Philippine Journal of Science
151 (3): 955-968, June 2022
ISSN 0031 - 7683
Date Received: 12 Oct 2021

In Silico Evaluation of the Taxonomic Resolution and Coverage of the COI Region and Alternative Barcode Markers for Echinoderms

Ardea M. Licuanan[1,2*] and **Ambrocio Melvin A. Matias**[1,2]

[1]Natural Sciences Research Institute, [2]Institute of Biology,
College of Science, University of the Philippines Diliman
Quezon City 1101 Philippines

DNA barcoding methods are potentially useful for the detection and identification of species difficult to identify using morphology-based taxonomy, such as the larvae of marine taxa. Combining barcoding methods with massively parallel sequencing and low-cost DNA library preparation improves the feasibility of processing several samples but limits the taxonomic resolution of resulting sequences. In light of these limitations, we looked for suitable barcode regions for echinoderm larvae, first by examining the cytochrome c oxidase subunit I (COI) region, the most commonly used barcode for echinoderms. Using a dataset containing over 4000 COI sequences representing nearly 400 echinoderm species across all five extant classes, the COI gene was found to lack a sufficiently conserved region for barcoding. Taxonomic resolution was also found lacking, with no clear barcode gap across classes and the results of Automatic Barcode Gap Discovery (ABGD) failing to reflect current taxonomic assignments. While sequences may be mislabeled, the aggregation of multiple species in one ABGD partition suggests that course-grained taxonomic resolution may still result from using COI alone. The search for short alternative mitochondrial regions across 110 genomes of distinct species using ecoPrimers software revealed that class-specific primers targeting 12S and 16S ribosomal RNA (rRNA) regions are prime candidates for barcoding echinoderm sequences. The results of *in silico* PCR across the 110 mitochondrial genomes indicate that compared to existing COI primers, the candidate 12S and 16S primers yield similar taxonomic coverage, with slightly lower resolution. Taking into consideration the previous analysis of COI, we suggest the use of multiple markers (COI, 12S, 16S) to adequately capture the understudied diversity of echinoderm larvae. *In vitro* PCR of the 12S and 16S primers is also needed to validate the results of the *in silico* analyses.

Keywords: DNA barcoding, echinoderms, *in silico* approach, primers

INTRODUCTION

Species detection and identification methods based on DNA barcode markers are becoming viable alternatives to traditional morphology-based approaches of species documentation. These approaches are especially useful when dealing with large amounts of samples that are difficult to identify, such as the larvae of marine taxa (Webb *et al.* 2006; Heimeier *et al*. 2010). However, the success of these methods is highly dependent on the barcode selected (Valentini *et al.* 2009). Specifically, the barcode regions targeted by primers must be sufficiently conserved within the target taxa so that barcodes can be retrieved from all relevant species represented in a sample

*Corresponding author: amlicuanan@up.edu.ph

(*i.e.* primers work for many taxa of interest). Moreover, the barcodes must be sufficiently differentiated between species to allow unambiguous species identification and discrimination (*i.e.* provide taxonomic resolution; Ficetola *et al*. 2010). Optimality criteria aside, the practicality of the method should also be considered. Standard barcoding typically involves per-individual amplification and sequencing, both of which are impractical for processing large quantities of samples (Kimmerling *et al.* 2018). Massively parallel sequencing (next-generation sequencing or NGS) in conjunction with low-cost DNA library preparation is a feasible alternative to individual Sanger-based sequencing (Glenn *et al.* 2019). Metabarcoding, which utilizes this technology, would allow simultaneous sequencing of barcodes from pooled samples (environmental DNA or bulk samples). Thus, species identification through metabarcoding is potentially useful for documenting the diversity of difficult taxa, provided that the selected barcodes have good taxonomic coverage and resolution.

The optimization of barcoding methods could benefit the study of marine invertebrate larvae, which are likely being undervalued as a food source in coral reefs because of difficulties in identification and biases towards sampling fish taxa (Riginos and Leis 2019). One of the invertebrate groups understudied in this respect is the phylum Echinodermata, whose species contribute a large portion of the epibenthic biomass in many marine habitats (McClintock 1994; Ellis and Rogers 2000). While echinoderm diversity can be subdivided into well-defined clades (Asteroidea, Crinoidea, Echinoidea, Holothuroidea, and Ophiuroidea), identification at deeper taxonomic levels is often difficult for larvae. At the family level, the identities of some echinoderm larvae are rarely discerned with certainty (Knott *et al.* 2003). The problem is further compounded by the difficulty in connecting the wide range of forms in the planktonic larval stage to the adult benthic stage, the phenotypic trophic plasticity exhibited by many larvae, as well as the considerable morphological similarities between the larvae of related species (Smith 1997; Heimeier *et al.* 2010; Wolfe *et al.* 2015). Since the identification of echinoderm larvae using morphological traits alone is challenging, barcoding methods could be useful for investigating the diversity and distribution of larval pools.

A number of standard barcoding studies have used the COI gene as the marker for exploring echinoderm diversity (Folmer *et al.* 1994; Arndt *et al.* 1996; Ward *et al.* 2008; Hoareau and Boissin 2010; Layton *et al.* 2016). Although some find that COI generally satisfies the criteria of primer universality, others report zero to low amplification success in select species (Folmer *et al.* 1994; Hoareau and Boissin 2010). Problems about alignment within the region have also arisen, with combined alignments for some species exhibiting gaps despite the absence of stop codons in the sequences aligned (Laakmann *et al.* 2016). These issues call into question the applicability of COI for barcoding surveys, especially if NGS is involved in the approach. While NGS improves the feasibility of processing several samples, many sequencing platforms are limited by short read lengths and will generate short barcodes, potentially limiting the taxonomic resolution of the resulting sequences. Thus, existing barcoding methods for echinoderms need to be reevaluated in light of the limitations of NGS.

As part of the effort to better document the diversity and distribution of understudied taxa like echinoderm larvae, the present study has three objectives. First, we aim to evaluate the suitability of the COI region for barcoding echinoderms by examining its taxonomic resolution and coverage. We focus on COI barcodes from species in the Indo-West Pacific region, the most species-diverse marine region in the world (Veron *et al.* 2009). Second, we search for other mitochondrial regions that can be used as alternatives to COI. Using ecoPrimers and ecoPCR software, we generate class-specific primers for short barcodes and test the performance of these primers *in silico* (Ficetola *et al.* 2010; Riaz *et al.* 2011). Lastly, we compare the class-specific primers to existing barcoding primers in terms of taxonomic coverage and resolution. Thus, considering evaluations of barcode suitability, we find and test class-specific primers potentially useful for barcoding surveys of echinoderm larvae.

MATERIALS AND METHODS

Examining the Resolution of the COI Region

Because COI is still one of the most employed genetic regions for DNA barcoding of echinoderms (*i.e.* approach for species identification), we first examined the resolution of COI by utilizing available genetic data from different databases. In examining the resolution, we focused on echinoderms found in the Indo-West Pacific region. A list of Indo-Pacific Echinodermata species was obtained from a Global Biodiversity Information Facility occurrence dataset, which was downloaded as a Darwin Core Archive File on 22 Apr 2020 (GBIF 2020). This dataset included occurrences identified at least up to the species level based on the Darwin Core term *taxonRank* (Wieczorek *et al.* 2012; see https://dwc.tdwg.org/terms/). A simplified species list was extracted from the *scientificName* term by excluding authorship and date information. Sequences containing the COI region for this set of taxa were retrieved from the NCBI Nucleotide database using the rentrez v1.2.2 package for R (Winter 2017).

The downloaded COI sequences were aligned by class using Multiple Sequence Alignment with External Applications (MUSCLE) v3.8.31 through the ape v5.3 package for R (Edgar 2004; Paradis and Schliep 2019). Each class alignment was read into R v3.6.1 as a matrix (R Core Team 2019). To maximize the number and length of sequences to be included in the downstream analyses, sequences less than 550 bases in length were excluded. The remaining sequences were sorted based on the position at which the first base occurred, and were visualized as colored lines using the raster v3.3.13 R package to identify large gaps (Hijmans 2020). If large gaps were present, they were used to group sequences together for separate realignment. To remove the remaining gaps after realignment, positions towards the center of the alignment with less than 100 base reads were identified. Sequences that had base reads at these positions were removed, then the remaining sequences were realigned. All resulting gap-free alignments were trimmed at both ends, with final sequence alignment lengths being at least 559 bases long.

The taxonomic resolution or resolution capacity of a barcode refers to the ability of the marker to accurately distinguish different species based on interspecific differences in DNA sequences (Ficetola *et al.* 2010). We investigated the taxonomic resolution of the COI by employing classical DNA barcoding, which is reliant on a barcoding gap identified based on prior taxonomic information. Because species misidentification can heavily influence the barcoding gap, we also employed ABGD, which does not use prior taxonomic information. In investigating the barcoding gaps for the group of sequences that we analyzed, the pairwise genetic distances between sequences in each alignment were calculated with ape v5.3 using the Kimura 2-Parameter substitution model (K2P; Kimura 1980), as used by Hebert *et al.* (2003a). Frequency distributions for the genetic distances of intraspecies, interspecies, and intergeneric pairwise comparisons were generated and examined for overlaps. An expected arbitrary limit between intra- and interspecies divergence was also calculated by multiplying the average of nonzero intraspecific distances by 10 (Hebert *et al.* 2004).

The barcode gap for each alignment was also examined using the ABGD method. ABGD calculates a model-based confidence limit for intraspecific divergence based on a given prior limit to intraspecific diversity (P) and minimum gap width (X). A barcode gap is inferred from this limit, which is used to partition a given set of sequences into preliminary operational taxonomic units (OTUs). Final partitions are obtained after the method is recursively applied to preliminary groupings (Puillandre *et al.* 2012). Partitions for each alignment were determined using P values 10 steps across 0.001 and 0.1 and a value of 1.5 for X. K2P genetic distances were reported based on a substitution model with a transition/transversion ratio of 2. For each alignment, the graph of ranked pairwise differences was inspected and the *P*-value corresponding to the first steep slope was determined. The final partitions per alignment were selected based on this optimal *P*.

Searching for Alternative Regions for Barcoding

The ideal DNA barcode has the capacity to discriminate between closely related species but is flanked by highly conserved regions so that the barcode can be retrieved across a wide range of taxa. High-throughput sequencing platforms like Illumina have limited read lengths; thus, in addition to the two traits mentioned, a barcode to be used for metabarcoding applications must be short. Because we were unable to find a suitable barcode within the COI region across all echinoderm sequences included in the analyses. Searches for other potentially suitable metabarcodes in the mitochondrial genome were conducted using ecoPrimers v0.3 (Riaz *et al.* 2011). To this end, complete mitochondrial genome sequences of Echinodermata were downloaded from GenBank. The OBITools python package was used to annotate the sequences by their species taxid and to select a single sequence per species to reduce overrepresentation in the reference database (Boyer *et al.* 2016). The resulting database was used to generate class-specific primers for retrieving barcodes, which are 200–280 bases long. A maximum of three mismatches between each primer and its target sequence was allowed. In addition, two nucleotides on the 3'-end of each primer were required to strictly match the target sequence.

Up to 10 primer pairs per class were selected from the output generated by ecoPrimers. The minimum requirements for a primer pair to be selected were as follows: [1] the expected melting temperatures (without mismatch) must be between 52–61 °C; [2] the melting temperatures of the forward and reverse primers are within 5 °C of each other, and [3] each primer has a GC content between 40–70%. Primer pairs with high taxonomic coverage and resolution capacity – measured by the B_c and B_s indices, respectively – were also preferred (Ficetola *et al.* 2010).

***In Silico* Testing of Candidate Primers for Barcoding Echinoderms**

To evaluate the quality of barcodes targeted by the newly generated class-specific primers, *in silico* PCR was performed on the reference database using the ecoPCR v0.2 program (Ficetola *et al.* 2010). Amplicon lengths were restricted between 200–280 bases, ideal fragment lengths for Illumina sequencing. Additionally, a maximum of five mismatches between each primer and its target sequence was allowed. The output of ecoPCR lists all

barcodes expected to be retrieved by a specified primer pair. To tentatively identify the mitochondrial regions where barcodes were located, sequence comparisons were performed between the amplicons and a custom nucleotide database of annotated mitochondrial genomes representing each of the five classes (Appendix Table I). Alignments were carried out using Nucleotide-Nucleotide BLAST 2.7.1+ (Altschul *et al.* 1990).

The outputs of ecoPCR were used to calculate the taxonomic coverage (coverage index B_c) and taxonomic resolution (specificity index B_s) associated with each primer pair at the species rank. Taxonomic coverage is measured as a ratio of the number of target species recovered over the number of target species in the input database. On the other hand, taxonomic resolution is the number of taxa in the input database unambiguously identified by a barcode over the number of target species recovered (Ficetola *et al.* 2010). The B_c and B_s indices were computed through the ecoTaxStat and ecoTaxSpecificity scripts, respectively, of the OBITools package (Boyer *et al.* 2016). For the ecoTaxSpecificity script, the maximum number of base errors between two sequences for them to be considered the same barcode – the error parameter – was set to five. This number was based on the product of the average amplicon lengths (per primer pair) with 0.021, the intraspecies divergence limit obtained for the COI sequences using ABGD.

Comparison of New Primers to Existing Primers

To evaluate the performance of the new primers against metazoan universal primers and existing primers for echinoderms (Appendix Table II), *in silico* PCR was performed using the existing primers. For the parameters of ecoPCR, amplicon lengths were limited to 400–1000 bases (up to 1600 bases for the primer targeting the 16S region), and a maximum of five mismatches between primer and target sequence was allowed. B_c and B_s indices were computed for the generated amplicons. The error parameter for the specificity index was based on average amplicon length scaled by the previously determined intraspecific divergence limit. For full-length barcodes, the error parameter for ecoTaxSpecificity ranged from 13–18. Taxonomic resolution was also evaluated for the following cases: [1] only the first 150 bases each barcode are retained, and [2] the first 150 bases are merged with the last 150 bases of each barcode. These truncated barcodes simulate the typical sequence length of Illumina next-generation sequencing. Error parameters of 3 and 6, respectively, were used in computing the specificity index associated with the truncated barcodes. The new primers were compared with existing primers in terms of taxonomic coverage and resolution.

RESULTS

Description of the COI Sequence Dataset Analyzed

The GBIF dataset from which a species list was obtained contained records of 3009 Echinodermata species in the Indo-West Pacific region. There were also records for 145 species further categorized to either subspecies, variety, or form. However, no COI sequences were retrieved for taxa identified at ranks more specific than the species level. At the species level, 518 of the 3009 species were linked to COI sequences. A total of 6601 sequences containing the COI region were downloaded from the NCBI Nucleotide database, with lengths ranging from 205–1942 bases.

Initial attempts to align all 6601 sequences resulted in extremely patchy alignments – that is, alignments were frequently punctuated by gaps of varied sizes. These gaps were also present in some alignments performed by class, although less frequently. In classes Asteroidea, Echinoidea, and Ophiuroidea, gaps common across many base positions were used to further split alignments into smaller groups. Asteroidea sequences were divided into five sequence groups, whereas Echinoidea and Ophiuroidea alignments were each split into two. After splitting alignments and filtering out sequences, the 6601 sequences (518 species) were reduced to 4383 sequences (399 species). Eleven (11) clean alignments were generated, each containing 52–889 sequences and representing between 1–123 species (Appendix Table III), and alignment lengths ranged from 559–706 bases. These 11 groups were the basis for examining the resolution of the COI region. Their positions along the COI region of *Ophiura lutkeni*'s mitogenome, also used by Hoareau and Boissin (2010) as a reference sequence, are shown in Figure 1.

Species Delineation Using Classical Barcoding and ABGD

Species delineation through classical DNA barcoding is effective when intraspecific variation within the barcode gene is smaller than interspecific variation. To quantify variation, Kimura 2-Parameter genetic distances were calculated within and between species. Across the 11 sequence groups, intraspecies distances ranged from 0–37.4%, with a mean of 1.8%. Interspecies divergence averaged at 13.5%, nearly eight-fold greater than the mean intraspecies divergence. However, in eight of nine groups with interspecies comparisons, the maximum intraspecies distance exceeded the minimum interspecific distance. Interspecies divergence ranged from 0–39.8%. Comparisons with 0% interspecies divergence were found in Asteroidea (*Psilaster acuminatus vs. P. charcoti* and among some members of *Luidia*), Crinoidea (*Phanogenia typica vs. P. gracilis* and *Comaster schlegelii vs. C.*

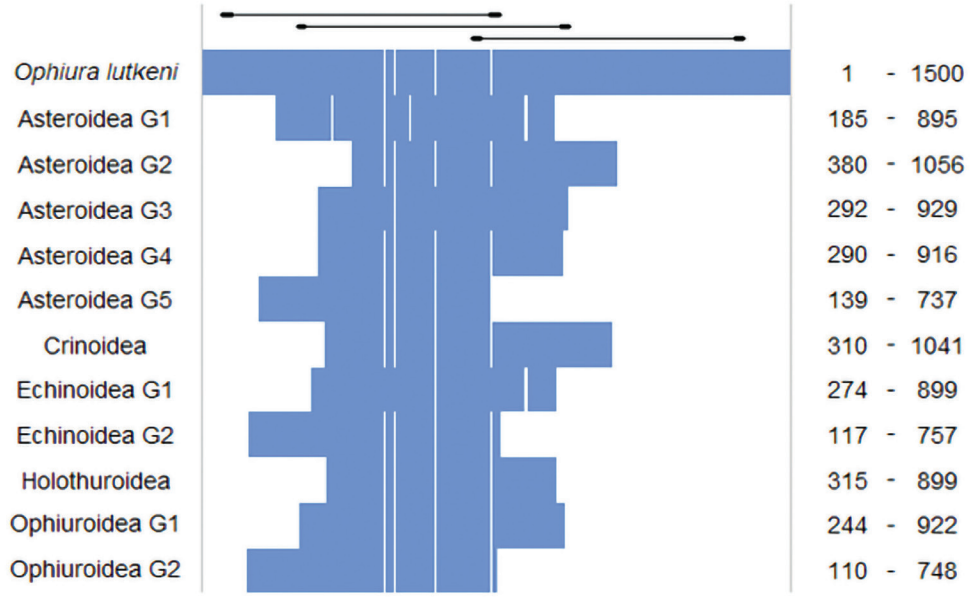

Figure 1. Positions of representative sequences from each sequence group along the first 1500 base pairs of the *Ophiura lutkeni* mitochondrial genome (accession AY184223). The first 1605 bases of the reference genome corresponds to the COI gene. Positions of existing primers by Folmer *et al.* (1994), Arndt *et al.* (1996), and Palumbi (1996) are presented at the top of the figure in that order.

audax), and Holothuroidea (*Stichopus horrens vs. S. monotuberculatus*). As for comparisons between genera, pairwise distances across groups ranged from 0.2–50.2%, with a mean intergeneric divergence of 21.7%. Extensive overlaps between the distribution of interspecific (or intraspecific) and intergeneric distances were found in Echinoidea, Holothuroidea, and Ophiuroidea (Appendix Figure I).

Ideally, intraspecific and interspecific variations are sufficiently different so that thresholds can be used in classifying specimens into species. For example, the "10x rule" proposes that pairwise distances exceeding 10 times the intraspecific variation are found between different species (Hebert *et al.* 2004). Across the sequence groups with multiple species, values for the 10x-threshold ranged from 3.3–38.1%, with a mean of 23.0%. Except for some groupings within Asteroidea and the Holothuroidea group, the 10x-threshold was not effective in delineating different species (Appendix Figure I). As the 10x-threshold itself is based on current taxonomic information, incorrect classification biases the presumed location of the barcoding gap. To preclude this bias, sequence groups were also subjected to ABGD. In each of the 11 sequence groups, the relations between partitions generated by ABGD and the species information of the sequences were assessed. If all the sequences of one species exclusively fell under one partition, the sequences were presumed to have correct species identities. Generally, the partitions generated by ABGD did not consistently reflect current taxonomic assignments. The only exception was an Asteroidea group, but this consisted of sequences representing only two species. For the rest of the sequence groups, the proportion of correctly partitioned species ranged from 0–80%.

Species whose collective sequences did not form one-to-one correspondence with ABGD partitions were categorized into three: candidates for splitting, candidates for collapse into more inclusive groups with other species, or species involved in an ambiguous grouping. Candidates for splitting were those whose sequences monopolized two or more partitions. Candidates for collapse were species whose sequences fell under one partition and shared it with species whose sequences grouped similarly. Species with ambiguous grouping were those with sequences that split into partitions but shared at least one partition with other species. Assuming the captured sequence information reflects species identity, species categorized under this last group may have misidentified sequences.

Candidates for splitting were found in all classes except Crinoidea, with the number of candidates ranging from 3–21 species per class (Figure 2). Thus, for a handful of species, the ABGD approach captures variability in the COI region not reflected in current species assignments. While a total of 43 species were candidates for splitting, a total of 91 species were candidates for collapse into more inclusive groups. Candidates for collapse were found across all classes. Most collapsible groups contained 2–5 species. One notable exception was the sequence group

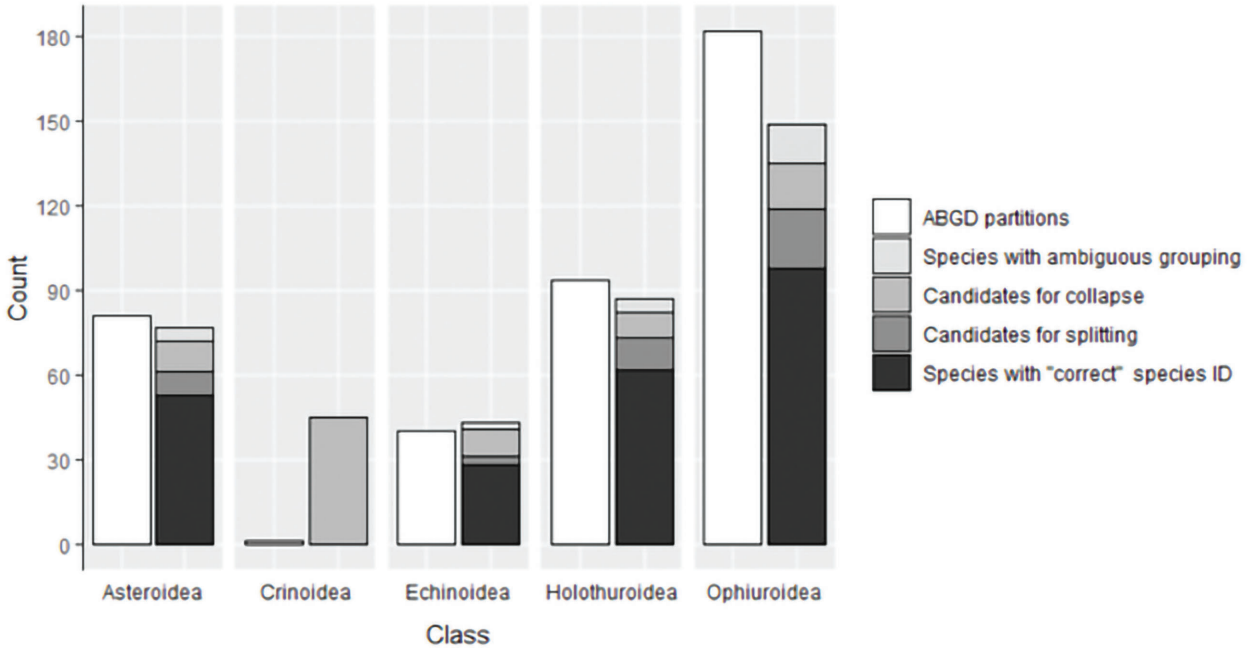

Figure 2. The non-congruence between the number of ABGD partitions found and the number of species examined per class. Species count is further categorized into counts for [1] species with one-to-one correspondence with ABGD partitions, [2] candidates for splitting, [3] candidates for collapse into more inclusive groups with other species, and [4] species involved in ambiguous groupings.

of Crinoidea that represented 45 different species, all of which were grouped into one ABGD partition.

Alternative Regions for Barcoding

A region within the COI gene with good taxonomic resolution and coverage was not found across the sequences examined. Thus, we sought alternative regions in the mitochondrial genome for barcoding echinoderms. To this end, a reference database of echinoderm mitochondrial genomes was created. The database consisted of 110 mitochondrial genomes, which represented 16 Asteroidea species, six Crinoidea species, 37 Echinoidea species, 25 Holothuroidea species, and 26 Ophiuroidea species. The ecoPrimers software could not identify primers universal across all 110 genomes, so class-specific primers were generated instead. A total of 3809 unique 18-base pair (bp) primer pairs were identified. Each pair was given a primer ID indicating the class for which they were generated (Asteroidea: "ast", Crinoidea: "cri", Echinoidea: "ech", Holothuroidea: "hol", Ophiuroidea: "oph") and a serial number.

Candidate primers pairs were selected per class on the basis of melting temperature (T_m), GC content, and initial measures for taxonomic coverage (B_c) and resolution capacity (B_s) (Ficetola *et al.* 2010). Of the 3809 primer pairs, 26 showed good potential with respect to the criteria above (Table 1). Melting temperatures for the candidate primers ranged from 52–61 °C, whereas GC content ranged from 44–61% and averaged at 56%. The primers targeted regions 201-bp to 262-bp long.

Almost all candidate primers retrieved regions within non-protein-coding sequences. Primers generated for Crinoidea, Holothuroidea, and Ophiuroidea targeted sequences within the rRNA 16S gene, while primers generated for Asteroidea and Echinoidea targeted sequences within the rRNA 12S gene. One Echinoidea primer – namely, ech.11 – targeted sequences within both the 12S gene and the COI gene (Table 1).

***In Silico* Testing of Candidate and Existing Primers**

The 26 candidate class-specific primers were tested over the entire reference database of 110 mitochondrial genomes. All primers targeted sequences of species outside the class they were generated for, with varying levels of coverage (Figure 3). Generally, sequences of species under Crinoidea, Echinoidea, and Holothuroidea were recovered uniformly by the candidate primers. Sequences under Asteroidea and Ophiuroidea were not as easily retrieved *in silico*, especially by cri and hol primers. Among existing primers, the primer combination arndt1 performed best in terms of taxonomic coverage (Figure 3; Appendix Table II). A handful of ast primers (ast.85, ast.86, ast.93, ast.94) and ech primers (ech.9, ech.10, ech.11, ech.12) performed similarly.

Table 1. Candidate class-specific primers with potential for amplifying short barcode regions in mitochondrial genomes of echinoderms.

Primer ID	Sequences (5'-3')		T_m (°C)		Mean amplicon length (bp)	Region
	Forward (F)	Reverse (R)	F	R		
ast.101	AACCGCCAAGTCCTTTGA	GCCACCGCGGTTATACAT	57.3	58.4	230.5	12S rRNA
ast.102	AAACCGCCAAGTCCTTTG	GCCACCGCGGTTATACAT	56.6	58.4	231.5	12S rRNA
ast.13	AACCGCCAAGTCCTTTGA	ATTATGTGCCAGCCACCG	57.3	58.6	241.5	12S rRNA
ast.14	AAACCGCCAAGTCCTTTG	ATTATGTGCCAGCCACCG	56.6	58.6	242.5	12S rRNA
ast.85	AACCGCCAAGTCCTTTGA	CAGCCACCGCGGTTATAC	57.3	59.3	232.5	12S rRNA
ast.86	AAACCGCCAAGTCCTTTG	CAGCCACCGCGGTTATAC	56.6	59.3	233.5	12S rRNA
ast.93	AACCGCCAAGTCCTTTGA	AGCCACCGCGGTTATACA	57.3	59.3	231.5	12S rRNA
ast.94	AAACCGCCAAGTCCTTTG	AGCCACCGCGGTTATACA	56.6	59.3	232.5	12S rRNA
cri.1591	CTGACCTGACTTGCGTCG	GGGGCAACCACGAAGAAA	59.2	58.7	244.5	16S rRNA
cri.1592	CTGACCTGACTTGCGTCG	TGGGGCAACCACGAAGAA	59.2	59.7	245.5	16S rRNA
cri.1593	CTGACCTGACTTGCGTCG	TTGGGGCAACCACGAAGA	59.2	59.7	246.5	16S rRNA
cri.1651	CTGACTTGCGTCGGTCTG	GGGGCAACCACGAAGAAA	59.2	58.7	239.5	16S rRNA
cri.1652	CTGACTTGCGTCGGTCTG	TGGGGCAACCACGAAGAA	59.2	59.7	240.5	16S rRNA
cri.1653	CTGACTTGCGTCGGTCTG	TTGGGGCAACCACGAAGA	59.2	59.7	241.5	16S rRNA
cri.1736	GCGTCGGTCTGAACTCAG	TGGGGCAACCACGAAGAA	58.9	59.7	233.5	16S rRNA
cri.1737	GCGTCGGTCTGAACTCAG	TTGGGGCAACCACGAAGA	58.9	59.7	234.5	16S rRNA
cri.2137	CGGTCCAACATCGAGGTC	GAAGACCCTGTCGAGCTT	58.5	57.1	245.83	16S rRNA
cri.2147	GAAGACCCTGTCGAGCTT	GGTCCAACATCGAGGTCG	57.1	58.5	244.83	16S rRNA
ech.10	AGCCACCGCGGTTATACG	TGGAAAACCGCCAAGTCC	60.7	58.7	229.41	12S rRNA
ech.11	GCCACCGCGGTTATACGT	GGAAAACCGCCAAGTCCT	61	58.3	227.41	12S rRNA, COX1
ech.12	GCCACCGCGGTTATACGT	TGGAAAACCGCCAAGTCC	61	58.7	228.41	12S rRNA
ech.9	AGCCACCGCGGTTATACG	GGAAAACCGCCAAGTCCT	60.7	58.3	228.41	12S rRNA
hol.11	AGACGAGAAGACCCTGTC	CAATCCAACATCGAGGTC	56.1	53.8	222.6	16S rRNA
hol.12	AGACGAGAAGACCCTGTC	CCCAATCCAACATCGAGG	56.1	56.1	224.6	16S rRNA
hol.13	AGACGAGAAGACCCTGTC	CCCCAATCCAACATCGAG	56.1	56.1	225.6	16S rRNA
oph.11	ATAGGGTCTTCTCGTCCC	CGCCTGTTTACCTAAAAC	55.5	52	203.33	16S rRNA

Notably, the universal metazoan primers designed by Folmer *et al.* (1994) – which target the same region as primers by Ward *et al.* (2008) and Layton *et al.* (2016) (see Figure 1) – did not recover as many sequences as all other primers. Also, two of the nine existing primers tested did not recover any sequences in the database at all. These primers were the 16S primer pair identified by Arndt *et al*. (1996) and the COI primer pair identified by Hoareau and Boissin (2010). The latter was specifically designed for COI amplification in the Echinodermata and targets the same region as the COI primer of Arndt *et al.* (1996).

In addition to their performance in recovering sequences from different echinoderm classes, the 26 candidate primers were evaluated based on two indices defined by Ficetola *et al.* (2010) – the coverage index B_c and the resolution index B_s. A measure of taxonomic coverage, $B_{c,}$ is the proportion of echinoderm species recovered *in silico* in the database of 110 species. On the other hand, $B_{s,}$ a measure of resolution capacity, is the proportion of species accurately identified by their barcodes in the sequences retrieved by a primer pair. A species is defined to be unambiguously identified by a primer pair if it can be mapped to a barcode region amplified by the primer pair that is not shared by other species (Ficetola *et al.* 2010).

For the candidate primers, values for B_c ranged from 60.9–100%, whereas values for B_s ranged from 76.8–82.7% (Figure 4). The best performing primers were ast, ech, and oph pairs. Primers ast.85, ast.86, ast.93, and ast.94 recovered all species in the database (B_c = 100%). These primers also had the highest resolution capacity (B_s =

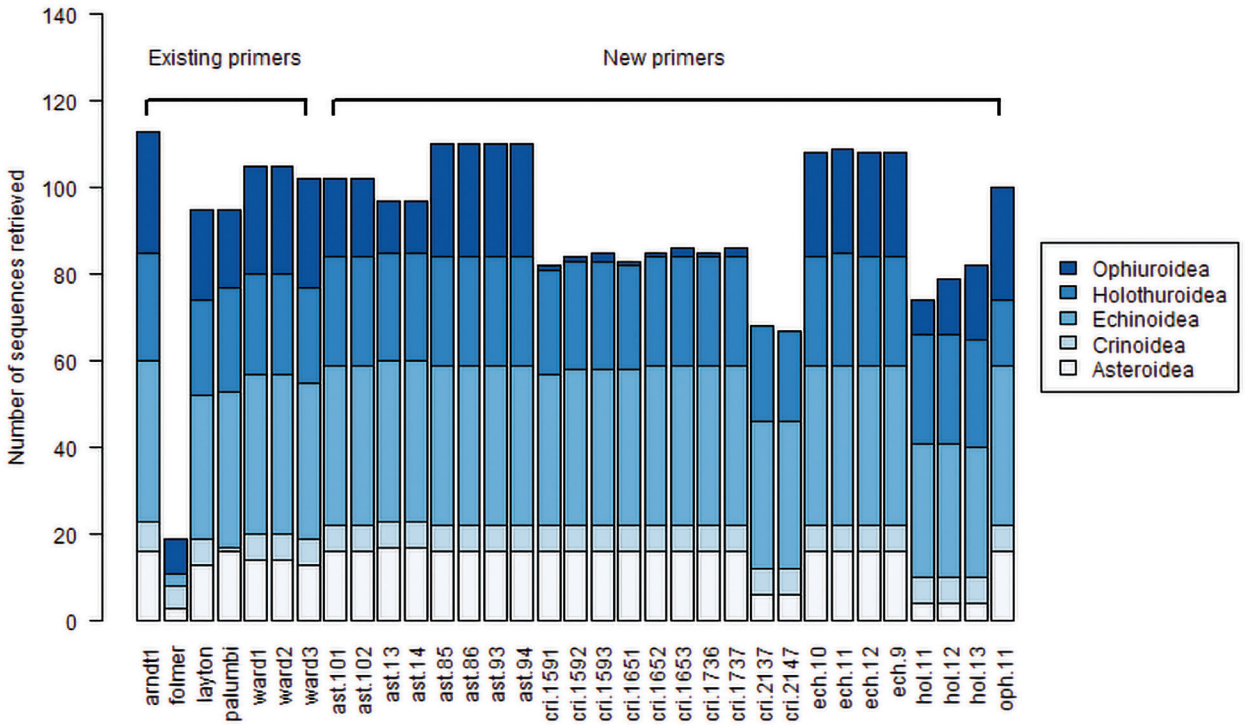

Figure 3. The taxonomic coverage of candidate primers beyond the classes they were generated for. Several candidate primers performed similarly to the best-performing existing COI primers in terms of taxonomic coverage.

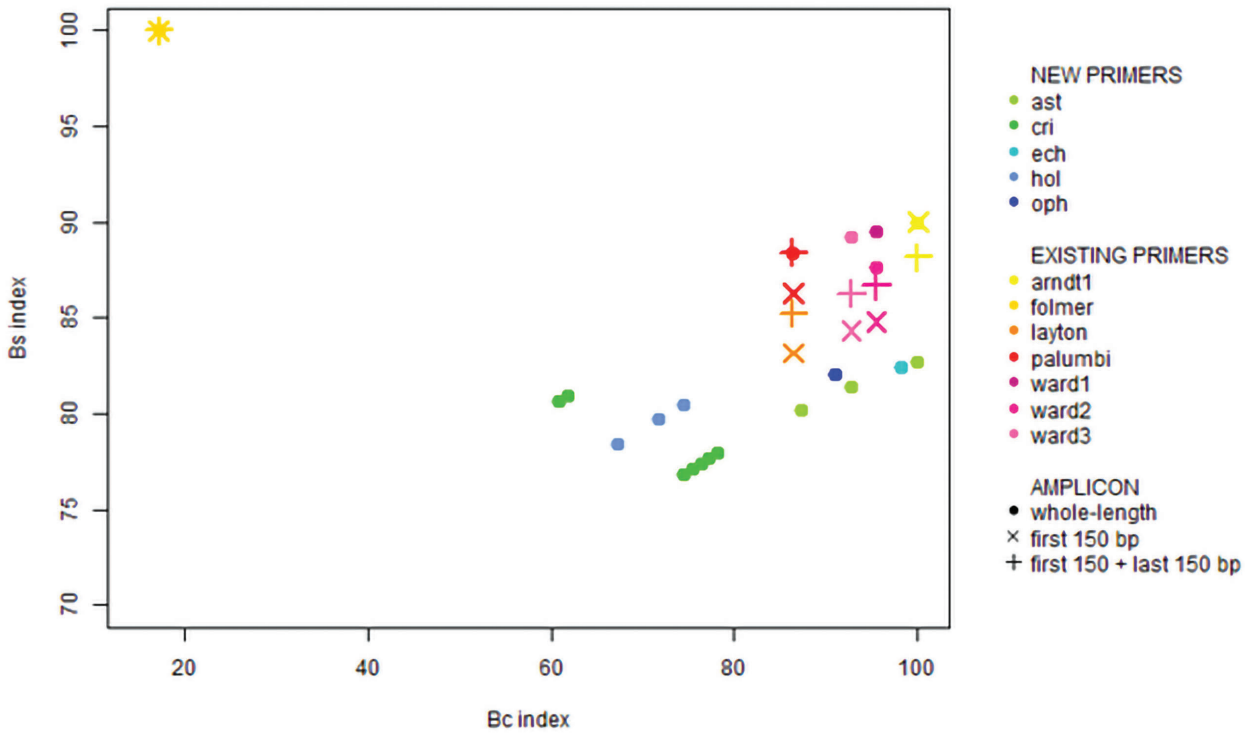

Figure 4. The comparable performance of existing and candidate primers in terms of indices proposed by Ficetola *et al.* (2010) for taxonomic coverage (B_c) and resolution capacity (B_s).

82.7%) among the candidate primers. All ech primers (ech.9, ech.10, ech.11, and ech.12) also had high values for B_c and B_s at 98.18% and 82.41%, respectively. The above-mentioned ast and ech primers mostly targeted barcodes within the 12S rRNA gene. The best performing primer which targeted barcodes within the 16S rRNA gene was oph.11. Its taxonomic coverage was at 90.9%, whereas resolution capacity was at 82%.

The taxonomic coverage of the best performing candidate primers was very comparable with that of existing COI primers (Figure 4). The B_c index values for some candidate primers exceeded those of primers designed by Folmer *et al.* (1994), Palumbi (1996), and Layton *et al.* (2016). In terms of resolution capacity, existing primers were unmatched by any of the candidate primers, likely due to the differences in preset amplicon lengths. Existing primers were designed to retrieve barcodes at least 600-bp long. In the *in silico* PCR, amplicon lengths recovered by the existing primers ranged from 579–969 base pairs. Amplicon lengths for the candidate primers were much shorter, ranging from 199–280 base pairs. Truncating the amplicons retrieved by existing primers to either 150 or 300 base pairs mostly resulted in lower B_s indices. These still had higher resolution capacity than candidate primers, but target the COI region instead of 12S or 16S rRNA genes.

DISCUSSION

To be effective in detecting and identifying species, larval surveys that employ barcoding should retrieve sequences with good taxonomic coverage and resolution. We evaluated the commonly used COI region and found it lacking in both criteria for echinoderms. In assessing taxonomic resolution, classical barcoding and ABGD indicated that the COI region alone may not consistently delineate different echinoderm species. We searched for alternative regions in the mitochondrial genome of echinoderm species and found that class-specific primers targeting short 12S and 16S ribosomal regions can potentially be used with metabarcoding or massively parallel sequencing. *In silico* testing over a database of 110 genomes showed these class-specific primers perform similarly to existing COI primers. The results suggest barcoding surveys for echinoderms could benefit from using a combination of markers, targeting regions within the COI gene plus 12S and 16S rRNA genes.

Most DNA barcoding studies for echinoderms have adopted the COI gene (Arndt *et al.* 1996; Ward *et al.* 2008; Hoareau and Boissin 2010; Layton *et al.* 2016) and so have metabarcoding surveys for fish larvae and planktonic communities (Kimmerling *et al.* 2018; Garcia-Vazquez *et al.* 2021). Hence, the applicability of COI for metabarcoding echinoderm larvae warranted closer examination. By attempting to align over 6,000 COI barcodes (representing over 500 echinoderm species) from the NCBI nucleotide database, we found that there was no region within the COI gene universally conserved across all sequences. This result was validated with the ecoPrimers software, which did not yield any primers universal across all 110 echinoderm mitochondrial genomes. Lack of universal regions within COI has been observed in other taxa and is attributed to the high level of nucleotide diversity within the gene (Deagle *et al.* 2014). COI divergence rate estimates for echinoderms do not exceed that of other marine invertebrate groups (mollusks, arthropods, polychaetes) (Loeza-Quintana *et al.* 2019), but even for taxa with lower percentage sequence divergence (*e.g.* vertebrates), the taxonomic coverage yielded by COI primers can be low (Hebert *et al.* 2003b; Ficetola *et al.* 2010). Folmer primers, designed to amplify COI from metazoan invertebrates, are widely used but also frequently fail (Folmer *et al.* 1994; Geller *et al.* 2013). We observed as much in the present study, with Folmer primers recovering the least number of amplicons from *in silico* PCR. Compared to the nuclear 18S gene or mitochondrial 16S and 12S genes, primer-binding sites in COI are more variable, necessitating degeneracy in primers like Folmer's to account for mismatches with the target position (Geller *et al.* 2013). Lack of universality risks unreliable amplification, which is not ideal for investigatory surveys of larval composition, especially if samples are pooled and taxonomic assignments cannot be performed per individual larva.

In examining the taxonomic resolution of COI, average intraspecific (1.8%) and interspecific (13.5%) genetic distances were found to be roughly in the same order of magnitude as those previously obtained for echinoderms: past barcoding studies report mean intraspecies divergence between 0.6–1.3% and mean interspecies divergence between 12.0–20.6% (Ward *et al.* 2008; Hoareau and Boissin 2010; Layton *et al.* 2016). In the present study, we found that in most sequence groups, the maximum intraspecific divergence exceeded the minimum interspecific divergence – that is, no clear barcoding gap was found through classical barcoding. As overlaps in the distributions of intraspecific and interspecific divergences may be caused by the taxonomic misidentification of sequences, ABGD was also conducted. ABGD, which partitions sequences into OTUs and lessens the bias of current taxonomic assignments, showed that in many cases COI divergences grouped sequences with different species labels into one OTU. An extreme case of sequences collapsing into one OTU was found in the Crinoidea, the most basal class in Echinodermata, where sequences representing 45 different species (based on

the taxonomic labels of the sequences) formed one ABGD partition. While discordances between ABGD partitions and current species assignments may be due to mislabeled sequences, this is unlikely to be the case for all Crinoidea sequences, given the relatively low rate of taxonomic errors in GenBank. For example, at the genus level, the error rate is likely < 1% (Leray *et al.* 2019). The aggregation of multiple species into more inclusive ABGD partitions suggests that variation within the COI region examined may be insufficient to delineate some species based on sequence information alone. Thus, course-grained taxonomic resolution may result from larval metabarcoding surveys that use COI alone as the marker.

The search for alternative mitochondrial regions using ecoPrimers software showed that non-protein-coding 12S and 16S rRNA genes contain candidate barcodes that maximize taxonomic coverage and resolution despite barcode lengths being limited to 200–280 base pairs. These rRNA genes have slower evolutionary rates than the COI gene and contain relatively more conserved regions optimal for primer design (Machida and Tsuda 2010) but offer a similar taxonomic resolution to COI (Deagle *et al.* 2014). Expectedly, the candidate primers targeting short 12S or 16S barcodes yielded lower taxonomic resolution than the existing primers tested, which were designed for recovering COI barcodes greater than 500-bp long. When we simulated the outcome of next-generation sequencing using COI primers (*e.g.* the first 150-bp or the first 150-bp joined with the last 150-bp), the taxonomic resolution of the shorter COI barcodes still exceeded that of 12S or 16S barcodes (Figure 4). However, this measure of taxonomic resolution may change if more mitochondrial genomes are available for *in silico* PCR. At present, comparisons of the resolution were done based on 110 genomes available (110 species). The coverage of species per class in this genome database is not as comprehensive as that of the COI sequences examined. As discussed earlier, our previous analyses of the COI region based on 4383 COI sequences (399 species) indicated that the COI region may not be sufficient, especially in terms of taxonomic coverage. Hence, our candidate 12S and 16S primers could potentially recover species information that COI primers may fail to detect. Studies wherein a single barcoding marker cannot sufficiently capture the diversity and breadth of the target taxon commonly suggest the use of multiple markers (Collins *et al*. 2019; van der Loos and Nijland 2021). In the case of echinoderms, COI barcodes can be augmented by the use of class-specific 12S and 16S barcodes. The more taxon-specific approach leaves less room for preferential amplification of some taxa. The candidate primers proposed in the present study warrant *in vitro* analyses to validate effectivity but show that 12S and 16S barcodes have potential applications for metabarcoding echinoderms.

In summary, *in silico* analyses suggest that the COI gene alone may be inadequate for amplifying echinoderm sequences – no region was sufficiently conserved to consistently match a universal primer and no clear barcoding gap was found. Barcodes within 12S and 16S rRNA genes that are targeted by class-specific primers could be useful as complementary markers. These offer a similar taxonomic resolution to some existing COI primers, and can potentially make up for lacking taxonomic coverage. Moreover, because these barcodes are short, they can be amplified through massively parallel sequencing, which can greatly improve how larval surveys for echinoderms are conducted.

ACKNOWLEDGMENTS

This study was funded by the Natural Science Research Institute (Project Code: BIO-20-1-02) and the University of the Philippines System Enhanced Creative Work and Research Grant (ECWRG-2019-2-10-R). We also extend our thanks to Dr. Ian Kendrich C. Fontanilla for his feedback on improving the manuscript.

REFERENCES

ALTSCHUL SF, GISH W, MILLER W, MYERS EW, LIPMAN DJ. 1990. Basic local alignment search tool. J Mol Biol 215(3): 403–410.

ARNDT A, MARQUEZ C, LAMBERT P, SMITH MJ. 1996. Molecular phylogeny of eastern Pacific sea cucumbers (Echinodermata: Holothuroidea) based on mitochondrial DNA sequence. Mol Phylogenet Evol 6(3): 425–437.

BOYER F, MERCIER C, BONIN A, LE BRAS Y, TABERLET P, COISSAC E. 2016. Obitools: A unix-inspired software package for DNA metabarcoding. Mol Ecol Resour 16(1): 176–182.

COLLINS RA, BAKKER J, WANGENSTEEN OS, SOTO AZ, CORRIGAN L, SIMS DW, MARIANI S *et al.* 2019. Non-specific amplification compromises environmental DNA metabarcoding with COI. Methods Ecol Evol 10(11): 1985–2001.

DEAGLE BE, JARMAN SN, COISSAC E, POMPANON F, TABERLET P. 2014. DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. Biol Lett 10(9): 20140562.

EDGAR RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5): 1792–1797.

ELLIS JR, ROGERS SI. 2000. The distribution, relative abundance and diversity of echinoderms in the eastern English Channel, Bristol Channel, and Irish Sea. J Mar Biolog Assoc 80(1): 127–138.

FICETOLA GF, COISSAC E, ZUNDEL S, RIAZ T, SHEHZAD W, BESSIÈRE J, POMPANON F. 2010. An *In silico* approach for the evaluation of DNA barcodes. BMC Genom 11: 434.

FOLMER O, BLACK M, HOEH W, LUTZ R, VRIJENHOEK R. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol Mar Biol Biotechnol 3(5): 294–299.

GARCIA-VAZQUEZ E, GEORGES O, FERNANDEZ S, ARDURA A. 2021. eDNA metabarcoding of small plankton samples to detect fish larvae and their preys from Atlantic and Pacific waters. Sci Rep 11(1): 1–13.

GELLER J, MEYER C, PARKER M, HAWK H. 2013. Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. Mol Ecol Resour 13(5): 851–861.

GLENN TC, PIERSON TW, BAYONA-VÁSQUEZ NJ, KIERAN TJ, HOFFBERG SL, THOMAS IV JC, FAIRCLOTH BC *et al*. 2019. Adapterama II: universal amplicon sequencing on Illumina platforms (TaggiMatrix). PeerJ 7: e7786.

[GBIF] Global Biodiversity Information Facility. 2020. GBIF Occurrence Download. Retrieved from https://doi.org/10.15468/dl.np4h6k on 22 Apr 2020.

HEBERT PD, CYWINSKA A, BALL SL, DEWAARD JR. 2003a. Biological identifications through DNA barcodes. Proc R Soc Lond B Biol Sci 270(1512): 313–321.

HEBERT PD, RATNASINGHAM S, DE WAARD JR. 2003b. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc R Soc Lond B Biol Sci 270: S96–S99.

HEBERT PD, STOECKLE MY, ZEMLAK TS, FRANCIS CM. 2004. Identification of birds through DNA barcodes. PLOS Biol 2(10): e312.

HEIMEIER D, LAVERY S, SEWELL MA. 2010. Using DNA barcoding and phylogenetics to identify Antarctic invertebrate larvae: lessons from a large scale study. Mar Genomics 3: 165–177.

HIJMANS RJ. 2020. raster: Geographic Data Analysis and Modeling. R package version 3.3-13.

HOAREAU TB, BOISSIN E. 2010. Design of phylum-specific hybrid primers for DNA barcoding: addressing the need for efficient COI amplification in the Echinodermata. Mol Ecol Resour 10: 960–967.

KIMMERLING N, ZUQERT O, AMITAI G, GUREVICH T, ARMOZA-ZVULONI R, KOLESNIKOV I, SOREK R *et al*. 2018. Quantitative species-level ecology of reef fish larvae *via* metabarcoding. Nat Ecol Evol 2(2): 306–316.

KIMURA M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16(2): 111–120.

KNOTT KE, BALSER EJ, JAECKLE WB, WRAY GA. 2003. Identification of asteroid genera with species capable of larval cloning. Biol Bull 204(3): 246–255.

LAAKMANN S, BOOS K, KNEBELSBERGER T, RAUPACH MJ, NEUMANN H. 2016. Species identification of echinoderms from the North Sea by combining morphology and molecular data. Helgol Mar Res 70(1): 1–18.

LAYTON KKS, CORSTORPHINE EA, HEBERT PDN. 2016. Exploring Canadian echinoderm diversity through DNA barcodes. PLoS One 11(11): 1–16.

LERAY M, KNOWLTON N, HO, SL, NGUYEN BN, MACHIDA RJ. 2019. GenBank is a reliable resource for 21st century biodiversity research. PNAS 116(45): 22651–22656.

LOEZA-QUINTANA T, CARR CM, KHAN T, BHATT YA, LYON SP, HEBERT PD, ADAMOWICZ SJ. 2019. Recalibrating the molecular clock for Arctic marine invertebrates based on DNA barcodes. Genome 62: 200–216.

MACHIDA RJ, TSUDA A. 2010. Dissimilarity of species and forms of planktonic Neocalanus copepods using mitochondrial COI, 12S, nuclear ITS, and 28S gene sequences. PLoS One 5(4): e10278.

MCCLINTOCK JB. 1994. Trophic biology of Antarctic shallow-water echinoderms. Mar Ecol Prog Ser 111(1): 191–202.

PALUMBI SR. 1996. Nucleic acids II: the polymerase chain reaction. In: Molecular Systematics (eds Hillis DM, Moritz C, Mable BK). Sunderland, MA: Sinauer Associates. p. 205–247.

PARADIS E, SCHLIEP K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35(3): 526–528.

PUILLANDRE N, LAMBERT A, BROUILLET S, ACHAZ G. 2012. ABGD, Automatic Barcode Gap Discovery for primary species delimitation. Mol Ecol 21(8): 1864–1877.

R CORE TEAM. 2019. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

RIAZ T, SHEHZAD W, VIARI A, POMPANON F, TABERLET P, COISSAC E. 2011. EcoPrimers: Inference of new DNA barcode markers from whole genome sequence analysis. Nucleic Acids Res 39(21): 1–11.

RIGINOS C, LEIS JM. 2019. Do tiny fish rule the reefs? Science 364(6446): 1128–1130.

SMITH AB. 1997. Echinoderm Larvae and Phylogeny. Annu Rev Ecol and Systemat 28: 219–241.

VALENTINI A, POMPANON F, TABERLET P. 2009. DNA barcoding for ecologists. Trends Ecology Evol 24(2): 110–117.

VAN DER LOOS LM, NIJLAND R. 2021. Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. Mol Ecol 30(13): 3270–3288.

VERON JE, DEVANTIER LM, TURAK E, GREEN AL, KININMONTH S, STAFFORD-SMITH M, PETERSON N. 2009. Delineating the coral triangle. Galaxea, Journal of Coral Reef Studies 11(2): 91–100.

WARD RD, HOLMES BH, O'HARA TD. 2008. DNA barcoding discriminates echinoderm species. Mol Ecol Resour 8: 1202–1211.

WEBB KE, BARNES DK, CLARK MS, BOWDEN DA. 2006. DNA barcoding: a molecular tool to identify Antarctic marine larvae. Deep Sea Res Part II Top Stud Oceanogr 53: 1053–1060.

WIECZOREK J, BLOOM D, GURALNICK R, BLUM S, DÖRING M, GIOVANNI R, VIEGLAIS D *et al*. 2012. Darwin Core: an evolving community-developed biodiversity data standard. PloS One 7(1): e29715.

WINTER DJ. 2017. rentrez: an R package for the NCBI eUtils API (No. e3179v2). PeerJ Preprints.

WOLFE K, GRABA-LANDRY A, DWORJANYN SA, BYRNE M. 2015. Larval phenotypic plasticity in the boom-and-bust crown-of-thorns seastar, *Acanthaster planci*. Mar Ecol Prog Ser 539: 179–189.

APPENDICES

Table I. Representative reference mitochondrial genomes per class for identifying the mitochondrial regions of candidate barcodes.

Class	Accession	Description
Asteroidea	NC_007788	*Acanthaster planci* mitochondrion, complete genome
Crinoidea	NC_010692	*Antedon mediterranea* mitochondrion, complete genome
Echinoidea	NC_033522	*Diadema setosum* mitochondrion, complete genome
Holothuroidea	NC_027086	*Holothuria scabra* mitochondrion, complete genome
Ophiuroidea	NC_013876	*Amphipholis squamata* mitochondrion, complete genome

Table II. Previously identified universal primers for metazoans and echinoderms. All primers target regions within the COI gene, except primer pair arndt2, which targets the 16S rRNA gene.

Primer pair	Primer name	Sequence (5'–3')	Reference
ward1	EchinoF1	TTTCAACTAATCATAAGGACATTGG	Ward *et al.* (2008)
	EchinoR1	CTTCAGGGTGTCCAAAAAATCA	
ward2	EchinoF1	TTTCAACTAATCATAAGGACATTGG	Ward *et al.* (2008)
	COIer	GCTCGTGTRTCTACRTCCAT	Arndt *et al.* (1996)
ward3	EchinoF1	TTTCAACTAATCATAAGGACATTGG	Ward *et al.* (2008)
	HCO2198	TAAACTTCAGGGTGACCAAAAAATCA	Folmer *et al.* (1994)
arndt1	COIef	ATAATGATAGGAGGRTTTGG	Arndt *et al.* (1996)
	COIer	GCTCGTGTRTCTACRTCCAT	
arndt2	16Scuc	TGACAAIIAGGATTGCGACC	
	16Sr	ACTTAGATAGAAACTGACCTG	
layton	LCOech1aF1	TTTTTTCTACTAAACACAAGGATATTGG	Layton *et al.* (2016)
	HCO2198	TAAACTTCAGGGTGACCAAAAAATCA	Folmer *et al.* (1994)
hoareau	COIceF	ACTGCCCACGCCCTAGTAATGATATTTTTTATGGTNATGCC	Hoareau and Boissin (2010)
	COIceR	TCGTGTGTCTACGTCCATTCCTACTGTRAACATRTG	
palumbi	COIF	CCTGCAGGAGGAGGAGAYCC	Palumbi (1996)
	COA	AGTATAAGCGTCTGGGTAGTC	
folmer	LCO1490	GGTCAACAAATCATAAAGATATTGG	Folmer *et al.* (1994)
	HCO2198	TAAACTTCAGGGTGACCAAAAAATCA	

Table III. The number of species and COI sequences represented in sequence group alignments before and after processing (filtering, realignment, and trimming). Counts made per class are enclosed in parentheses.

	Number of species		Number of sequences	
Sequence group	Pre-processing	Post-processing	Pre-processing	Post-processing
Asteroidea	(85)	(77)	(2561)	(1660)
Group 1		1		238
Group 2		1		166
Group 3		2		52
Group 4		15		5
Group 5		58		889
Crinoidea	(87)	(45)	(500)	(169)
Echinoidea	(60)	(43)	(793)	(474)
Group 1		7		151
Group 2		36		323
Holothuroidea	(103)	(85)	(1014)	(740)
Ophiuroidea	(183)	(149)	(1733)	(1340)
Group 1		26		854
Group 2		123		486
Total	518	399	6601	4383

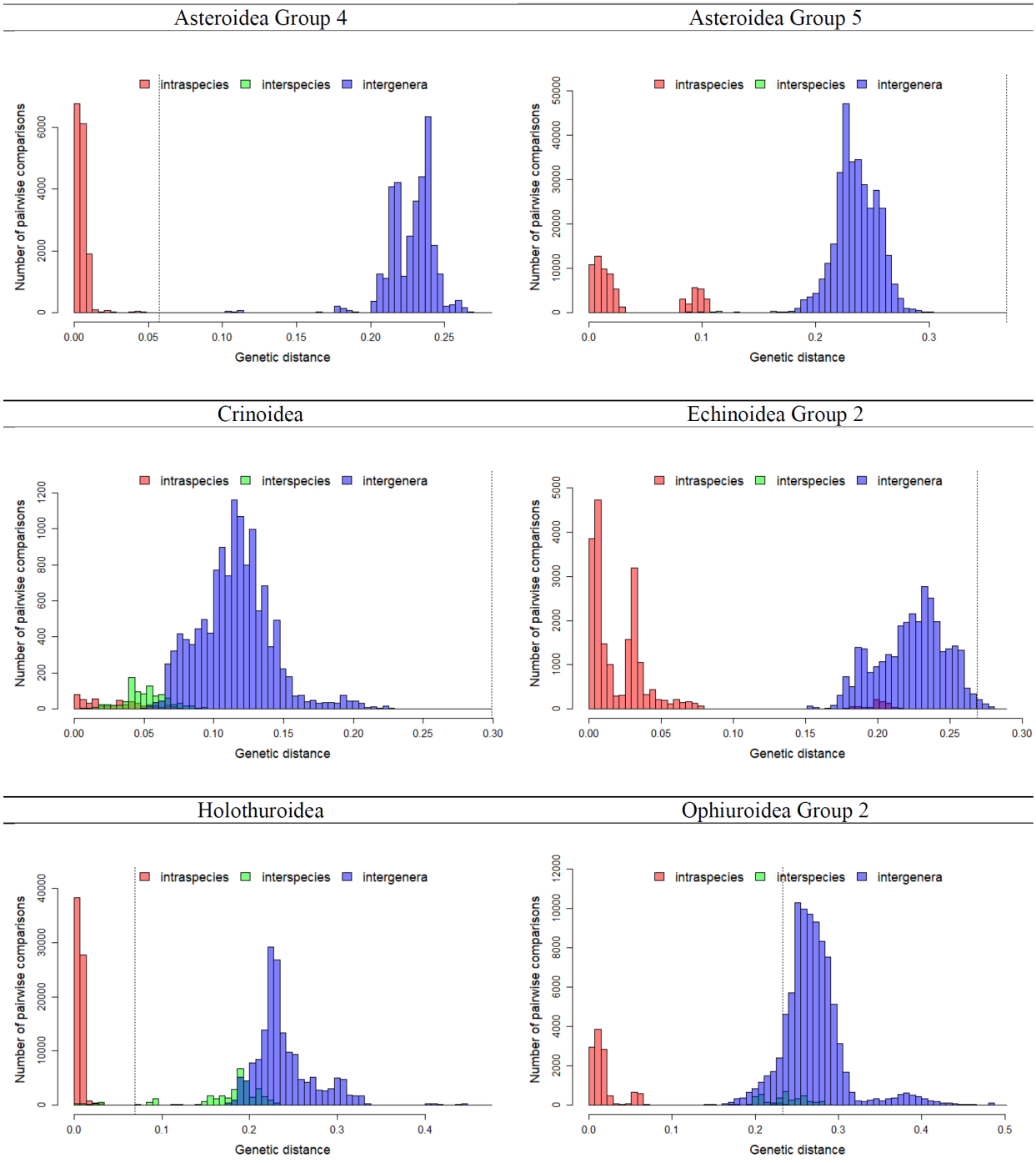

Figure I. Representative frequency distribution for Kimura 2-Parameter genetic distances of COI sequences showing the frequent overlaps between intraspecies, interspecies, and intergeneric pairwise comparisons. Dashed vertical lines mark the 10x-threshold of Hebert *et al.* (2004) – the mean of nonzero intraspecific distances multiplied by 10.