 and Rvb2 proteins. *Biochemistry and cell biology =*
681   *Biochimie et biologie cellulaire* 88:77–88. DOI: 10.1139/o09-159.

682 Clarke LJ, Beard JM, Swadling KM, Deagle BE. 2017. Effect of marker choice and thermal
683   cycling protocol on zooplankton DNA metabarcoding studies. *Ecology and Evolution*
684   7:873–883. DOI: https://doi.org/10.1002/ece3.2667.

685 Fujikura K, Lindsay D, Kitazato H, Nishida S, Shirayama Y. 2010. Marine Biodiversity in
686   Japanese Waters. *PLoS ONE* 5:e11836. DOI: 10.1371/journal.pone.0011836.

687 Fuller MS, Sándor G, Punske B, Taccardi B, MacLeod RS, Ershler PR, Green LS, Lux RL.
688   2001. Multiplexed species-specific PCR protocol to discriminate four N. Atlantic *Calanus*
689   species, with an mtCOI gene tree for ten *Calanus* species. *Marine Biology* 139:279–287.
690   DOI: 10.1007/s002270100548.

691 Gallienne CP, Robins DB. 2001. Is *Oithona* the most important copepod in the world's oceans?
692   *Journal of Plankton Research* 23:1421–1432. DOI: 10.1093/plankt/23.12.1421.

693 Goetze E. 2003. Cryptic speciation on the high seas; global phylogenetics of the copepod family
694   Eucalanidae. *Proceedings of the Royal Society B: Biological Sciences* 270:2321–2331. DOI:
695   10.1098/rspb.2003.2505.

696   Hamaoka S, Katakura S, Nagata R, Hirakawa K. 2010. Seasonal variability in zooplankton
697   community, with special reference to copepods in the inshore waters. In: *Proceedings of the*
698   *25th International Symosium on Okhotsk Sea & Sea Ice*. Mombetsu, 169–172.
699   Harvey JBJ, Fisher JL, Ryan JP, Johnson SB, Peterson WT, Vrijenhoek RC. 2018. Changes in
700   zooplankton assemblages in northern Monterey Bay, California, during a fall transition.
701   *Marine Ecology Progress Series* 604:99–120. DOI: 10.3354/meps12742.
702   Hirai J, Hidaka K, Nagai S, Ichikawa T. 2017a. Molecular-based diet analysis of the early
703   post-larvae of Japanese sardine *Sardinops melanostictus* and Pacific round herring
704   *Etrumeus teres*. *Marine Ecology Progress Series* 564:99–113. DOI: 10.3354/meps12008.
705   Hirai J, Katakura S, Kasai H, Nagai S. 2017b. Cryptic Zooplankton Diversity Revealed by a
706   Metagenetic Approach to Monitoring Metazoan Communities in the Coastal Waters of the
707   Okhotsk Sea, Northeastern Hokkaido. *Frontiers in Marine Science* 4:379. DOI:
708   10.3389/fmars.2017.00379.
709   Hirai J, Kuriyama M, Ichikawa T, Hidaka K, Tsuda A. 2015. A metagenetic approach for
710   revealing community structure of marine planktonic copepods. *Molecular Ecology*
711   *Resources* 15:68–80. DOI: 10.1111/1755-0998.12294.
712   Hirai J, Shimode S, Tsuda A. 2013. Evaluation of ITS2-28S as a molecular marker for
713   identification of calanoid copepods in the subtropical western North Pacific. *Journal of*
714   *Plankton Research* 35:644–656. DOI: 10.1093/plankt/fbt016.
715   Hirai J, Tachibana A, Tsuda A. 2020. Large-scale metabarcoding analysis of epipelagic and
716   mesopelagic copepods in the Pacific. *PLoS ONE* 15:e0233189. DOI:
717   10.1371/journal.pone.0233189.
718   Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
719   Improvements in performance and usability. *Molecular Biology and Evolution* 30:772–780.
720   DOI: 10.1093/molbev/mst010.
721   Kiesling TL, Wilkinson E, Rabalais J, Ortner PB, McCabe MM, Fell JW. 2002. Rapid
722   identification of adult and naupliar stages of copepods using DNA hybridization
723   methodology. *Marine Biotechnology* 4:30–39. DOI: 10.1007/pl00021689.
724   Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular evolutionary
725   genetics analysis across computing platforms. *Molecular Biology and Evolution* 35:1547–
726   1549. DOI: 10.1093/molbev/msy096.
727   Kunin V, Engelbrektson A, Ochman H, Hugenholtz P. 2010. Wrinkles in the rare biosphere:
728   Pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental*
729   *Microbiology* 12:118–123. DOI: 10.1111/j.1462-2920.2009.02051.x.
730   Lindeque PKP, Parry HE, Harmer RA, Somerfield PJP, Atkinson A. 2013. Next generation
731   sequencing reveals the hidden diversity of zooplankton assemblages. *PLoS ONE* 8:e81327.
732   DOI: 10.1371/journal.pone.0081327.
733   Machida RJ, Hashiguchi Y, Nishida M, Nishida S. 2009. Zooplankton diversity analysis through
734   single-gene sequencing of a community sample. *BMC Genomics* 10:438. DOI:
735   10.1186/1471-2164-10-438.
736   Machida RJ, Knowlton N. 2012. PCR Primers for Metazoan Nuclear 18S and 28S Ribosomal
737   DNA Sequences. *PLoS ONE* 7:e46180. DOI: 10.1371/journal.pone.0046180.
738   Nagai S, Sildever S, Nishi N, Tazawa S, Basti L, Kobayashi T, Ishino Y. 2022. Comparing PCR-
739   generated artifacts of different polymerases for improved accuracy of DNA metabarcoding.
740   *Metabarcoding and Metagenomics 6: e77704* 6:e77704-. DOI: 10.3897/MBMG.6.77704.

741   Neigel J, Domingo A, Stake J. 2007. DNA barcoding as a tool for coral reef conservation. *Coral*
742   *Reefs* 26:487–499. DOI: 10.1007/s00338-007-0248-4.
743   Ohtsuka S, Nishida S. 2017. Copepod biodiversity in Japan: recent advances in Japanese
744   copepodology. In: Springer, Tokyo, 565–602. DOI: 10.1007/978-4-431-56432-4_22.
745   Questel JM, Hopcroft RR, DeHart HM, Smoot CA, Kosobokova KN, Bucklin A. 2021.
746   Metabarcoding of zooplankton diversity within the Chukchi Borderland, Arctic Ocean:
747   improved resolution from multi-gene markers and region-specific DNA databases. *Marine*
748   *Biodiversity* 51:4. DOI: 10.1007/s12526-020-01136-x.
749   R Core team. 2022. R: A language and environment for statistical computing. R Foundation for
750   Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. *R: A Language*
751   *and Environment for Statistical Computing. R Foundation for Statistical Computing ,*
752   *Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.* DOI:
753   10.2788/95827.
754   Raupach MJ, Astrin JJ, Hannig K, Peters MK, Stoeckle MY, Wägele JW. 2010. Molecular
755   species identification of Central European ground beetles (Coleoptera: Carabidae) using
756   nuclear rDNA expansion segments and DNA barcodes. *Frontiers in Zoology* 7:26. DOI:
757   10.1186/1742-9994-7-26.
758   Razouls C, de Bovée F, Kouwenberg J, Desreumaux N. 2020.Diversity and Geographic
759   Distribution of Marine Planktonic Copepods. *Available at https://copepodes.obs-*
760   *banyuls.fr/en/* (accessed June 22, 2020).
761   Sildever S, Kawakami Y, Kanno N, Kasai H, Shiomoto A, Katakura S, Nagai S. 2019. Toxic
762   HAB species from the Sea of Okhotsk detected by a metagenetic approach, seasonality and
763   environmental drivers. *Harmful Algae* 87:101631. DOI: 10.1016/j.hal.2019.101631.
764   Sonnenberg R, Nolte A, Tautz D. 2007. An evaluation of LSU rDNA D1-D2 sequences for their
765   use in species identification. *Frontiers in Zoology* 4:6. DOI: 10.1186/1742-9994-4-6.
766   Steinberg DK, Van Mooy BAS, Buesseler KO, Boyd PW, Kobari T, Karl DM. 2008. Bacterial
767   vs. zooplankton control of sinking particle flux in the ocean's twilight zone. *Limnology and*
768   *Oceanography* 53:1327–1338. DOI: 10.4319/lo.2008.53.4.1327.
769   Tang CQ, Leasi F, Obertegger U, Kieneke A, Barraclough TG, Fontaneto D. 2012. The widely
770   used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity
771   surveys of the meiofauna. *Proceedings of the National Academy of Sciences* 109:16208–
772   16212. DOI: 10.1073/pnas.1209160109.
773   Tittensor DP, Mora C, Jetz W, Lotze HK, Ricard D, Berghe E Vanden, Worm B. 2010. Global
774   patterns and predictors of marine biodiversity across taxa. *Nature* 466:1098–1101. DOI:
775   10.1038/nature09329.
776   Turner JT. 2004. The Importance of Small Planktonic Copepods and Their Roles in Pelagic
777   Marine Food Webs. *Zoological Studies* 43:255–266.
778   de Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R, Lara E, Berney C, Le Bescot N,
779   Probert I, Carmichael M, Poulain J, Romac S, Colin S, Aury J-M, Bittner L, Chaffron S,
780   Dunthorn M, Engelen S, Flegontova O, Guidi L, Horak A, Jaillon O, Lima-Mendez G, Luke
781   J, Malviya S, Morard R, Mulot M, Scalco E, Siano R, Vincent F, Zingone A, Dimier C,
782   Picheral M, Searson S, Kandels-Lewis S, Acinas SG, Bork P, Bowler C, Gorsky G,
783   Grimsley N, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S, Raes J, Sieracki ME,
784   Speich S, Stemmann L, Sunagawa S, Weissenbach J, Wincker P, Karsenti E, Boss E,
785   Follows M, Karp-Boss L, Krzic U, Reynaud EG, Sardet C, Sullivan MB, Velayoudon D.

786   2015. Eukaryotic plankton diversity in the sunlit ocean. *Science* 348:1261605. DOI:
787   10.1126/science.1261605.
788 Vassart G, Georges M, Monsieur R, Brocas H, Lequarre AS, Christophe D. 1987. A Sequence in
789   M13 Phage Detects Hypervariable Minisatellites in Human and Animal DNA. *Science*
790   235:683–684. DOI: 10.1126/SCIENCE.2880398.
791 Ward BA, Dutkiewicz S, Jahn O, Follows MJ. 2012. A size-structured food-web model for the
792   global ocean. *Limnology and Oceanography* 57:1877–1891. DOI:
793   10.4319/lo.2012.57.6.1877.
794 WoRMS Editorial Board. 2021.World Register of Marine Species. DOI: 10.14284/170.
795 Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W,
796   Glöckner FO. 2014. The SILVA and "All-species Living Tree Project (LTP)" taxonomic
797   frameworks. *Nucleic Acids Research* 42:D643–D648. DOI: 10.1093/nar/gkt1209.

798
799
800

Figures

802 **Figure 1: Sampling map of the stations in six sea areas surrounding Japan.**
803 The square shows enlarged views of research stations in the Nansei Islands. Triangle (△), plus (+),
804 circle (○), square (□), diamond (◇), and inverted triangle (▽) indicate six sea areas, i.e. the Sea of
805 Okhotsk, the Sea of Japan, the Cold Temperate Western North Pacific (CTWNP), the Warm
806 Temperate Western North Pacific (WTWNP), the East China Sea, and the Nansei Islands,
807 respectively. Black circle (●) shows the Okhotsk Tower which is the metabarcoding sampling
808 station. The coordinates of all stations are available in Table 1. The large arrow denotes the
809 Kuroshio current path in May 2018 based on the Quick Bulletin of Ocean Conditions issued by
810 the Japan Coast Guard. (https://www1.kaiho.mlit.go.jp/ KANKYO/KAIYO/qboc/index_E.html).
811

812 **Figure 2: Number of morphologically verified species per genus in registered zooplankton**
813 **sequences.**
814 The number of morphologically identified species per genus based on the 18S (left, 166 sequences
815 from 96 species) and the 28S (right, 165 sequences from 95 species) rRNA genes registered in this
816 study. The family and order (right side) match with the genus (left side). Colors indicate that the
817 sequences matched with a higher rank (family or more, orange), genus (blue), and species (i.e. pre-
818 exist sequences, grey) in the public database using BLAST.
819

820 **Figure 3: Number of morphologically verified zooplankton species from six sea areas for**
821 **which sequences were registered.**
822 18S (left) and 28S (right) rRNA gene sequences from all sampling stations (Total) and six sea
823 areas. Colors indicate that the sequences match with a higher rank (family or more, orange), genus
824 (blue), and species (i.e. pre-exist sequences, grey) in the public database using BLAST.
825

826 **Figure 4: Boxplots of OUT numbers associated with a single zooplankton species detected**
827 **based on old and new reference databases.**
828 Old (white) and New (orange) indicate the reference database without and with taxonomically
829 verified and newly registered sequences, respectively. Diamond (◇) denotes the mean value of
830 the number of OTUs. Asterisk (**) means a significant difference ($p < 0.01$), and NS abbreviates
831 no significant difference.

832
833 **Figure 5: Seasonal variation of the sea surface temperature and the difference of (delta)**
834 **OTUs detected based on the rRNA genes from metabarcoding samples from the Okhostk**
835 **Tower based on the new and old reference databases.**
836 (A) seasonal variation of the sea surface temperature (SST). Seasonal variation of delta OTUs
837 based on the 18S rRNA gene (B) and based on the 28S rRNA gene (C). The delta OTUs were
838 obtained by subtracting the number of OTUs in the old database from the number of OTUs in the
839 new database. The delta OTUs greater than 0 indicates that the number of OTUs in the New
840 database is larger than in the Old database. The horizontal axis indicates the sampling period (10
841 April 2012 to 29 March 2016). The vertical dotted lines separate the seasons every three months.
842 All data is shown in Supplemental Table S2.
843
844 **Figure 6: Boxplot of the difference (delta) of OTUs from rRNA genes metabarcoding samples**
845 **between new and old reference databases by seasons at the Okhotsk Tower.**
846 The delta OTUs from 18S (left) and 28S (right) rRNA gene sequences were compared between
847 seasons. The seasons were separated by spring, March-May; summer, June-August; autumn,
848 September-November; winter, December-February. Diamond ($\diamondsuit$) indicates the mean value of the
849 delta OTUs by season. Asterisk (**) means a significant difference ($p < 0.01$), and NS abbreviates
850 no significant difference.
851
852
853
854 **Tables**
855 **Table 1:**
856 **Sampling information for the locations of taxonomically verified sequences and**
857 **metabarcoding of zooplankton.**
858 Sample indicates which samples were used for analysis. Sea area shows the sampling area (and
859 abbreviations) for this study. Depth shows the depth of the sampling station, not the net towing
860 depth for zooplankton collecting. SST: Sea Surface Temperature. * See Supplemental Table S2
861 for details of the sampling data.
862
863 **Table 2:**
864 **Summary of the taxonomically verified and newly registered zooplankton sequences from**
865 **this study.**
866 'Sequence' indicates the number of sequences, and 'Species' indicates the number of groups by
867 species which includes duplicate sequences of the same species from other specimens. The table
868 denoted our registered sequence was classified to what taxonomic level on the public database by
869 BLAST. 'Pre-exist' indicates a sequence with a complete match at a species level based on the
870 public database, and 'New' a match with taxonomic criteria of genus or higher level (i.e., the
871 sequences for which the species did not exist in the public database) database. Abbreviations:
872 CTWNP, Cold Temperate Western North Pacific; WTWNP, Warm Temperate Western North
873 Pacific.
874
875 **Table 3:**
876 **Overview of OTUs associated with a single zooplankton species detected from**
877 **metabarcoding samples based on the old and new reference databases.**

878   Zooplankton composition of metabarcoding samples from Okhotsk Tower based on the reference
879   databases with (new) and without (old) newly registered sequences, performed separately for the
880   18S and 28S rRNA gene sequences. 'Total' is the number of OTUs in the old and new databases
881   combined; 'New' and 'Old' are the number of OTUs in each database. 'Difference' is the number
882   of OTUs in the new database minus the number of OTUs in the old database, representing the
883   increase in the number of detections in the new database. The OTU number detected in both
884   databases ('Shared OTUs') or only in the respective database are also shown ('Unique OTUs').
885   However, most OTUs that appeared in the old database were also detected at the species level in
886   the new database (see text for details). 'OTU per sample' shows the respective OTU number
887   (mean, standard deviation, minimum, and maximum) per sample.
888

Supplemental Information

Supplemental Table S1:
Data of sequenced and registered individuals of all zooplankton in this study.

892   List of zooplankton sequences registered in this study, ordered by rDNA, sea area, and taxonomy
893   (species name), with sampling information (date, cruise, and station), accession number,
894   taxonomic information (kingdom to genus) and BLAST search results. In the BLAST column,
895   'Other', 'Genus', and 'Species' indicate that the sequences matched at a higher rank (family or
896   higher), genus level, and species level (i.e. fully matched to pre-existing sequences) in the public
897   database, respectively.
898

Supplemental Table S2: List of Okhotsk tower samples used in the metabarcoding analysis.

900   Metabarcoding sample information is listed in order of sampling date (total 199 samples, 18S
901   rRNA 117 samples, 28S rRNA 195 samples). SST is the abbreviation for sea surface temperature
902   (Celsius degrees). The values in 'Detected OTUs' are the number of detected OTUs; Figures 4, 5,
903   6, and Table 3 are based on these data. The 'Old' and 'New' columns within the rRNA columns
904   indicate the reference database used without (old) and with (new) our newly registered sequences,
905   respectively.
906

Supplemental Table S3: List of zooplankton OTUs associated with species from Arthropoda
and Chaetognath present in the Okhotsk Tower metabarcoding analysis.

909   All zooplankton OTUs used in the data analysis (Figures 4, 5, 6, and Table 3) are shown. The 'Top-
910   hit name' column is the species name of the OTU hit in the homology search. The asterisk on the
911   left column of the top-hit name indicates the species containing multiple sequences; * indicates
912   detection in the new database, ** in the old database, and *** in the new and old databases.
913   'Similarity' is the homology value in the metabarcoding analysis. 'Occurrence frequency' is the
914   number of times an OTU was detected. 'Sum of reads' is the total number of sequences. 'Reads per
915   sample' is the mean value of sequences per sample. 'Zooplankton sampling area' is the sea area
916   where the species registered in this study were sampled; Values are the number of sequences
917   registered.

Figure 1

Sampling map of the stations in six sea areas surrounding Japan.

The square shows enlarged views of research stations in the Nansei Islands. Triangle (△), plus (+), circle (○), square (□), diamond (◇), and inverted triangle (▽) indicate six sea areas, i.e. the Sea of Okhotsk, the Sea of Japan, the Cold Temperate Western North Pacific (CTWNP), the Warm Temperate Western North Pacific (WTWNP), the East China Sea, and the Nansei Islands, respectively. Black circle (●) shows the Okhotsk Tower which is the metabarcoding sampling station. The coordinates of all stations are available in Table 1. The large arrow denotes the Kuroshio current path in May 2018 based on the Quick Bulletin of Ocean Conditions issued by the Japan Coast Guard. (https://www1.kaiho.mlit.go.jp/ KANKYO/KAIYO/qboc/index_E.html).

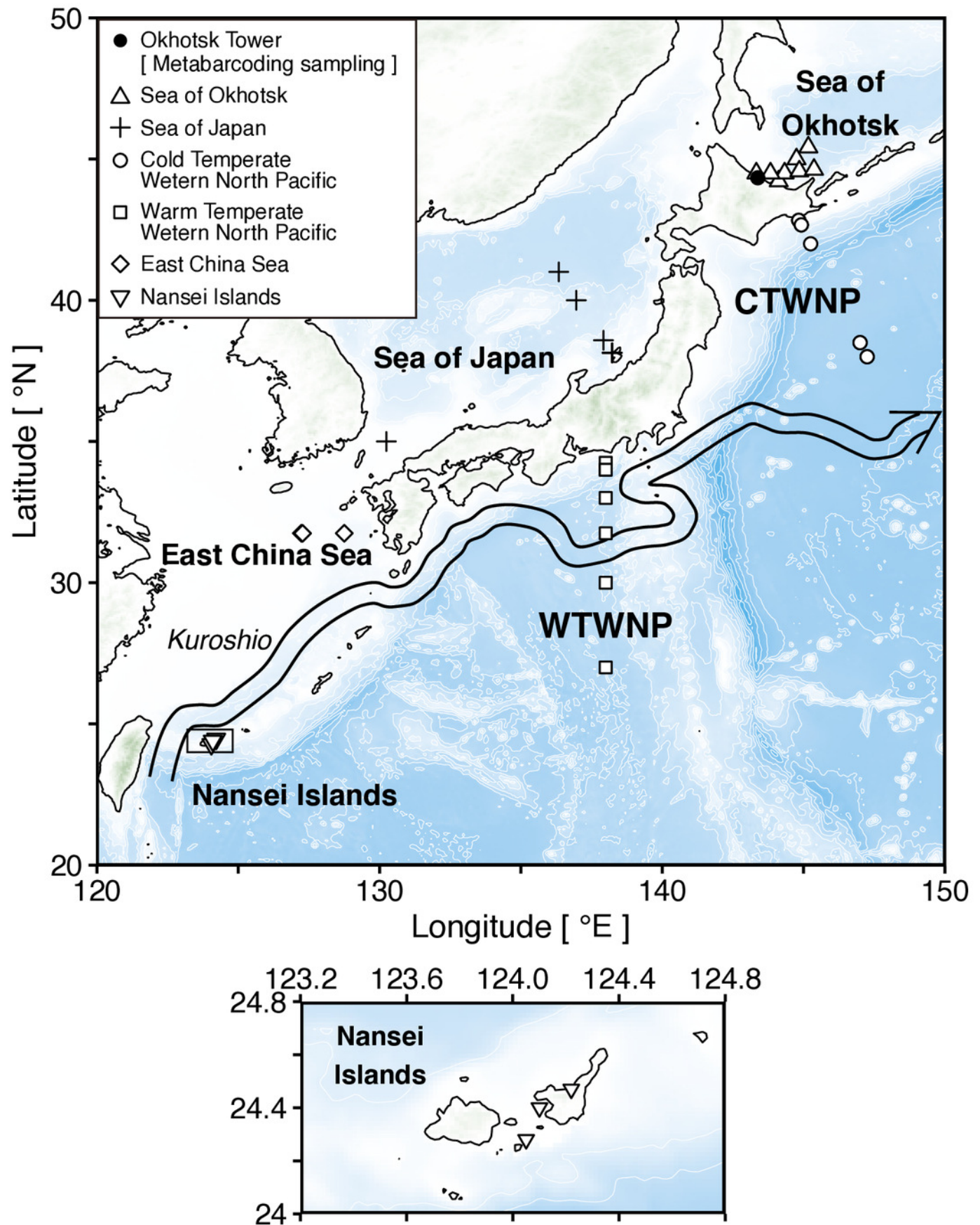



Figure 2

Number of morphologically verified species per genus in registered zooplankton sequences.

The number of morphologically identified species per genus based on the 18S (left, 166 sequences from 96 species) and the 28S (right, 165 sequences from 95 species) rRNA genes registered in this study. The family and order (right side) match with the genus (left side). Colors indicate that the sequences matched with a higher rank (family or more, orange), genus (blue), and species (i.e. pre-exist sequences, grey) in the public database using BLAST.

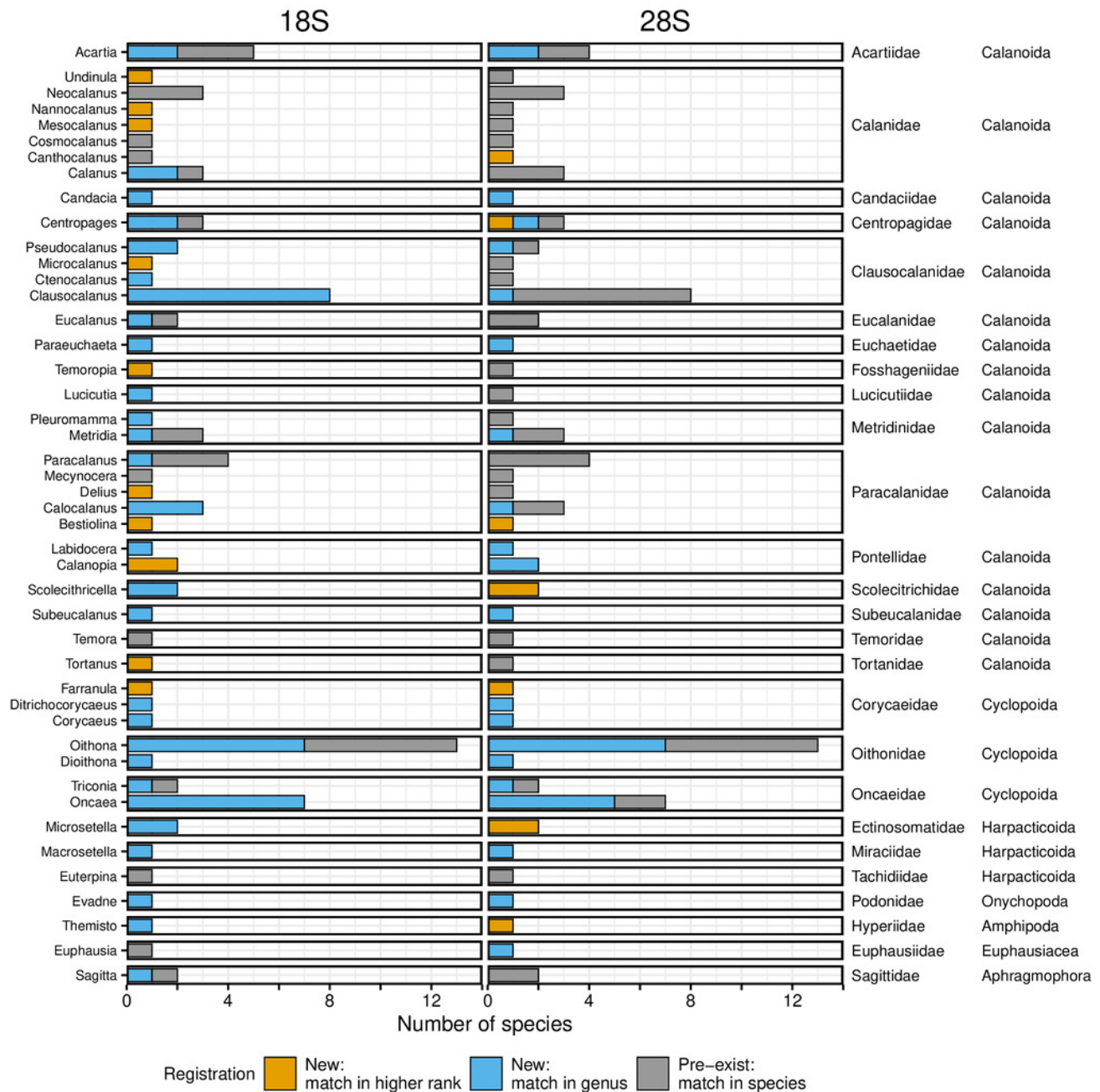



Figure 3

Number of morphologically verified zooplankton species from six sea areas for which sequences were registered.

18S (left) and 28S (right) rRNA gene sequences from all sampling stations (Total) and six sea areas. Colors indicate that the sequences match with a higher rank (family or more, orange), genus (blue), and species (i.e. pre-exist sequences, grey) in the public database using BLAST.

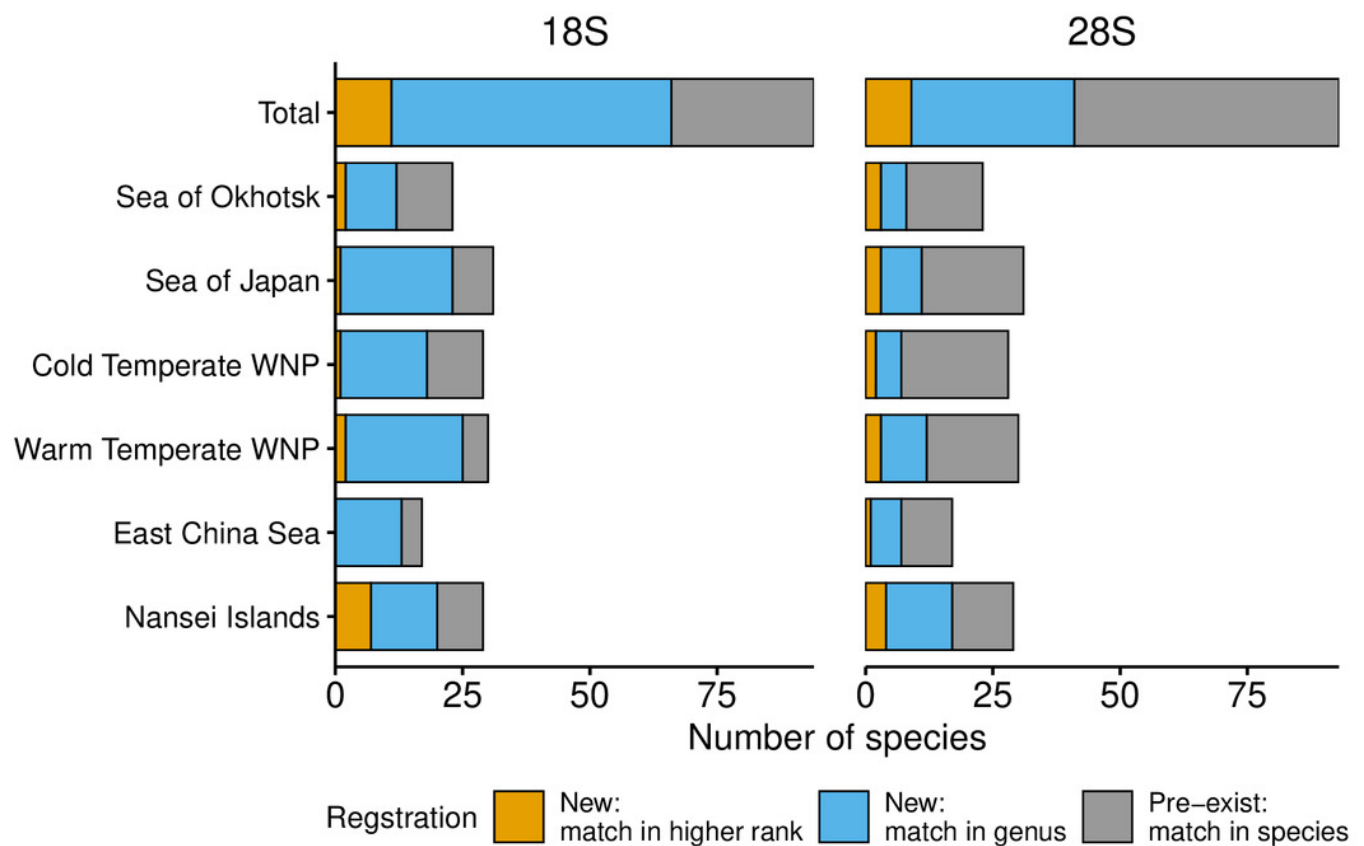



Figure 4

Boxplots of OUT numbers associated with a single zooplankton species detected based on old and new reference databases.

Old (white) and New (orange) indicate the reference database without and with taxonomically verified and newly registered sequences, respectively. Diamond (◇) denotes the mean value of the number of OTUs. Asterisk (**) means a significant difference ($p <$ 0.01), and NS abbreviates no significant difference.

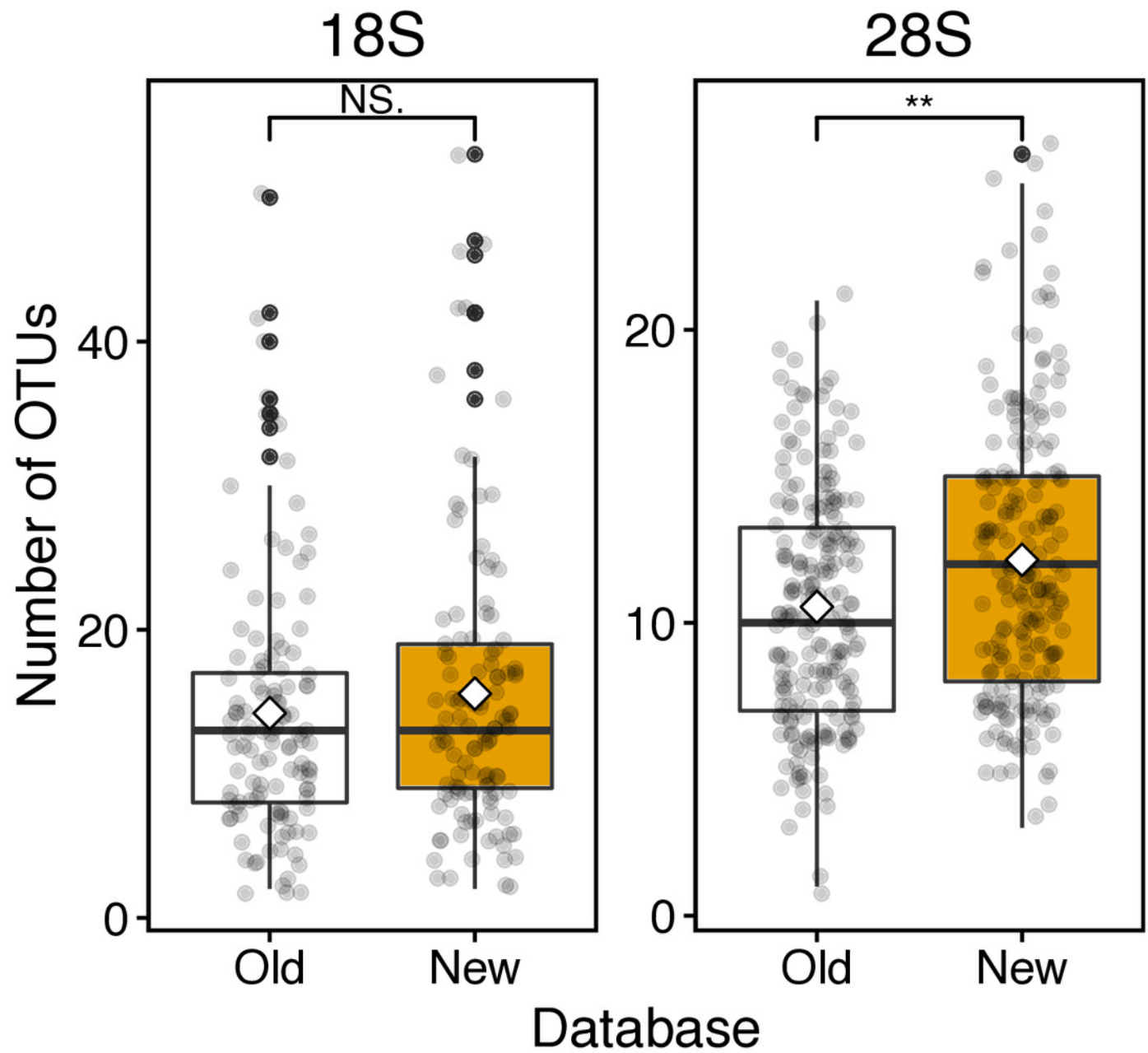



Figure 5

Seasonal variation of the sea surface temperature and the difference of (delta) OTUs detected based on the rRNA genes from metabarcoding samples from the Okhostk Tower based on the new and old reference databases.

(A) seasonal variation of the sea surface temperature (SST). Seasonal variation of delta OTUs based on the 18S rRNA gene (B) and based on the 28S rRNA gene (C). The delta OTUs were obtained by subtracting the number of OTUs in the old database from the number of OTUs in the new database. The delta OTUs greater than 0 indicates that the number of OTUs in the New database is larger than in the Old database. The horizontal axis indicates the sampling period (10 April 2012 to 29 March 2016). The vertical dotted lines separate the seasons every three months. All data is shown in Supplemental Table S2.

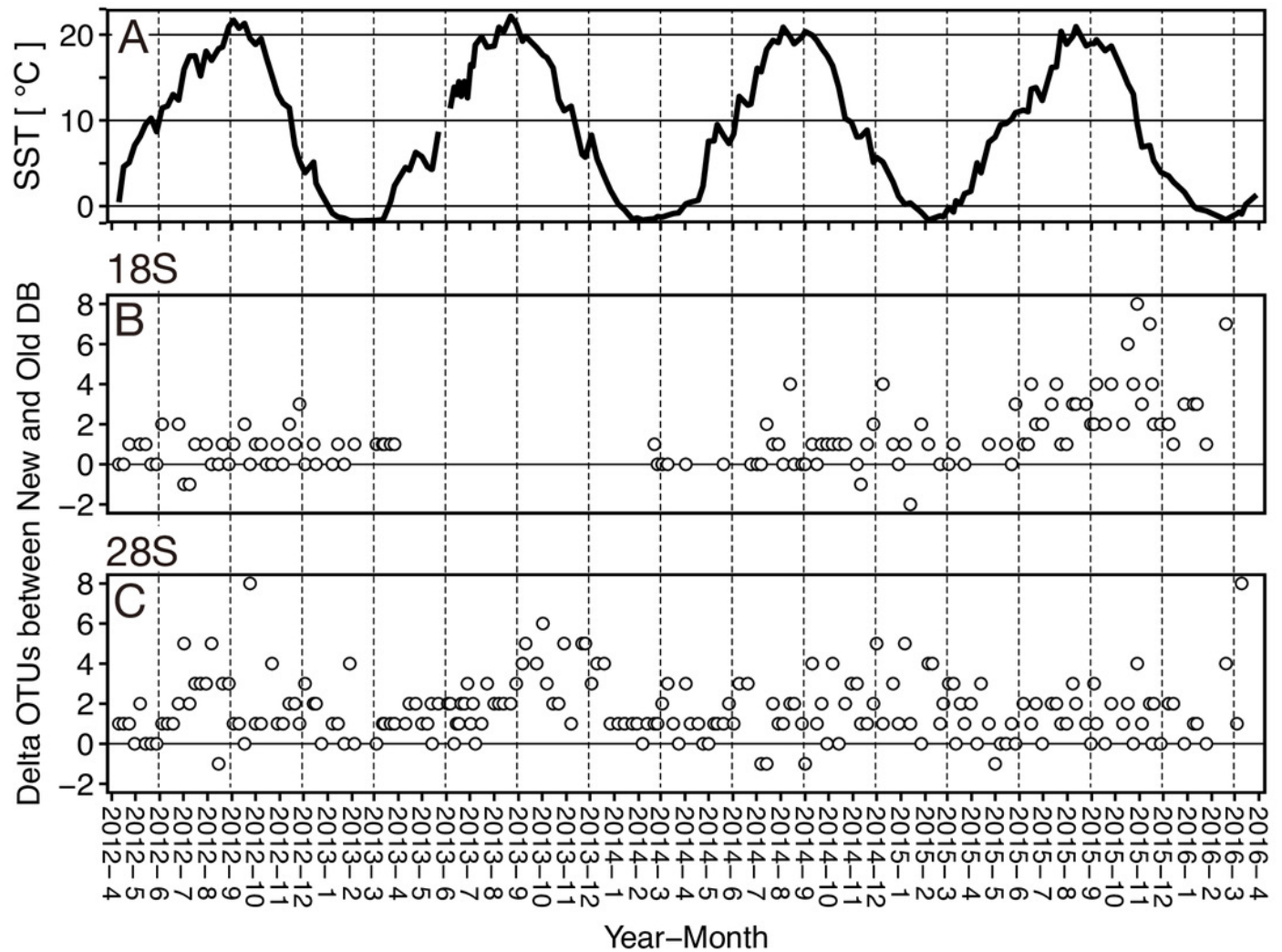
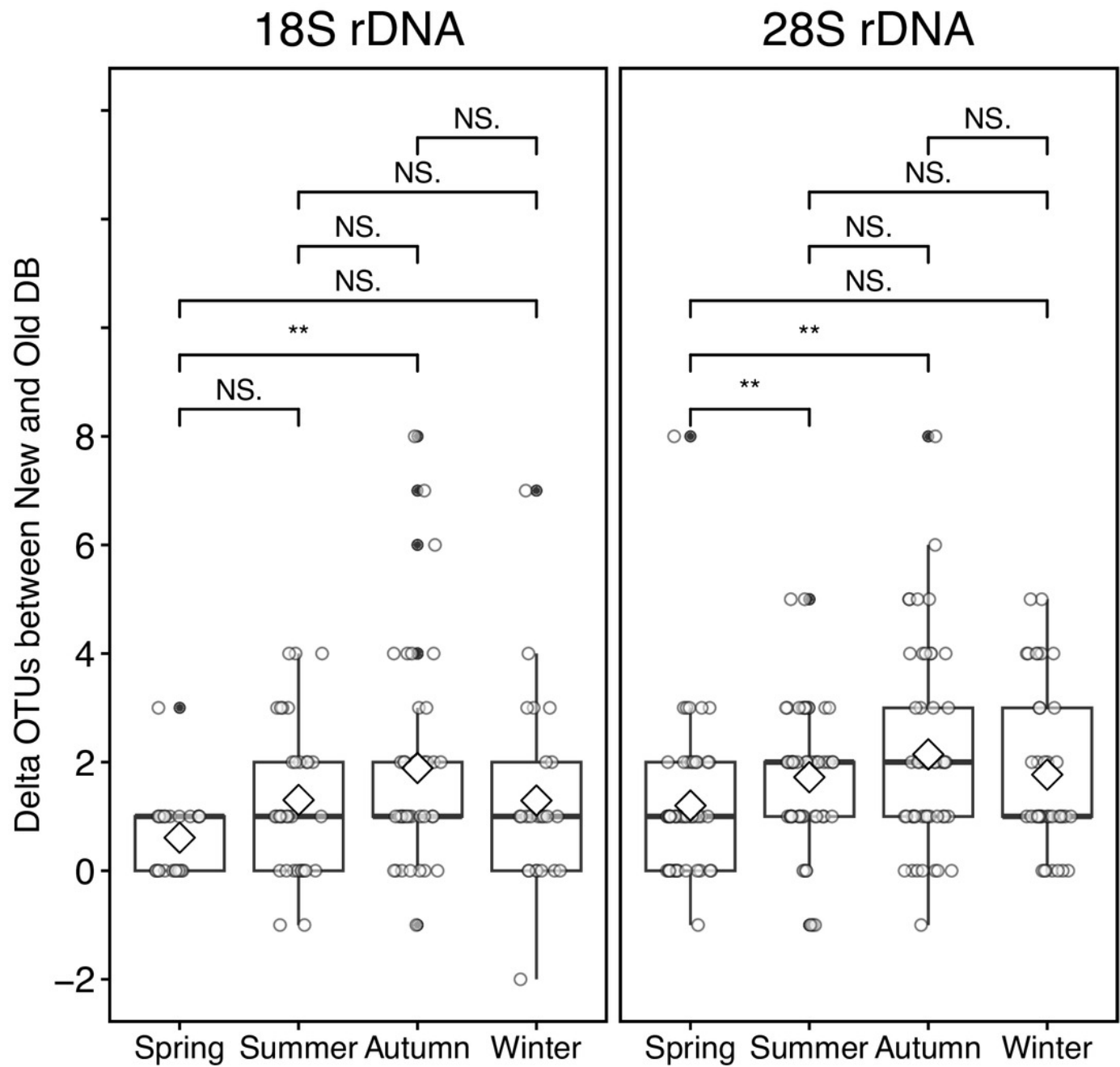



Figure 6

Boxplot of the difference (delta) of OTUs from rRNA genes metabarcoding samples between new and old reference databases by seasons at the Okhotsk Tower.

The delta OTUs from 18S (left) and 28S (right) rRNA gene sequences were compared between seasons. The seasons were separated by spring, March-May; summer, June-August; autumn, September-November; winter, December-February. Diamond (◇) indicates the mean value of the delta OTUs by season. Asterisk (**) means a significant difference ($p < 0.01$), and NS abbreviates no significant difference.

Sampling information for the locations of taxonomically verified sequences and metabarcoding of zooplankton.

Sample indicates which samples were used for analysis. Sea area shows the sampling area (and abbreviations) for this study. Depth shows the depth of the sampling station, not the net towing depth for zooplankton collecting. SST: Sea Surface Temperature. * See Supplemental Table S2 for details of the sampling data.

1   **Table 1:**
2   **Sampling information for data registration and metabarcoding of zooplankton.**
3   Sample indicates which samples were used for analysis. Sea area shows the sampling area (and
4   abbreviations) for this study. Depth shows the depth of the sampling station, not the net towing
5   depth for zooplankton collecting. SST: Sea Surface Temperature. * See Supplemental Table S2
6   for details of the sampling data.

Sample	Sea area	Cruise or location	Station	Longitude [°E]	Latitude [°N]	Depth [m]	Date	SST [℃]
Individual zooplankton	Sea of Okhotsk	HK1805 or WK1809	N1	144.12	44.22	-55	May and Sep 2018	8.5, 16.7
			N2	144.33	44.50	-734	May and Sep 2018	6.7, 12.2
			N3	144.75	45.00	-1,553	May and Sep 2018	7.2, 13.2
			N4	145.17	45.42	-2,977	May and Sep 2018	6.3
			S1	143.33	44.50	-57	May and Sep 2018	9.1, 17.4
			S2	143.83	44.50	-164	May and Sep 2018	8.1, 15.2
			S3	144.85	44.57	-1,655	May and Sep 2018	6.3, 13.4
			S4	145.37	44.63	-2,239	May and Sep 2018	8.4
	Sea of Japan	SHU1804	SI01	138.23	38.14	-130	18 Apr 2018	10.7
			SI07	136.96	40.00	-2,060	17 Apr 2018	7.7
			SI09	136.34	41.00	-3,310	17 Apr 2018	4.5
		TY1806	SI01	138.23	38.14	-130	13 Jun 2018	17.4
		TY1809	SI01	138.23	38.14	-130	21 Sep 2018	23.7
			SI03	137.92	38.58	-1,890	20 Sep 2018	23.4
		YK1801	St.89	130.24	35.00	-125	30 Apr 2018	16.1
	Cold Temperate Western North Pacific (CTWNP)	WK1805	A01	144.83	42.83	-94	18 May 2018	5.0
			A02	144.92	42.67	-764	18 May 2018	6.1
			A19	147.00	38.50	-5,438	13 May 2018	15.4
		WK1807	A01	144.83	42.83	-94	18 Jul 2018	14.3
			A02	144.92	42.67	-764	18 Jul 2018	15.3
			A05	145.25	42.00	-4,382	17 Jul 2018	15.4
			A21	147.25	38.00	-5,559	13 Jul 2018	21.4
		HK1809	A19	147.00	38.50	-5,438	3 Oct 2018	20.8
			A21	147.25	38.00	-5,559	2 Oct 2018	22.6
	Warm Temperate Western North Pacific (WTWNP)	SY1804	C2700	138.00	27.00	-4,864	23 Apr 2018	23.0
			C3300	138.00	33.00	-4,145	19 Apr 2018	18.1
			C3415	138.00	34.25	-824	18 Apr 2018	17.4
		SY1811	C3000D	138.00	30.00	-4,056	12 Nov 2018	26.0
			C3145D	138.00	31.75	-3,781	10 Nov 2018	25.2
			C3400N	138.00	34.00	-1,501	15 Nov 2018	21.5
			C3400D	138.00	34.00	-1,501	16 Nov 2018	21.5
	East China Sea	YK1801	St.74	127.23	31.75	-124	27 Apr 2018	18.5
		YK1805	St.24	127.28	31.75	-128	30 Jun 2018	24.3
		YK1806	St.9	128.75	31.75	-767	20 Jul 2018	29.6
	Nansei Islands		U5	124.22	24.47	-25	23 May, 25 Jun, 24 Jul, 21 Aug, 2 Oct, 6 Nov, 6 Dec 2018	24.5–29.6
			S11	124.05	24.28	-36	24 May, 8 Aug,	25.4–30.3
			N1	124.10	24.40	-55	15 Nov 2018	25.0–30.4
Metabarcoding	Sea of Okhotsk*	Mombetsu	Okhotsk Tower	143.38	44.34	-10	Apr 2012– Mar 2016	-1.7–22.2

Summary of the taxonomically verified and newly registered zooplankton sequences from this study.

'Sequence' indicates the number of sequences, and 'Species' indicates the number of groups by species which includes duplicate sequences of the same species from other specimens. The table denoted our registered sequence was classified to what taxonomic level on the public database by BLAST. 'Pre-exist' indicates a sequence with a complete match at a species level based on the public database, and 'New' a match with taxonomic criteria of genus or higher level (i.e., the sequences for which the species did not exist in the public database) database. Abbreviations: CTWNP, Cold Temperate Western North Pacific; WTWNP, Warm Temperate Western North Pacific.

1 **Table 2**:
2 **Summary of the taxonomically verified and newly registered zooplankton sequences from**
3 **this study.**

4 'Sequence' indicates the number of sequences, and 'Species' indicates the number of groups by
5 species which includes duplicate sequences of the same species from other specimens. The table
6 denoted our registered sequence was classified to what taxonomic level on the public database by
7 BLAST. 'Pre-exist' indicates a sequence with a complete match at a species level based on the
8 public database, and 'New' a match with taxonomic criteria of genus or higher level (i.e., the
9 sequences for which the species did not exist in the public database) database. Abbreviations:
10 CTWNP, Cold Temperate Western North Pacific; WTWNP, Warm Temperate Western North
11 Pacific.

rDNA	Sea area	Species				Sequences			
		Total	Pre-exist	New		Total	Pre-exist	New	
				genus	higher rank			genus	higher rank
18S	Total	94	28	55	11	162	49	100	13
	Sea of Okhotsk	23	11	10	2	24	12	10	2
	Sea of Japan	31	8	22	1	31	8	22	1
	CTWNP	29	11	17	1	29	11	17	1
	WTWNP	30	5	23	2	30	5	23	2
	East China Sea	18	4	14	0	19	4	15	0
	Nansei Islands	29	9	13	7	29	9	13	7
28S	Total	93	52	32	9	161	97	48	16
	Sea of Okhotsk	23	15	5	3	24	16	5	3
	Sea of Japan	31	20	8	3	31	20	8	3
	CTWNP	28	21	5	2	28	21	5	2
	WTWNP	30	18	9	3	30	18	9	3
	East China Sea	18	10	7	1	19	10	8	1
	Nansei Islands	29	12	13	4	29	12	13	4

Overview of OTUs associated with a single zooplankton species detected from metabarcoding samples based on the old and new reference databases.

Zooplankton composition of metabarcoding samples from Okhotsk Tower based on the reference databases with (new) and without (old) newly registered sequences, performed separately for the 18S and 28S rRNA gene sequences. 'Total' is the number of OTUs in the old and new databases combined; 'New' and 'Old' are the number of OTUs in each database. 'Difference' is the number of OTUs in the new database minus the number of OTUs in the old database, representing the increase in the number of detections in the new database. The OTU number detected in both databases ('Shared OTUs') or only in the respective database are also shown ('Unique OTUs'). However, most OTUs that appeared in the old database were also detected at the species level in the new database (see text for details). 'OTU per sample' shows the respective OTU number (mean, standard deviation, minimum, and maximum) per sample.

1 **Table 3**:
2 **Overview of zooplankton OTUs detected in metabarcoding samples based on old and new**
3 **reference databases.**

4 Zooplankton composition of metabarcoding samples from Okhotsk Tower based on the reference
5 databases with (new) and without (old) newly registered sequences, performed separately for the
6 18S and 28S rRNA gene sequences. 'Total' is the number of OTUs in the old and new databases
7 combined; 'New' and 'Old' are the number of OTUs in each database. 'Difference' is the number of
8 OTUs in the new database minus the number of OTUs in the old database, representing the increase
9 in the number of detections in the new database. The OTU number detected in both databases
10 ('Shared OTUs') or only in the respective database are also shown ('Unique OTUs'). However,
11 most OTUs that appeared in the old database were also detected at the species level in the new
12 database (see text for details). 'OTU per sample' shows the respective OTU number (mean,
13 standard deviation, minimum, and maximum) per sample.

rRNA	Number of detected OTUs		Shared OTUs	Unique OTUs	OTUs per sample			
					Mean	SD	Min	Max
18S	Total	98	73					
	New	92		19	15.6	10.2	2	53
	Old	79		6	14.2	8.9	2	50
	Difference	13			1.4	1.6	-2	8
28S	Total	109	44					
	New	89		45	12.1	4.7	3	26
	Old	64		20	10.5	4	1	21
	Difference	25			1.7	1.5	-1	8