Title

Improving metabarcoding taxonomic assignment: A case study of fishes in a large marine ecosystem

Journal

Molecular Ecology Resources, 21(7)

ISSN

1755-098X

Authors

Gold, Zachary
Curd, Emily E
Goodwin, Kelly D
et al.

Publication Date

2021-10-01

DOI

10.1111/1755-0998.13450

Peer reviewed

MOLECULAR ECOLOGY
RESOURCES

Improving Metabarcoding Taxonomic Assignment: A Case Study of Fishes in a Large Marine Ecosystem

Journal:	*Molecular Ecology Resources*
Manuscript ID	MER-21-0082.R1
Manuscript Type:	Resource Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Gold, Zachary; UCLA, Ecology and Evolutionary Biology Curd, Emily; UCLA, Ecology and Evolutionary Biology Goodwin, Kelly; Atlantic Oceanographic and Meteorological Laboratory, Ocean Chemistry and Ecosystem Division; Southwest Fisheries Science Center Choi, Emma; University of California San Diego Scripps Institution of Oceanography Frable, Benjamin; University of California San Diego Scripps Institution of Oceanography Thompson, Andrew; Southwest Fisheries Science Center Walker Jr, Harold J; University of California San Diego Scripps Institution of Oceanography Burton, Ronald; University of California San Diego Scripps Institution of Oceanography Kacev, Dovi; University of California San Diego Scripps Institution of Oceanography Martz, Lucas Barber, Paul; UCLA, Ecology and Evolutionary Biology
Keywords:	metabarcoding, MiFish primers, California Current Large Marine Ecosystem, eDNA, Environmental DNA, reference database

Improving Metabarcoding Taxonomic Assignment:

A Case Study of Fishes in a Large Marine Ecosystem

12S Taxonomic Assignment Performance

Zachary Gold[1*], Emily E. Curd[1], Kelly D. Goodwin[3], Emma S. Choi[2], Benjamin W. Frable[2], Andrew R. Thompson[4], H. J. Walker, Jr.[2], Ronald S. Burton[2], Dovi Kacev[2], Lucas D. Martz[2], Paul H. Barber[1]

[1] Department of Ecology and Evolutionary Biology, University of California Los Angeles, Los Angeles, California, USA
[2] Scripps Institution of Oceanography, University of California San Diego, La Jolla, California, USA
[3] Ocean Chemistry and Ecosystems Division, Atlantic Oceanographic and Meteorological Laboratory, National Oceanic and Atmospheric Administration, stationed at Southwest Fisheries Science Center, La Jolla, California, USA
[4] Southwest Fisheries Science Center, National Oceanic and Atmospheric Administration, La Jolla, California, USA

[*]Corresponding Author; email: zjgold@g.ucla.edu

ABSTRACT

DNA metabarcoding is an important tool for molecular ecology. However, its effectiveness hinges on the quality of reference sequence databases and classification parameters employed. Here we evaluate the performance of MiFish *12S* taxonomic assignments using a case study of California Current Large Marine Ecosystem fishes to determine best practices for metabarcoding. Specifically, we use a taxonomy cross-validation by identity framework to compare classification performance between a global database comprised of all available sequences and a curated database that only includes sequences of fishes from the California Current Large Marine Ecosystem. We demonstrate that the curated, regional database provides higher assignment accuracy than the comprehensive global database. We also document a tradeoff between accuracy and misclassification across a range of taxonomic cutoff scores, highlighting the importance of parameter selection for taxonomic classification. Furthermore, we compared assignment accuracy with and without the inclusion of additionally generated reference sequences. To this end, we sequenced tissue from 597 species using the MiFish *12S* primers, adding 252 species to GenBank's existing 550 California Current Large Marine Ecosystem fish sequences. We then compared species and reads identified from seawater environmental DNA samples using global databases with and without our generated references, and the regional database. The addition of new references allowed for the identification of 16 additional native taxa representing 17.0% of total sequence reads from eDNA samples, including species with vast ecological and economic value. Together these results demonstrate the importance of comprehensive and curated reference databases for effective metabarcoding and the need for locus-specific validation efforts.

KEYWORDS: metabarcoding, MiFish primers, California Current Large Marine Ecosystem, eDNA, environmental DNA, reference database

INTRODUCTION

Metabarcoding is a process in which multiple species are identified from bulk DNA (e.g. homogenized gut contents, settlement tile scrapings, etc.) or environmental samples (Bohmann et al., 2014; Deiner et al., 2017; Taberlet, Coissac, Pompanon, et al., 2012). Metabarcoding is increasingly used to study marine ecosystems as the ability to sequence tens to hundreds of millions of reads in a single sequencing run allows the development of novel research questions, including species mapping, biomonitoring, gut content analyses, and population genomics, all of which aid understanding of the ecology of marine ecosystems (Baetscher et al., 2019; Closek et al., 2019; Goodwin et al., 2017; Guo, 2017; Kelly, Port, Yamahara, Martone, et al., 2014; Sanders et al., 2015; Thompson et al., 2017; Yamahara et al., 2019). In particular, metabarcoding of environmental DNA (eDNA), freely associated DNA obtained from environmental samples, is an increasingly attractive approach for marine ecosystem characterization because it can detect a broad range of diversity from a single liter of seawater, and has the potential to transform marine biomonitoring efforts (Kelly, Port, Yamahara, Martone, et al., 2014).

Metabarcoding typically employs PCR amplification and sequencing of a target gene (Goodwin et al., 2017) followed by comparison of these sequences to a database of known reference sequences to identify species present in the sample (Taberlet, Coissac, Hajibabaei, et al., 2012). Incomplete databases cannot identify all species present, leading to a lack of assignment despite the actual detection and capture of the sequences, potentially biasing the

interpretation of results (Boyer et al., 2016; Deiner et al., 2017; Machida et al., 2017). Thus, building complete and accurate reference databases is paramount to the success of molecular ecology monitoring efforts (Schenekar et al., 2020).

One approach for maximizing metabarcoding taxonomic assignment is to compare query sequences to a global database of archived sequences (Camacho et al., 2009; Edgar, 2018b). Global databases, such as GenBank, include nearly all publicly available sequences for specific barcode loci and are thus inherently comprehensive (Benson et al., 2018). However, the inclusion of reference barcodes from non-target or biologically irrelevant species may potentially bias taxonomic assignment algorithms (Curd et al., 2019). This issue is particularly problematic for lowest common ancestor taxonomic assignment methods that make inherent assumptions that each best sequence alignment is equally valid, irrespective of the geographic distributions and ecologies of these taxa (Curd et al., 2019; Gao et al., 2017), potentially leading to assignments of biologically implausible species. This problem can be compounded by the occurrence of mis-annotated sequences, a well-known problem in global reference databases (Heller et al., 2018; Leray et al., 2019; Nobre et al., 2016; Wakeling et al., 2019).

An alternative approach to using global databases for taxonomic classification is to employ a curated reference database that includes only appropriately annotated sequences for taxa that occur in a given region (Macheriotou et al., 2019; Poloczanska et al., 2013; Richardson et al., 2018). However, the inclusion or exclusion of barcodes from a reference database can affect metabarcoding taxonomic assignments (Macheriotou et al., 2019; Poloczanska et al., 2013; Richardson et al., 2018), yet few studies systematically addressed this problem (Bergsten et al., 2012; Stoeckle et al., 2020). As such, it is currently unclear whether global or regional reference

databases produce more accurate taxonomic assignments. Systematically quantifying error and bias associated with global and curated database is essential to identifying best practices for metabarcoding taxonomic assignment.

Critical to such assessments are methods that validate taxonomy prediction and evaluate the sensitivity to bioinformatic and database parameters. One key method for comparing the performance of taxonomic classification across different reference databases or classification parameters is the taxonomy cross-validation by identity (TAXXI) framework (Edgar, 2018a). The TAXXI framework is executed by using a reference database with known taxonomic identities that is split into test and training sets and then assigning taxonomy to the training set using the test set. The TAXXI framework can then be applied to allow taxonomic assignment performance to be compared across different metabarcodes, reference databases, and different assignment parameters.

Critically, TAXXI approaches allow for comparing the performance of bioinformatic pipelines within and across loci, including informing the proper selection of classifier parameters for a given metabarcoding locus (Boyer et al., 2016; Machida et al., 2017). Taxonomic assignments made by metabarcoding classifiers are particularly influenced by taxonomic cutoff scores (e.g., exact alignment match or 97% identity threshold) (Edgar, 2018c, 2018a). Using this cross-validation approach to evaluate the performance of taxonomic assignments for *16S* and fungal *ITS* metabarcoding loci across a range of classification parameters revealed that percent identities below 95% had poor classification performance (Edgar 2018a), and highlighted key tradeoffs between assignment confidence and taxonomic resolution (Edgar, 2018c, 2018a). Frequently, attempts to balance confidence–resolution tradeoffs leads to the selection of

conservative taxonomic cutoff scores to avoid over-classification errors (Alberdi et al., 2018; Camacho et al., 2009; Port et al., 2015; Siegwald et al., 2017; Wood & Salzberg, 2014). However, parameter selection is rarely systematically evaluated across different taxonomic groups or metabarcoding loci, inadvertently leading to poorer quality taxonomic assignments (Curd et al., 2019; Edgar, 2018a, 2018c). Importantly, the few studies that explored classification parameter performance across metabarcoding loci found that a "one size fits all" approach (e.g., 97% identity threshold) is inappropriate across different metabarcoding loci (Curd et al., 2019; Edgar, 2018c, 2018a). Thus, evaluating the performance of taxonomic assignments across a range of cutoff scores for a given metabarcoding target is important for maximizing the accuracy of metabarcoding efforts (Balakirev et al., 2017; Bokulich et al., 2018; Hassanin et al., 2010).

Using the TAXXI framework, Curd *et al.* (2019) compared the performance of reference databases for taxonomic assignment, demonstrating the utility of custom reference libraries. The *Creating Reference libraries Using eXisting tools* (*CRUX*) module of the *Anacapa Toolkit* constructs custom reference databases by querying public sequence archives based on primer sets defined by the user. Curd et al. (2019) showed that *CRUX*-generated custom reference databases were more comprehensive and provided improved taxonomic assignment compared to previously published *CO1* reference databases [Midori (Machida et al., 2017) and CO-Arbitrator (Heller et al., 2018)], yielding results nearly equal to heavily curated reference databases for *16S* [SILVA (Quast et al., 2012)] and *12S* [MitoFish (Sato et al., 2018)] metabarcodes. The TAXXI framework thus provides a critical set of tools to evaluate the performance of taxonomic assignment across classification parameters and reference databases for any metabarcoding locus of interest.

The MiFish Universal Teleost and MiFish Elasmobranch primer sets (Miya et al., 2015) target the same portions of the mitochondrial *12S* RNA gene, but differ by a few critical base pairs on the forward primer. These metabarcodes are vertebrate specific, provide species-level resolution for many fishes, and are well suited to short read-length next-generation DNA sequencing, such as Illumina platforms (Collins et al., 2019; Jo et al., 2017; Miya et al., 2015; Valsecchi et al., 2019). As such, they are becoming the standard barcode locus for marine vertebrate metabarcoding studies (Bista et al., 2017; Closek et al., 2019; Miya et al., 2015; Thomsen et al., 2016; Valsecchi et al., 2019; Yamamoto et al., 2017). However, *12S* fish reference databases are relatively incomplete compared to traditional barcoding loci, such as the 655 bp region of the mitochondrial Cytochrome Oxidase I (*COI*) gene (Ardura et al., 2013; Duke & Burton, 2020; Hastings & Burton, 2008; Ward et al., 2009). For example, there is an extensive *CO1* barcode database of fishes of the California Current Large Marine Ecosystem (Hastings & Burton, 2008) that, according to the MitoHelper query of the MitoFish database (accessed April 2021) includes 878 of 1,144 (76.7%) species (Iwasaki et al., 2013; Lim & Thompson, 2021)] facilitating numerous recent metabarcoding studies (Closek et al., 2019; Djurhuus et al., 2020; Pitz et al., 2020). However, there are relatively few reference *12S* sequences that overlap with the MiFish primer sets, limiting the utility of *12S* metabarcoding approaches in this region.

The California Current Large Marine Ecosystem is a highly productive coastal ecosystem that extends approximately 3,000 km across most of the Northeast Pacific from Baja California, Mexico to British Columbia, Canada (Checkley Jr & Barth, 2009; Coleman, 2008; Ekstrom, 2009; Koslow & Davison, 2016). This large marine ecosystem has enormous regional and global importance (Ekstrom, 2009; Sherman, 1991; Wells et al., 2020), driving an ocean economy

valued at over $56 billion USD, employing over 675,000 people (Block et al., 2011; Koslow & Davison, 2016; NMFS, 2017) and supporting food security of the region. The California Current Large Marine Ecosystem also plays a vital role in the cultures and traditional practices of coastal North American tribes and First Nations by supporting species such as Pacific salmon (*Oncorhynchus* spp.), orcas (*Orcinus orca*), eulachon (*Thaleichthys pacificus*), and abalone (*Haliotis* spp.) (Armstrong, 2017; Braje et al., 2017; Brooks et al., 2012; Lepofsky et al., 2017; Norgaard, 2019; Wadewitz, 2012).

Unfortunately, this ecosystem is increasingly facing numerous threats including overexploitation (Koslow & Davison, 2016), ocean acidification and hypoxia (Chan et al., 2008; Crozier et al., 2019; Hofmann et al., 2014; Samhouri et al., 2017), pollution (Good et al., 2020; Halpern et al., 2009), and climate change induced marine heat waves (Rogers-Bennett & Catton, 2019; Santora et al., 2020). Metabarcoding has the power to address many critical management questions in this region, ranging from shifting species distributions, effectiveness of marine protected areas, and seasonal patterns of larval fish recruitment, among others (Duke & Burton, 2020; Kelly, Port, Yamahara, Martone, et al., 2014; Port et al., 2015). However, the ability of metabarcoding efforts to address these important questions hinges on the availability of comprehensive reference databases and appropriate methods of bioinformatic analysis.

To improve the utility of *12S* metabarcoding of marine fishes for the California Current Large Marine Ecosystem and to address larger questions regarding the impact of bioinformatic processes on taxonomic classification, we 1) generated and contributed 741 additional MiFish *12S* sequences representing 597 fish species to global sequence databases; 2) used these additional sequences to create a reference database curated specifically for the California Current

Large Marine Ecosystem; 3) compared the performance of taxonomic assignments made by this regional curated reference database to those made by global marine vertebrate reference databases; and 4) assessed the effect of classifier parameters on phylum through species level assignments of MiFish *12S* sequences to identify optimal locus-specific bioinformatic parameters.

METHODS

Reference Barcode Generation from Fish Tissue Samples

To generate a more complete *12S* barcode reference database for California Current Large Marine Ecosystem fishes, we assembled a list of the 1,144 marine teleost and elasmobranch species that occur in this system (Allen & Horn, 2006; Froese & Pauly, 2010; Hastings & Burton, 2008; Love, & Passarelli, 2020) (Table S1). From this list, we acquired 741 ethanol-preserved voucher specimens representing 597 species (Table S1, Table S2) from the Scripps Institution of Oceanography Marine Vertebrate Collection at the University of California San Diego. DNA was extracted from each tissue sample using a Chelex 100 extraction method (Walsh, Metzger, & Higuchi, 1991), as described in the Supplemental Methods. We amplified all teleost DNA extracts (n=701) using the MiFish Universal Teleost Primers (Miya et al., 2015), and all elasmobranchs (n=55) using the MiFish Elasmobranch primers (Miya et al., 2015) following the thermocycler profile of Curd et al., (2019) (Table S3). We Sanger sequenced purified amplicons (see Supplemental Methods for details), and aligned and trimmed forward and reverse sequences in Sequencher version 5.4.6 (Nishimura, 2000). We used *R* package *taxize* (version 0.9.99) (Chamberlain & Szöcs, 2013) to synonymize taxonomic names of all vouchered

specimens and GenBank. We then checked the accuracy of generated reference barcodes by building a UPGMA phylogenetic tree of all reference sequences and California Current Large Marine Ecosystem fishes using *phangorn* (2.5.5). In addition, we queried each sequence using *blastn* (Camacho et al., 2009) and removed any sequence that did not cluster or align to known taxonomic lineages (data available at https://doi.org/10.5068/D1H963). The resulting *12S* reference barcodes were deposited into GenBank (SAMN19289093–SAMN19289810; Table S2).

Reference Database Creation

To test variation in taxonomic assignment among reference databases, we generated three distinct reference sequence databases: "CRUX-GenBank", "global", and "regional" (Table 1 and Table 2). CRUX-GenBank is a custom *12S* reference database generated using Creating Reference libraries Using eXisting tools (*CRUX*) module of the Anacapa Toolkit to query GenBank for reference barcodes conducted with standard search parameters (Benson et al., 2018; Curd et al., 2019) and MiFish Universal *12S* sequences (Table S1) as the user-defined primers. Briefly, we created this reference database by running *in silico* PCR (Ficetola et al., 2010) on the European Molecular Biology Laboratory (EMBL) standard nucleotide database (Stoesser et al., 2002) to generate a seed library of *12S* references. Next, we used *blastn* (Camacho et al., 2009) to capture reference barcodes without included primer sequences and to query the seed database against the NCBI non-redundant nucleotide database (Gold, 2020; Pruitt et al., 2005; sequences downloaded in October 2019). The resulting *blastn* hits were de-replicated by retaining only the longest version of each sequence and taxonomy for each accession was retrieved using

Entrez-qiime (Baker, 2016). The resulting set of reference sequences in the CRUX-GenBank database included any GenBank reference barcodes that *in silico* amplified to the MiFish *12S* primers at the time of this analysis.

We created the global database to evaluate whether increasing database completeness improves taxonomic assignment. To create the global database, we supplemented the CRUX-GenBank database with 741 additional California Current Large Marine Ecosystem fish *12S* barcodes generated for this study (Table S2). Thus, the global database includes all fish *12S* reference sequences available at the time of download. From this global database, we created the regional database, including only *12S* sequences of fishes known to occur in the California Current Large Marine Ecosystem. We created this database to specifically test whether databases curated to specific ecosystems enhance taxonomic assignment performance relative to more comprehensive databases ("global"). Because of the high degree of similarity between the MiFish Universal and Elasmobranch loci and the flexibility built into *CRUX,* a single *CRUX* generated *12S* reference database performs well for both markers (Curd et al., 2019), so we did not create separate teleost and elasmobranch databases. Additionally, because the MiFish primer set amplifies nearly all vertebrate taxa (Miya et al., 2015; Valsecchi et al., 2019), the global database include teleosts, elasmobranchs, mammals, reptiles, amphibians, birds, etc. All databases are available at https://doi.org/10.5068/D1H963.

Taxonomy cross-validation by identity comparisons

We implemented the taxonomy cross-validation by identity (TAXXI) framework developed by (Edgar, 2018a) to 1) compare taxonomic assignment performance metrics for global versus

regional reference databases, 2) determine the resolution of taxonomic assignments for all available MiFish barcodes in the global database, and 3) understand the performance of the MiFish barcode across taxonomic classifier cutoff scores. Although we use three databases (global, CRUX-GenBank and regional) on our test dataset below, we did not include the CRUX-GenBank database in taxonomic cross validation comparisons because the global database contains all these sequences.

The TAXXI analyses were implemented using scripts from Curd et al. (2019) which adapted TAXXI to the *Anacapa Toolkit* (https://drive5.com/taxxi/doc/index.html and https://github.com/limey-bean/Anacapa). We conducted taxonomic assignments using the *Anacapa Toolkit classifier* which implements the Bayesian Lowest Common Ancestor (BLCA) classifier (Gao et al., 2017) modified to incorporate sequences from *Bowtie2* (Langmead & Salzberg, 2012). In brief, amplicon sequence variants (ASVs; exact unique sequences dereplicated from generated metabarcoding data) are first aligned to reference barcodes using *Bowtie2* retaining the top 100 alignments. Then the BLCA classifier conducts multiple sequence alignment for each query ASV to inform a weighted Bayesian posterior probability of taxonomic assignment. Taxonomy is then ultimately assigned based on the lowest common ancestor of the total weighted reference database matches; reliability is evaluated through bootstrap confidence scores which are analogous to percent identity metrics provided by other metabarcoding classifiers (Gao et al., 2017; See Curd et al. 2019 for full description).

We evaluated taxonomic assignment performance by comparing the following metrics: 1) true positive rate – the number of correct taxonomic assignments divided by the total opportunities for correct classification, 2) over-classification rate - the number of assignments

incorrectly made to additional lower taxonomic ranks divided by the total opportunities to make an over-classification error, 3) under-classification rate - the number of assignments incorrectly made to fewer taxonomic ranks divided by the total opportunities to make an under-classification error, 4) misclassification rate - the number of assignments incorrectly predicted divided by the opportunities for correct classification, and 5) accuracy - the number of correct assignments divided by the taxonomic assignment opportunities for which correctness can be determined (R. C. Edgar, 2018a). The 6) sensitivity was calculated as the true positive rate / (true positive rate + under-classification rate) as under-classification is analogous to a false negative rate. The 7) specificity was calculated as 1- (misclassification rate + over-classification rate) as the combination of the misclassification rate and over-classification rate is analogous to the false positive rate.

Taxonomic Resolution of the MiFish *12S* primer

To provide insights into which fishes can be resolved to species level using the MiFish *12S* primer set, we conducted TAXXI comparisons using the global database as both the test and training database to assign taxonomy to itself. We then calculated the seven taxonomic assignment metrics described above. Additionally, we identified families and genera of fishes for which the MiFish *12S* locus performed poorly, defined as frequently failing to assign species level identification. Although all vertebrate sequences in the global database were used in the taxonomic cross validation, only results for fishes are discussed here.

Regional vs. global reference databases

To compare the relative ability of regional versus global reference databases to accurately assign taxonomy, we conducted two additional TAXXI comparisons using the reference databases created for this study. First, we used the global reference database as a training database to assign taxonomy to the regional reference database that only contained sequences for fishes known from the California Current Large Marine Ecosystem. Second, we used the regional reference database as both the test and training database to assign taxonomy against itself. The taxonomic assignments made by the global and regional reference databases were compared across the taxonomic assignment metrics described above.

Effect of Bootstrap Confidence Scores on Taxonomic Assignment

To understand the performance of the MiFish barcode across a range of taxonomic classifier cutoff scores, we repeated each of the three TAXXI analyses described above (global-regional, regional-regional, global-global) using bootstrap confidence cutoff scores of 40, 50, 60, 70, 80, 90, 95, and 100. We then evaluated the effect of bootstrap confidence cutoff scores across the various taxonomic assignment metrics, as described above.

eDNA Metabarcoding Case Study

Seawater Sample Collection, DNA Extraction, and Library Generation

To specifically test the impact of *12S* database design on taxonomic assignment in real world applications, we compared the performance of the three databases in assigning taxonomy to existing eDNA sequence data as a test case. Briefly, we used MiFish *12S* metabarcoding

sequence data generated from three seawater samples collected from 10 m depth from three sites off eastern Santa Cruz Island, CA in 2017 that were part of a larger ecological study of biodiversity patterns within rocky reef ecosystems. These sequences were generated using standard eDNA collection, processing, and sequencing methods, as outlined in Gold et al., (2021).

We processed this eDNA metabarcoding data three separate times using the *Anacapa Toolkit* (Curd et al., 2019), assigning taxonomy using the CRUX-GenBank, global, and regional reference databases (Table 2). We used the default *Anacapa Toolkit* parameters and a bootstrap confidence cutoff score of 60. We then examined the total number of ASVs and taxonomic ranks identified by each of the three reference databases. We also investigated differences in taxonomic assignment between single direction ASVs (comprised of forward- and reverse-only sequence reads) and merged ASVs (merged paired-end sequence reads) to understand the importance of full length vs. partial length sequences for taxonomic assignment (See Supplemental Results and Discussion).

RESULTS

Generation of Novel Barcodes and 3 References Databases

We generated 741new *12S* MiFish barcode sequences for 597 California Current Large Marine Ecosystem fishes (Table S1 and Table S2), 545 teleosts (bony fishes), 49 elasmobranchs (cartilaginous fishes), and 3 cyclostomatan (jawless fishes) (Table S2). This dataset includes 252 that had no previous *12S* reference barcodes (Table S1).

CRUX created a custom *12S* database comprised of 14,066 taxa and 44,140 sequences with existing entries in GenBank. Adding the 741 novel sequences, above, resulted in a global database comprised of 14,321 species and 44,882 sequences. Restricting these sequences to only fishes from the California Current Large Marine Ecosystem resulted in a curated regional database that includes 706 out of 1,144 (61.7%) reference *12S* barcodes from fishes known from this region. Excluding 382 species missing from the database that are rare in California (n=357) or not coastal (n=25), resulted in a total coverage of 92.7% of the 763 common coastal fishes in this region.

Taxonomy Cross-validation by Identity Comparisons

Regional Versus Global Reference Database Comparisons

The TAXXI quality metrics indicate that the regional reference database yielded more reliable taxonomy at genus and species ranks relative to the global reference databases across all bootstrap confidence scores; regional database species level accuracy ranged from 64.2-94.2% compared to 51.3-90.8% for the global database (Table 2 & Table S4; Figures S1 and S2). This difference was driven by higher misclassification and under-classification rates for the global reference databases. In particular, database misclassification rates were higher for the global compared to the regional reference database across all bootstrap confidence cutoff scores less than 60 (global reference database misclassification 1.8-4.5%, regional database misclassification rate 1.3-3.1%) (Table S4). Likewise, global reference database under-classification rates were higher than regional reference database under-classification rates across

all bootstrap confidence cutoff scores (global reference database under-classification 4.8-48.7%, regional database under-classification rate 2.8-35.8%).

Taxonomic Resolution of the MiFish *12S* primer

Cross validation of the 44,896 sequences within the global database demonstrated that the MiFish primer set delivered 88.0% sensitivity [true positive rate / (true positive rate + under-classification rate)] and 98.2% specificity [1- (misclassification rate + over-classification rate)] at a bootstrap cutoff score of 60 (Table 2, Table S5), providing species level taxonomic assignments to 6,762 fish species, genus level resolution to 923 fish species, family level assignments to 180 fish species, and class level assignments to 2 fish species while overclassifying 214 fish species (Table S5). While poor taxonomic resolution with the MiFish primer sets (e.g. assigned taxonomic rank above species) spanned a large number of genera and families, the genus *Sebastes* and families Cichlidae, Cyprinidae, and Pleuronectidae were particularly problematic (Figures 4 and 5). Of these, *Sebastes* and Pleuronectidae are highly prevalent within the California Current Large Marine Ecosystem. A full breakdown of taxonomic assignment resolution is provided in the Supplemental Results.

Effect of Bootstrap Confidence Scores on Taxonomic Assignment

Across all TAXXI comparisons, accuracy and true positive rates increased with decreasing bootstrap confidence cutoff scores (Figure 1, Figures S1 and S2, Table S4). Likewise, the proportion of species level assignments also increased with decreasing bootstrap confidence score (Figure 2, Figures S3 and S4). We also found that misclassification rates increased with

decreasing bootstrap confidence cutoff score, but at much lower rate (Figure 3, Figures S5 and S6). These results indicate a clear tradeoff between under-classification and misclassification across bootstrap confidence cutoff scores.

eDNA Metabarcoding Example

Unassigned MiFish *12S* ASVs

The *Anacapa Toolkit* failed to assign taxonomy to 49.6% (169/341) of ASVs representing 24.5% (81,002/330,877) of all reads using all three reference databases investigated in this study (Table S6). Of the 169 unassigned ASVs, 16 were forward-only reads, and 153 were merged reads. To explore the origins of these unassigned reads, we used BLAST to query all GenBank sequences, revealing that 94.7% (160/169) of these ASVs aligned to marine prokaryotic and eukaryotic *16S* sequences (Max Alignment Scores 87.9-475). Of these aligned ASVs, 85% (136/160) matched to uncultured sequences generated from marine metagenomic studies. 80.0% (128/160) of successfully aligned ASVs matched to bacterial barcodes including those from *Psychromonas* sp., *Photococcus caeruleum*, *Loktanella* sp., *Leucothrix* sp., and *Gimesia* sp., and cyanobacteria. A smaller fraction of assigned ASVs (18.8%; 30/160) best aligned to eukaryotic sequences including those from diatoms (e.g. *Nitzschia alba* and *Eucampia antarctica*) and other marine microalgae (e.g. Picobiliphytes, *Heterosigma akashiwo*, *Mesopedinella arctica*, and *Phacus warszewiczii*). Given that these 169 unassigned sequences were non-vertebrate, we excluded these ASVs from all subsequent comparisons. All remaining 172 ASVs were assigned to a class of vertebrates by at least one of the three reference databases used. Of these vertebrate ASVs, 58 were merged, 107 were forward-only, and 7 were reverse only reads.

372

373 **Comparisons of CRUX-GenBank, Global, and Regional Reference Database Taxonomic**
374 **Assignments**

375 The inclusion of additional reference barcodes increased the total number of ASVs and reads
376 assigned to marine fishes resident in the California Current Large Marine Ecosystem (Tables 2 &
377 Table S7). Importantly, the inclusion of novel voucher sequences within the global database
378 resulted in species-level identification for 11 additional California Current Large Marine
379 Ecosystem fishes including Kelp Bass (*Paralabrax clathratus*), California Moray (*Gymnothorax*
380 *mordax*), Opaleye (*Girella nigricans*), Giant Kelpfish (*Heterostichus rostratus*), Ocean
381 Whitefish (*Caulolatilus princeps*), and California Halibut (*Paralichthys californicus*) (Table S8).
382 Use of the regional database largely increased accuracy of taxonomic assignments. The
383 regional database assigned an ASV to the Black Croaker (*Cheilotrema saturnum*) that was only
384 assigned to the family Sciaenidae by the global database. Additionally, the regional database
385 assigned one ASV as Bat Ray (*Myliobatis californica*) and another as Jack Mackerel (*Trachurus*
386 *symmetricus*), species native to the California Current Large Marine Ecosystem, that the global
387 database assigned to the non-native species, Common Eagle Ray (*Myliobatis aquila*) and Rough
388 Scad (*Trachurus lathami*), respectively. However, the regional reference database failed to
389 resolve the taxonomy of one ASV that the global database assigned to the family of Delphinidae.

390

391 # DISCUSSION

392 Taxonomic assignment in metabarcoding studies typically employ large public sequence
393 databases such as GenBank or Barcode of Life (Leray & Knowlton, 2015; Schenekar et al.,

2020; Stat et al., 2017), or databases that are curated to specific barcoding markers or taxonomic groups without consideration of species distributions (e.g. Curd et al. 2019). However, systematic comparison of these approaches to a curated, region-specific reference database shows that the region-specific database outperforms the global databases in metabarcoding taxonomic assignment (Table 1). Accuracy of eDNA metabarcoding only improved by including GenBank sequences from fishes native to the California Current Large Marine Ecosystem and supplementing these sequences with additional reference barcodes. Furthermore, examination of taxonomic assignment over a range of bootstrap cutoff scores revealed key tradeoffs, with lower bootstrap confidence cutoffs yielding more accurate species assignment, but at the cost of higher misclassification rates. Combined, these results highlight the importance of reference database and bootstrap cutoff selection in obtaining the best results from metabarcoding studies.

In a test dataset for fish eDNA extracted from seawater collected from three sites on Santa Cruz Island, the regional database performed the best. The regional database identified 16 additional ASVs to species not identified by the CRUX-GenBank database, and an additional 3 fishes that were misidentified by the global database (Table 2). Higher accuracy with increased database completeness echoes previous research on the importance of complete reference databases in metabarcoding (Leray et al., 2012; Machida et al., 2017), and greatly improves the utility of eDNA for monitoring the California Current Large Marine Ecosystem.

Although the *12S* barcodes and reference databases tested here performed well with regard to annotating fish species (e.g., 91.3% sensitivity and 98.3% specificity across MiFish reference barcodes), almost half of the ASVs and a quarter of all reads generated in our eDNA test datasets were not assigned to any fish reference barcode (Table 2). While other

metabarcoding studies report similar levels of unassigned taxa (Leray & Knowlton, 2017) and others have encountered this issue (Goodwin, personal communication), this issue isn't widely reported in the literature, particularly considering the popularity of the MiFish primers. Further investigation showed that the vast majority of unassigned ASVs were uncultured bacteria *16S* loci (Table S6) derived from marine shotgun sequencing metagenomic studies (Bork et al., 2015). This result highlights that the MiFish Teleost *12S* primer set, while extremely useful for targeting vertebrate *12S* loci, can also amplify non-target *16S* genes, raising the possibility that non-target amplification may at best result in lower returns of target sequences, and at worst artificially increase estimates of fish diversity.

Importance of Regional reference databases

Given that increased reference database completeness increases the ability to assign ASV's to species (Table 2), it is logical to assume that databases with more taxonomic coverage are universally better (Curd et al., 2019). However, our results suggest an unexpected trade-off between greater diversity of barcodes and ecologically informed taxonomic assignment. For example, using only the regional database specific to California Current Large Marine Ecosystem marine fishes, we identified important native taxa like Black Croaker (*Cheilotrema saturnum*) and Bat Ray (*Myliobatis californica*) in eDNA isolated from seawater samples. However, while the global database contained the largest total number of barcodes, including all taxa in the regional database, Black Croaker was not identified and Bat Ray was inconsistently identified across multiple ASVs. The global database failed to identify Black Croaker due to the high similarity of *12S* barcode sequences within the Family Sciaenidae, specifically within the

clade that includes *Cheilotrema,* a genus native to California, as well as *Equetus* and *Pareques,* non-native coral reef-associated genera (Table S8). Similarity of barcode sequences also explains the loss of taxonomic resolution in *Myliobatis*.

By excluding highly similar non-native *12S* barcodes, the database curated for the region of interest provided more accurate species-level assignments and far fewer under-classifications and misclassifications, demonstrating that a database comprised of only local taxa is preferred to maximize identification of local species. Yet, this improvement was not universal. For example, the regional database failed to classify one ASV belonging to the family Delphinidae that was identified by both the CRUX-GenBank and global databases. This result stems from the regional database being specific to California Current Large Marine Ecosystem fishes, and could thus not identify a marine mammal. This shortcoming easily could be overcome, however, by appending the regional database with barcodes for other marine-associated vertebrate taxa of regional management interests (Valsecchi et al., 2019). An alternative and taxon agnostic approach currently employed by the co-authors is to conduct taxonomic assignments twice. First, taxonomic assignments are conducted using a regional reference database to get the best taxonomic assignment for focal taxa of interest, and second using a global reference database to identify as many remaining unidentified ASVs as possible (Gold et al., 2021). We did not directly report the results of the two-step taxonomic assignment method here as the only difference between this approach and the taxonomic assignments made using the regional database alone is the additional assignment of the single Delphinidae ASV.

These results highlight the tradeoff between identifying local species from clades with little genetic variation and providing taxonomic coverage across a broad range of species. As

such, researchers need to identify their research priorities when deciding on which reference databases to use, with a particular focus on defining the scope of the target taxa. Future work could alleviate this tradeoff by building bioinformatic pipelines that prioritize assignments to a reference set of resident species, perhaps by including information on species ranges and sample locations in the assignment algorithm. However, an advantage of the two-step approach outlined above is that it allows for eDNA studies to address specific ecological questions without having a specific target list in mind. This approach is particularly important for eDNA studies which directly test for the presence of invasive species or range shifts associated with climate change (Bohmann et al., 2014; Klymus et al., 2017).

Importance of Taxonomic Cutoff Scores

Taxonomic cutoff scores, or percent identity, strongly influenced taxonomic assignments (Edgar, 2018c). Patterns for the MiFish *12S* locus showed a similar pattern with higher true positive and misclassification rates and lower under-classification rates at lower bootstrap confidence cutoff scores (Edgar, 2018a). These results highlight a key tradeoff between under-classification and misclassification for metabarcoding taxonomic assignment, and demonstrate that the decision of which taxonomic cutoff score can strongly influence results (Edgar, 2018a). Lower bootstrap confidence cutoffs ensure a higher overall accuracy in species-level identification but come at the cost of higher misclassification rates to an incorrect species-level assignment.

Our results suggest that a TAXXI bootstrap confidence cutoff score of 60 provides a balance between maximizing species-level assignment accuracy (89.7%, global reference database) while minimizing misclassification rates (1.7%, global reference database), matching

the general findings of Curd et al. (2019). However, in instances in which metabarcoding results may influence management or health decisions with substantial legal or economic ramifications (i.e., detection of an endangered or invasive species or discriminating a putative disease causing microbe) a misclassification error may be valued as a far less desirable outcome than an under-classification error (Bohmann et al., 2014; Lodge et al., 2012; Wakeling et al., 2019). In such cases, results indicate that there isn't one single bootstrap confidence cutoff score that completely ameliorates these tradeoffs (Figure 1).

Given that previous work demonstrates that results may not be consistent across loci (Curd et al., 2019), we can only generalize our results to the MiFish *12S* primer set. Determining confidence–resolution tradeoffs in other widely used primer sets will be fundamental for effectively interpreting metabarcoding results from those loci. Combining the capabilities of *CRUX* with the TAXXI framework provides a critical set of tools to both generate and evaluate the performance of a range of metabarcoding loci and reference databases (Table 1; Curd et al., 2019; Edgar, 2018a), facilitating such studies. Given the growing number of metabarcoding applications across a broad range of ecosystems and taxa (Curd et al., 2019; Deiner et al., 2017; Edgar, 2018a), assessing the performance of barcoding markers in the taxonomic group of interest is critical.

Importance of Complete Reference Databases

Previous eDNA metabarcoding efforts in the California Current Large Marine Ecosystem report poor species-level identification and frequent taxonomic assignment to non-native sister taxa (Closek et al., 2019; Kelly, Port, Yamahara, & Crowder, 2014; Port et al., 2015). For example,

an eDNA metabarcoding study in Southern California (Curd et al., 2019) assigned multiple *12S* ASVs to *Girella simplicidens*, the Gulf Opaleye, a fish that does not occur in California Current Large Marine Ecosystem coastal waters (Froese & Pauly, 2010; Love & Passarelli, 2020). This incorrect assignment occurred due to the lack of *12S* reference sequences for the local native Opaleye, *G. nigricans*. By maximizing the number of local reference barcodes, regional databases allow the reads to be correctly assigned to ecologically and geographically relevant species.

In our eDNA samples, the regional database improved species-level assignments, identifying an additional 17.0% of the total vertebrate sequence reads. Much of this improvement was due to the inclusion of reference barcodes for Kelp Bass (*Paralabrax clathratus*), one of the most abundant marine species in Southern California kelp forest ecosystems and an important sport fishery target (Pondella II et al., 2015). By including a reference barcode for this species, the regional database assigned 20 previously unidentified ASVs to *P. clathratus,* which accounted for 16.4% of our total sequence reads. Thus, even the inclusion of reference barcodes for a few key native taxa can dramatically improve metabarcoding efforts.

Taxonomic Assignment Limitations of MiFish primers

Of the 8,084 fishes represented in the global database, the MiFish primers were unable to provide species level taxonomic assignments to 1,322 species (See Table S5 for complete list of putative *in silico* taxonomic assignments). Thus, although the MiFish primer set has broad utility for fish metabarcoding, this portion of *12S* cannot resolve many fishes to species (Miya et al., 2015). These results highlight the tradeoff between breadth and specificity of any metabarcoding

primer set, a result consistent with previous investigations of the MiFish primer set and universal barcodes in general (Deiner et al., 2017; Miya et al., 2015). Critically, these results provide much needed insights into taxonomic blind spots of the MiFish primers, informing primer selection for future fish metabarcoding applications both in the California Current Large Marine Ecosystem and globally (Figures 4 and 5).

Another key limitation to metabarcoding taxonomic assignment is the prevalence of sequence misannotations in public sequence repositories. Misannotations arise predominantly from subtle incidental issues, such as mislabeling of sequences, and thus are particularly difficult to address bioinformatically (Heller et al., 2018; Nobre et al., 2016; Wakeling et al., 2019). To date, the onus of identifying and preventing misannotations are on the user and research community and there remain few systematic methods for identifying and removing misannotated sequences although Kozlov et al., 2016 is a notable exception (we also note there is a process to flag and report such sequences are available through GenBank). One potential solution to the issue of misannotated sequences is the development and maintenance of global curated datasets (e.g., MitoFish, Silva, and UNITE) (Nilsson et al., 2018; Quast et al., 2012; Sato et al., 2018). However, while these approaches may work well for a handful of key loci and taxonomic targets, these approaches are not scalable with the rapid development of additional metabarcoding loci and targets of interest (Curd et al., 2019). Thus, further efforts to systemically prevent and address mis-annotations in public sequence repositories clearly are warranted.

Limitations of Barcoding Efforts

The regional database did not include barcodes for all California Current Large Marine Ecosystem fishes (Table S1) due to a combination of limited resources, difficulties amplifying vouchered tissue samples, the onset of the COVID-19 pandemic (Omary et al., 2020), and a lack of some vouchered reference material within the Marine Vertebrates Collection of the Scripps Institution of Oceanography. In total, our regional database did not include 438 of 1,144 (38.3%) California Current Large Marine Ecosystem fishes. However, the vast majority of these (n=357) are rare in the state of California (the focus of the collection and study), others (n=25) are common but not coastal species. Discounting these, our barcoding efforts provide coverage for 92.7% of the 763 marine fishes common in this ecosystem, making it an important tool for metabarcoding studies, despite a small number (n=53) of common coastal species missing from the database (Table S9).

The one major shortcoming of our barcoding efforts is that 7.3% (n=32) of the missing taxa are rockfishes in the genus *Sebastes*. Rockfishes are ecologically important (Hyde & Vetter, 2007), form the basis of many commercial and recreational fisheries (Lea et al., 1999; Williams et al., 2010), and declines in rockfish stocks led to the establishment of the largest marine protected areas in southern California, the Cowcod Conservation Areas (Thompson et al., 2017). Unfortunately, this shortcoming cannot be easily overcome through additional *12S* barcoding because rockfish are a recent and diverse radiation comprised of 110 species (Ingram & Kai, 2014) and *12S* fails to resolve most *Sebastes* to species-level (Hyde & Vetter, 2007; Yamamoto et al., 2017). Thus, effective metabarcoding of *Sebastes* will require designing novel *Sebastes*-specific metabarcoding primers that target a more rapidly evolving region of the mitochondrial

568 genome (e.g. *CytB*) (Min et al., 2020; Thompson et al., 2017). Importantly, this *Sebastes*
569 example highlights the importance of comprehensively evaluating the taxonomic performance of
570 a particular locus (here MiFish *12S*) for a given taxonomic group and the difficulty of using
571 metabarcoding methods for delineating species within an adaptive radiation.

572 Despite these limitations, however, the current regional California Current Large Marine
573 Ecosystem *12S*-specific reference database includes all but one non-*Sebastes* nearshore species
574 monitored by the Channel Islands National Kelp Forest Monitoring Program (n=80, Sprague et
575 al., 2013), as well as by PISCO, the Partnership for Interdisciplinary Studies of Coastal Oceans
576 (n=76; the only missing species is White Sea Bass *Atractoscion nobilis*; Caselle, Rassweiler,
577 Hamilton, & Warner, 2015; Pondella II et al., 2015). Further, there is now a *12S* reference
578 sequence for 98 of the 100 most abundant ichthyoplankton species collected by the California
579 Cooperative Oceanic Fisheries Investigation (CalCOFI) from the California Current Large
580 Marine Ecosystem between 1951-2019 (only missing Showy Bristlemouth *Cyclothone signata*
581 and White Barracudina, *Arctozenus risso*) (Moser, 1993). Moreover, in real world application,
582 this reference barcode database assigned taxonomy to over 90% of vertebrate ASVs detecting a
583 broad range of ecologically and commercially important nearshore rocky reef species (Pondella
584 II et al., 2019). As such, our barcoding efforts represents an important genetic resource for
585 coastal California marine metabarcoding monitoring efforts.
586

587 ## Off Target Limitations of MiFish primers

588 High numbers of unidentified ASVs are a common feature of barcoding and metabarcoding
589 studies (e.g. Leray & Knowlton, 2017). These unidentified ASVs are typically attributed to

incomplete reference databases (Curd et al., 2019; Ransome et al., 2017; Schenekar et al., 2020) and/or novel biodiversity (Barber & Boyce, 2006; Boussarie et al., 2018). However, given that the regional database includes 92.7% of fishes common in this coastal ecosystem, it was extremely surprising that half of all ASVs and a quarter of all sequences generated in our eDNA test datasets could not be assigned.

In fact, the vast majority of these sequences and ASVs did not belong to vertebrates, but instead uncultured marine bacteria, specifically matching to *16S*, rather than the target *12S* locus. Since mitochondria represent the capture of microbial endosymbionts by ancient eukaryotes (Roger et al., 2017) and that this capture occurred in the sea, it perhaps is not surprising that primers designed to target vertebrate *12S* might also capture marine prokaryotes. Similarly, the homology between vertebrate *12S* and prokaryotic and bacterial *16S* genes is well known (Crews & Attardi, 1980) suggesting capturing microbial *16S* with vertebrate *12S* primers is not surprising. However, this particular feature of the MiFish primer set previously has not been widely reported in the scientific literature (Minamoto et al., 2020), potentially impacting the interpretation of unidentified ASVs in other fish metabarcoding studies.

These findings highlight the importance of accurate universal metabarcoding primer design, especially in outlining both target and non-target sequences. In the design of the MiFish Teleost *12S* primers, uncultured marine microbe *16S* sequences were not considered as potential alternative targets for the primer set, resulting in the selection of a metabarcoding locus with a high degree of non-target amplification (Miya et al., 2015). This finding is important for the marine vertebrate eDNA community, which has recently converged on the MiFish *12S* primers as the vertebrate barcode of choice (Closek et al., 2019; Miya et al., 2020; O'Donnell et al.,

612 2017; Valsecchi et al., 2019; Yamahara et al., 2019). At best, this non-target amplification of
613 microbial DNA will lead to wasted sequencing effort, as every microbial sequence generated
614 reduces the number of vertebrate sequences captured. Such a situation would be particularly
615 problematic for relatively rare targets. At worst, it could result in incorrect interpretation of
616 unidentified ASVs and lead to incorrect biomonitoring assessments (Cordier et al., 2018). This
617 problem is of particular concern in biodiversity hotspots such as the Coral Triangle where
618 reference databases are incomplete, as well as in environments with high abundance of bacteria
619 relative to vertebrate biomass such as in some pelagic midwater and deep-sea habitats where
620 recent eDNA sample collection efforts have struggled to detect vertebrate sequences (K. Pitz
621 personal communication).

622 Previous applications of MiFish *12S* primer sets did not identify high rates of non-
623 homologous sequences (Collins et al., 2019; Miya et al., 2015). Interestingly, these studies used
624 higher annealing temperatures (60-65˚C) and fewer PCR cycles than those used in this study
625 which may potentially explain why we observed high rates of *16S* amplification. We note that we
626 used the touchdown PCR method in order to successfully amplify eDNA from sea water
627 samples. A white paper from *The eDNA Society* in Japan using the original 65˚C annealing
628 methods highlighted that the application of a size-selection step during library preparation (either
629 via gel extraction or dual size-selection bead clean up) can be used remove off-target sequences
630 and help ameliorate this issue (Minamoto et al., 2020; Miya & Sado, 2019). We also confirm
631 here that non-target ASVs are substantially longer in length than vertebrate *12S* fragments.
632 Incorporating these practices to reduce microbial cross-amplification will improve the
633 application of MiFish *12S* metabarcoding efforts. Ultimately, understanding the full scope of

taxa that can be amplified by a given metabarcoding primer is critical for the successful application and interpretation of results and concerted efforts to validate markers are clearly warranted.

Towards improved metabarcoding efforts

The curated California Current Large Marine Ecosystem *12S*-specific reference database was designed to improve effectiveness of metabarcoding of California Current Large Marine Ecosystem fishes. To further improve and expand the taxonomic coverage of the database, we generated a website that identifies species needing *12S* reference barcodes and provides the research community targets for additional barcoding efforts (Zack Gold, 2020). The ability to update and expand the regional reference database will be especially important as climate change leads to range expansions of sub-tropical species that may become resident within the California Current Large Marine Ecosystem (Gentemann et al., 2017; Harvell et al., 2019; Sanford et al., 2019; Walker et al., 2020). The importance of expanding the database is highlighted by our detection of Finescale Triggerfish, *Balistes polylepis,* in the eDNA samples, a species that has only recently become more common off Santa Cruz Island and La Jolla since the 2014-2016 marine heatwave (B. Frable & S. McMillan, personal communication).

Additionally, while the MiFish Teleost and Elasmobranch *12S* loci are important targets for current marine metabarcoding studies, future efforts and different applications of marine metabarcoding will likely rely on additional barcoding targets. Here we used the same primer set to both generate reference barcodes as well as conduct metabarcoding. Although, this choice limits the applicability and usefulness of our reference barcode generation beyond

metabarcoding efforts, it allowed us to more easily and rapidly sequence and generate barcodes for our intended purpose. Furthermore, recent efforts have found success multiplexing *CO1* and *16S* loci simultaneously provides more species-level identifications than either marker alone, demonstrating complimentary genetic loci can improve metabarcoding assignments (Duke & Burton, 2020). Thus future efforts to develop rapid and affordable multilocus barcoding or whole mitogenomic tools will provide greater resources for marine metabarcoding and population genomic efforts (Coissac et al., 2016). As these new barcode loci are developed (e.g., *Sebastes*-specific barcodes), the California Current Large Marine Ecosystem specific reference database can be expanded to include these loci. Additionally, resources like the SIO Marine Vertebrate Collection will continue to provide important voucher specimens for advancing marine molecular ecology resources as they accession new material.

ACKNOWLEDGEMENTS

We acknowledge NSF GRFP and GRIP [DEG No. 2015204395], UCLA La Kretz Center for Conservation Genomics, Howard Hughes Medical Institute, and UCLA Department of Ecology and Evolutionary Biology for funding this research. In addition, we would like to thank Regina Wetzer, Dean Pentcheff, Adam Wall, Janessa Wall, and Giacamo Bernardi for providing additional voucher specimens. We would like to thank Julianne K. Passarelli and Milton S. Love for sharing their comprehensive list of California fish species from their new edition of *Miller and Lea's Guide to the Coastal Marine Fishes of California* (Love & Passarelli, 2020). We would also like to thank Miya Masaki for sharing protocols and insights on the performance of the MiFish primer. Lastly, we would like to thank McKenzie Koch, Beverly Shih, Nikita Sridhar,

Lauren Man, Eric Caldera, Candice Cross, and Erick Zerecero for help conducting all laboratory work. We would like to thank William Watson, Phillip A. Hastings, Peggy Fong, Richard F. Ambrose, and Thomas B. Smith for providing thoughtful feedback and comments during the preparation of the manuscript.

REFERENCES

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution*, *9*(1), 134–147. https://doi.org/doi:10.1111/2041-210X.12849

Allen, L. G., & Horn, M. H. (2006). *The ecology of marine fishes: California and adjacent waters*. Univ of California Press.

Ardura, A., Planes, S., & Garcia-Vazquez, E. (2013). Applications of DNA barcoding to fish landings: authentication and diversity assessment. *ZooKeys*, *365*, 49–65. https://doi.org/10.3897/zookeys.365.6409

Armstrong, C. G. D. (2017). *Historical ecology of cultural landscapes in the Pacific Northwest*. Environment: Department of Archaeology.

Baetscher, D. S., Anderson, E. C., Gilbert-Horvath, E. A., Malone, D. P., Saarman, E. T., Carr, M. H., & Garza, J. C. (2019). Dispersal of a nearshore marine fish connects marine reserves and adjacent fished areas along an open coast. *Molecular Ecology*, *28*(7), 1611–1623. https://doi.org/10.1111/mec.15044

Baker, C. (2016). *bakerccm/entrez_qiime: entrez_qiime v2.0*. https://doi.org/10.5281/ZENODO.159607

Balakirev, E. S., Saveliev, P. A., & Ayala, F. J. (2017). Complete mitochondrial genomes of the Cherskii's sculpin Cottus czerskii and Siberian taimen Hucho taimen reveal GenBank entry errors: Incorrect species identification and recombinant mitochondrial genome. *Evolutionary Bioinformatics*, *13*, 1176934317726783.

Barber, P., & Boyce, S. L. (2006). Estimating diversity of Indo-Pacific coral reef stomatopods through DNA barcoding of stomatopod larvae. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1597), 2053–2061.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K. D., & Sayers, E. W. (2018). GenBank. *Nucleic Acids Research*, *46*(D1), D41–D47.

Bergsten, J., Bilton, D. T., Fujisawa, T., Elliott, M., Monaghan, M. T., Balke, M., Hendrich, L., Geijer, J., Herrmann, J., & Foster, G. N. (2012). The effect of geographical scale of sampling on DNA barcoding. *Systematic Biology*, *61*(5), 851–869.

Bista, I., Carvalho, G. R., Walsh, K., Seymour, M., Hajibabaei, M., Lallias, D., Christmas, M., & Creer, S. (2017). Annual time-series analysis of aqueous eDNA reveals ecologically relevant dynamics of lake ecosystem biodiversity. *Nature Communications*, *8*, 14087.

Block, B. A., Jonsen, I. D., Jorgensen, S. J., Winship, A. J., Shaffer, S. A., Bograd, S. J., Hazen, E. L., Foley, D. G., Breed, G. A., Harrison, A.-L., Ganong, J. E., Swithenbank, A.,

Castleton, M., Dewar, H., Mate, B. R., Shillinger, G. L., Schaefer, K. M., Benson, S. R., Weise, M. J., … Costa, D. P. (2011). Tracking apex marine predator movements in a dynamic ocean. *Nature*, *475*(7354), 86–90. https://doi.org/10.1038/nature10082

Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., Yu, D. W., & de Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution*, *29*(6), 358–367. https://doi.org/10.1016/j.tree.2014.04.003

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A., & Caporaso, J. G. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*, *6*(1), 90.

Bork, P., Bowler, C., De Vargas, C., Gorsky, G., Karsenti, E., & Wincker, P. (2015). *Tara Oceans studies plankton at planetary scale*. American Association for the Advancement of Science.

Boussarie, G., Bakker, J., Wangensteen, O. S., Mariani, S., Bonnin, L., Juhel, J.-B., Kiszka, J. J., Kulbicki, M., Manel, S., & Robbins, W. D. (2018). Environmental DNA illuminates the dark diversity of sharks. *Science Advances*, *4*(5), eaap9661.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). obitools: A unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, *16*(1), 176–182.

Braje, T. J., Rick, T. C., Szpak, P., Newsome, S. D., McCain, J. M., Smith, E. A. E., Glassow, M., & Hamilton, S. L. (2017). Historical ecology and the conservation of large, hermaphroditic fishes in Pacific Coast kelp forest ecosystems. *Science Advances*, *3*(2), e1601759.

Brooks, J. F., Carothers, C., Colombi, B. J., Diver, S., Kasten, E., Koester, D., Lien, M. E., Menzies, C. R., Reedy-Maschner, K., & Sharakhmatova, V. N. (2012). *Keystone nations: indigenous peoples and salmon across the north Pacific*. School for Advanced Research Press.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*(1), 421.

Caselle, J. E., Rassweiler, A., Hamilton, S. L., & Warner, R. R. (2015). Recovery trajectories of kelp forest animals are rapid yet spatially variable across a network of temperate marine protected areas. *Scientific Reports*, *5*, 14102. https://doi.org/10.1038/srep14102

Chamberlain, S. A., & Szöcs, E. (2013). taxize: taxonomic search and retrieval in R. *F1000Research*, *2*.

Chan, F., Barth, J. A., Lubchenco, J., Kirincich, A., Weeks, H., Peterson, W. T., & Menge, B. A. (2008). Emergence of anoxia in the California current large marine ecosystem. *Science*, *319*(5865), 920.

Checkley Jr, D. M., & Barth, J. A. (2009). Patterns and processes in the California Current System. *Progress in Oceanography*, *83*(1–4), 49–64.

Closek, C. J., Santora, J. A., Starks, H. A., Schroeder, I. D., Andruszkiewicz, E. A., Sakuma, K. M., Bograd, S. J., Hazen, E. L., Field, J. C., & Boehm, A. B. (2019). Marine vertebrate biodiversity and distribution within the central California Current using environmental DNA (eDNA) metabarcoding and ecosystem surveys. *Frontiers in Marine Science*, *6*, 732.

Coissac, E., Hollingsworth, P. M., Lavergne, S., & Taberlet, P. (2016). From barcodes to genomes: extending the concept of DNA barcoding. *Molecular Ecology*, *25*(7), 1423–1428.

761 Coleman, K. (2008). Research review of collaborative ecosystem-based management in the
762 California current large marine ecosystem. *Coastal Management*, *36*(5), 484–494.
763 Collins, R. A., Bakker, J., Wangensteen, O. S., Soto, A. Z., Corrigan, L., Sims, D. W., Genner,
764 M. J., & Mariani, S. (2019). Non-specific amplification compromises environmental DNA
765 metabarcoding with COI. *Methods in Ecology and Evolution*, *10*(11), 1985–2001.
766 https://doi.org/10.1111/2041-210X.13276
767 Cordier, T., Forster, D., Dufresne, Y., Martins, C. I. M., Stoeck, T., & Pawlowski, J. (2018).
768 Supervised machine learning outperforms taxonomy-based environmental DNA
769 metabarcoding applied to biomonitoring. *Molecular Ecology Resources*, *18*(6), 1381–1391.
770 Crews, S., & Attardi, G. (1980). The sequences of the small ribosomal RNA gene and the
771 phenylalanine tRNA gene are joined end to end in human mitochondrial DNA. *Cell*, *19*(3),
772 775–784.
773 Crozier, L. G., McClure, M. M., Beechie, T., Bograd, S. J., Boughton, D. A., Carr, M., Cooney,
774 T. D., Dunham, J. B., Greene, C. M., & Haltuch, M. A. (2019). Climate vulnerability
775 assessment for Pacific salmon and steelhead in the California Current Large Marine
776 Ecosystem. *PloS One*, *14*(7), e0217711.
777 Curd, E. E., Gold, Z., Kandlikar, G. S., Gomer, J., Ogden, M., O'Connell, T., Pipes, L.,
778 Schweizer, T. M., Rabichow, L., Lin, M., Shi, B., Barber, P. H., Kraft, N., Wayne, R., &
779 Meyer, R. S. (2019). Anacapa: an environmental DNA toolkit for processing multilocus
780 metabarcode datasets. *Methods in Ecology and Evolution*, *10*, 1469– 1475.
781 https://doi.org/https://doi.org/10.1111/2041-210X.13214
782 Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer,
783 S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E., & Bernatchez, L. (2017).
784 Environmental DNA metabarcoding: Transforming how we survey animal and plant
785 communities. *Molecular Ecology*, *26*(21), 5872–5895. https://doi.org/10.1111/mec.14350
786 Djurhuus, A., Closek, C. J., Kelly, R. P., Pitz, K. J., Michisaki, R. P., Starks, H. A., Walz, K. R.,
787 Andruszkiewicz, E. A., Olesin, E., & Hubbard, K. (2020). Environmental DNA reveals
788 seasonal shifts and potential interactions in a marine community. *Nature Communications*,
789 *11*(1), 1–9.
790 Duke, E. M., & Burton, R. S. (2020). Efficacy of metabarcoding for identification of fish eggs
791 evaluated with mock communities. *Ecology and Evolution*, *10*(7), 3463–3476.
792 https://doi.org/10.1002/ece3.6144
793 Edgar, R. C. (2018a). Accuracy of taxonomy prediction for 16S rRNA and fungal ITS
794 sequences. *PeerJ*, *6*, e4652.
795 Edgar, R. C. (2018b). Taxonomy annotation and guide tree errors in 16S rRNA databases. *PeerJ*,
796 *6*, e5030.
797 Edgar, R. C. (2018c). Updating the 97% identity threshold for 16S ribosomal RNA OTUs.
798 *Bioinformatics*, *34*(14), 2371–2375.
799 Ekstrom, J. A. (2009). California Current Large Marine Ecosystem: Publicly available dataset of
800 state and federal laws and regulations. *Marine Policy*, *33*(3), 528–531.
801 Ficetola, G. F., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessière, J., Taberlet, P., &
802 Pompanon, F. (2010). An in silico approach for the evaluation of DNA barcodes. *BMC*
803 *Genomics*, *11*(1), 434.
804 Froese, R., & Pauly, D. (2010). *FishBase*. Fisheries Centre, University of British Columbia.

Gao, X., Lin, H., Revanna, K., & Dong, Q. (2017). A Bayesian taxonomic classification method for 16S rRNA gene sequences with improved species-level accuracy. *BMC Bioinformatics*, *18*(1), 247.

Gentemann, C. L., Fewings, M. R., & García-Reyes, M. (2017). Satellite sea surface temperatures along the West Coast of the United States during the 2014–2016 northeast Pacific marine heat wave. *Geophysical Research Letters*, *44*(1), 312–319.

Gold, Zachary, Sprague, J., Kushner, D. J., Zerecero Marin, E., & Barber, P. H. (2021). eDNA metabarcoding as a biomonitoring tool for marine protected areas. *PLoS Biology*, *In press*.

Gold, Zack. (2020). *zjgold/FishCARD: California Fish Reference Database Version 0.0*. Zenodo. https://doi.org/10.5281/zenodo.4315278

Good, T. P., Samhouri, J. F., Feist, B. E., Wilcox, C., & Jahncke, J. (2020). Plastics in the Pacific: Assessing risk from ocean debris for marine birds in the California Current Large Marine Ecosystem. *Biological Conservation*, *250*, 108743.

Goodwin, K. D., Thompson, L. R., Duarte, B., Kahlke, T., Thompson, A. R., Marques, J. C., & Caçador, I. (2017). DNA sequencing as a tool to monitor marine ecological status. *Frontiers in Marine Science*, *4*, 107.

Guo, J. (2017). *Metabarcoding Analyses of Gut Microbiome Compositions in Red Abalone (Haliotis Rufescens, Swainson, 1822) Fed Different Macroalgal Diets*.

Halpern, B. S., Kappel, C. V, Selkoe, K. A., Micheli, F., Ebert, C. M., Kontgis, C., Crain, C. M., Martone, R. G., Shearer, C., & Teck, S. J. (2009). Mapping cumulative human impacts to California Current marine ecosystems. *Conservation Letters*, *2*(3), 138–148.

Harvell, C. D., Montecino-Latorre, D., Caldwell, J. M., Burt, J. M., Bosley, K., Keller, A., Heron, S. F., Salomon, A. K., Lee, L., Pontier, O., Pattengill-Semmens, C., & Gaydos, J. K. (2019). Disease epidemic and a marine heat wave are associated with the continental-scale collapse of a pivotal predator (Pycnopodia helianthoides). *Science Advances*, *5*(1), eaau7042. https://doi.org/10.1126/sciadv.aau7042

Hassanin, A., Bonillo, C., Nguyen, B. X., & Cruaud, C. (2010). Comparisons between mitochondrial genomes of domestic goat (Capra hircus) reveal the presence of numts and multiple sequencing errors. *Mitochondrial DNA*, *21*(3–4), 68–76.

Hastings, P. A., & Burton, R. S. (2008). *Establishing a DNA sequence database for the marine fish fauna of California*.

Heller, P., Casaletto, J., Ruiz, G., & Geller, J. (2018). A database of metazoan cytochrome c oxidase subunit I gene sequences derived from GenBank with CO-ARBitrator. *Scientific Data*, *5*.

Hofmann, G. E., Evans, T. G., Kelly, M. W., Padilla-Gamiño, J. L., Blanchette, C. A., Washburn, L., Chan, F., McManus, M. A., Menge, B. A., & Gaylord, B. (2014). Exploring local adaptation and the ocean acidification seascape–studies in the California Current Large Marine Ecosystem. *Biogeosciences*, *11*(4).

Hyde, J. R., & Vetter, R. D. (2007). The origin, evolution, and diversification of rockfishes of the genus Sebastes (Cuvier). *Molecular Phylogenetics and Evolution*, *44*(2), 790–811. https://doi.org/10.1016/j.ympev.2006.12.026

Ingram, T., & Kai, Y. (2014). The geography of morphological convergence in the radiations of Pacific Sebastes rockfishes. *The American Naturalist*, *184*(5), E115-31. https://doi.org/10.1086/678053

Iwasaki, W., Fukunaga, T., Isagozawa, R., Yamada, K., Maeda, Y., Satoh, T. P., Sado, T., Mabuchi, K., Takeshima, H., & Miya, M. (2013). MitoFish and MitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. *Molecular Biology and Evolution*, *30*(11), 2531–2540.

Jo, T., Murakami, H., Masuda, R., Sakata, M. K., Yamamoto, S., & Minamoto, T. (2017). Rapid degradation of longer DNA fragments enables the improved estimation of distribution and biomass using environmental DNA. *Molecular Ecology Resources*, *17*(6), e25–e33.

Kelly, R. P., Port, J. A., Yamahara, K. M., & Crowder, L. B. (2014). Using environmental DNA to census marine fishes in a large mesocosm. *PloS One*, *9*(1), e86175. https://doi.org/10.1371/journal.pone.0086175

Kelly, R. P., Port, J. a., Yamahara, K. M., Martone, R. G., Lowell, N., Thomsen, P. F., Mach, M. E., Bennett, M., Prahler, E., Caldwell, M. R., & Crowder, L. B. (2014). Harnessing DNA to improve environmental management. *Science*, *344*(6191). https://doi.org/10.1126/science.1251156

Klymus, K. E., Marshall, N. T., & Stepien, C. A. (2017). Environmental DNA (eDNA) metabarcoding assays to detect invasive invertebrate species in the Great Lakes. *PLOS ONE*, *12*(5), e0177643. https://doi.org/10.1371/journal.pone.0177643

Koslow, J. A., & Davison, P. C. (2016). Productivity and biomass of fishes in the California Current Large Marine Ecosystem: Comparison of fishery-dependent and-independent time series. *Environmental Development*, *17*, 23–32.

Kozlov, A. M., Zhang, J., Yilmaz, P., Glöckner, F. O., & Stamatakis, A. (2016). Phylogeny-aware identification and correction of taxonomically mislabeled sequences. *Nucleic Acids Research*, *44*(11), 5022–5033.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357.

Lea, R. N., McAllister, R. D., & VenTresca, D. A. (1999). *Biological aspects of nearshore rockfishes of the genus Sebastes from central California: with notes on ecologically related sport fishes* (Vol. 177). State of California, The Resources Agency, Department of Fish and Game.

Lepofsky, D., Armstrong, C. G., Greening, S., Jackley, J., Carpenter, J., Guernsey, B., Mathews, D., & Turner, N. J. (2017). Historical ecology of cultural keystone places of the Northwest Coast. *American Anthropologist*, *119*(3), 448–463.

Leray, M., Boehm, J. T., Mills, S. C., & Meyer, C. P. (2012). Moorea BIOCODE barcode library as a tool for understanding predator–prey interactions: insights into the diet of common predatory coral reef fishes. *Coral Reefs*, *31*(2), 383–388. https://doi.org/10.1007/s00338-011-0845-0

Leray, M., & Knowlton, N. (2015). DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences*, *112*(7), 201424997. https://doi.org/10.1073/pnas.1424997112

Leray, M., & Knowlton, N. (2017). Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina COI metabarcoding. *PeerJ*, *5*, e3006.

Leray, M., Knowlton, N., Ho, S.-L., Nguyen, B. N., & Machida, R. J. (2019). GenBank is a reliable resource for 21st century biodiversity research. *Proceedings of the National Academy of Sciences*, *116*(45), 22651–22656.

Lim, S. J., & Thompson, L. R. (2021). Mitohelper: A mitochondrial reference sequence analysis tool for fish eDNA studies. *Environmental DNA*.

Lodge, D. M., Turner, C. R., Jerde, C. L., Barnes, M. A., Chadderton, L., Egan, S. P., Feder, J. L., Mahon, A. R., & Pfrender, M. E. (2012). Conservation in a cup of water: estimating biodiversity and population abundance from environmental DNA. *Molecular Ecology*, *21*(11), 2555–2558. https://doi.org/10.1111/j.1365-294X.2012.05600.x

Love, M. S., & Passarelli, J. K. (2020). *Miller and Lea's Guide to the Coastal Marine Fishes of California* (2nd.). University of California Agriculture and Natural Resources.

Macheriotou, L., Guilini, K., Bezerra, T. N., Tytgat, B., Nguyen, D. T., Phuong Nguyen, T. X., Noppe, F., Armenteros, M., Boufahja, F., & Rigaux, A. (2019). Metabarcoding free-living marine nematodes using curated 18S and CO1 reference sequence databases for species-level taxonomic assignments. *Ecology and Evolution*, *9*(3), 1211–1226.

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

Min, M. A., Barber, P. H., & Gold, Z. (2020). MiSebastes: An eDNA metabarcoding primer set for rockfishes (genus Sebastes). *BioRxiv*, 2020.10.29.360859. https://doi.org/10.1101/2020.10.29.360859

Minamoto, T., Miya, M., Sado, T., Seino, S., Doi, H., Kondoh, M., Nakamura, K., Takahara, T., Yamamoto, S., & Yamanaka, H. (2020). An illustrated manual for environmental DNA research: Water sampling guidelines and experimental protocols. *Environmental DNA*.

Miya, M., Gotoh, R. O., & Sado, T. (2020). MiFish metabarcoding: a high-throughput approach for simultaneous detection of multiple fish species from environmental DNA and other samples. *Fisheries Science*, 1–32.

Miya, M., & Sado, T. (2019). *Environmental DNA Sampling and Experiment Manual (Version 2.1)*.

Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J. Y., Sato, K., Minamoto, T., Yamamoto, S., Yamanaka, H., Araki, H., Kondoh, M., & Iwasaki, W. (2015). MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. *Royal Society Open Science*, *2*(7), 150088. https://doi.org/10.1098/rsos.150088

Moser, H. G. (1993). *Distributional atlas of fish larvae and eggs in the California Current region: taxa with 1000 or more total larvae, 1951 through 1984* (Issue 31). Marine Life Research Program, Scripps Institution of Oceanography.

Nilsson, R. H., Larsson, K.-H., Taylor, A. F. S., Bengtsson-Palme, J., Jeppesen, T. S., Schigel, D., Kennedy, P., Picard, K., Glöckner, F. O., & Tedersoo, L. (2018). The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Research*, *47*(D1), D259–D264.

Nishimura, D. (2000). Sequencher 3.1. 1. *Biotech Software & Internet Report*, *1*(1–2), 24–30.

NMFS. (2017). *Fisheries economics of the United States, 2015*. NOAA Technical Memorandum.

Nobre, T., Campos, M. D., Lucic-Mercy, E., & Arnholdt-Schmitt, B. (2016). Misannotation awareness: a tale of two gene-groups. *Frontiers in Plant Science*, *7*, 868.

Norgaard, K. M. (2019). *Salmon and acorns feed our people: colonialism, nature, and social action*. Rutgers University Press.

O'Donnell, J. L., Kelly, R. P., Shelton, A. O., Samhouri, J. F., Lowell, N. C., & Williams, G. D. (2017). Spatial distribution of environmental DNA in a nearshore marine habitat. *PeerJ*, *5*, e3044. https://doi.org/10.7717/peerj.3044

Omary, M. B., Eswaraka, J., Kimball, S. D., Moghe, P. V, Panettieri, R. A., & Scotto, K. W. (2020). The COVID-19 pandemic and research shutdown: staying safe and productive. *The Journal of Clinical Investigation*, *130*(6).

Pitz, K. J., Guo, J., Johnson, S. B., Campbell, T. L., Zhang, H., Vrijenhoek, R. C., Chavez, F. P., & Geller, J. (2020). Zooplankton biogeographic boundaries in the California Current System as determined from metabarcoding. *Plos One*, *15*(6), e0235159.

Poloczanska, E. S., Brown, C. J., Sydeman, W. J., Kiessling, W., Schoeman, D. S., Moore, P. J., Brander, K., Bruno, J. F., Buckley, L. B., Burrows, M. T., Duarte, C. M., Halpern, B. S., Holding, J., Kappel, C. V., O'Connor, M. I., Pandolfi, J. M., Parmesan, C., Schwing, F., Thompson, S. A., & Richardson, A. J. (2013). Global imprint of climate change on marine life. *Nature Climate Change*, *3*(10), 919–925. https://doi.org/10.1038/nclimate1958

Pondella II, D. J., Caselle, J. E., Claisse, J. T., Williams, J. P., Davis, K., Williams, C. M., & Zahn, L. A. (2015). *Baseline Characterization of the Shallow Rocky Reef and Kelp Forest Ecosystems of the South Coast Study Region*. https://caseagrant.ucsd.edu/news/summaries-of-projects-selected-for-funding-through-the-south-coast-mpa-baseline-program

Pondella II, D. J., Piacenza, S. E., Claisse, J. T., Williams, C. M., Williams, J. P., Zellmer, A. J., & Caselle, J. E. (2019). Assessing drivers of rocky reef fish biomass density from the Southern California Bight. *Marine Ecology Progress Series*, *628*, 125–140.

Port, J. A., O'Donnell, J. L., Romero-Maraccini, O. C., Leary, P. R., Litvin, S. Y., Nickols, K. J., Yamahara, K. M., & Kelly, R. P. (2015). Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. *Molecular Ecology*. https://doi.org/https://doi.org/10.1111/mec.13481

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *33*(suppl_1), D501–D504.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), D590–D596.

Ransome, E., Geller, J. B., Timmers, M., Leray, M., Mahardini, A., Sembiring, A., Collins, A. G., & Meyer, C. P. (2017). The importance of standardization for biodiversity comparisons: A case study using autonomous reef monitoring structures (ARMS) and metabarcoding to measure cryptic diversity on Mo'orea coral reefs, French Polynesia. *PloS One*, *12*(4).

Richardson, R. T., Bengtsson-Palme, J., Gardiner, M. M., & Johnson, R. M. (2018). A reference cytochrome c oxidase subunit I database curated for hierarchical classification of arthropod metabarcoding data. *PeerJ*, *6*, e5126.

Roger, A. J., Muñoz-Gómez, S. A., & Kamikawa, R. (2017). The origin and diversification of mitochondria. *Current Biology*, *27*(21), R1177–R1192.

Rogers-Bennett, L., & Catton, C. A. (2019). Marine heat wave and multiple stressors tip bull kelp forest to sea urchin barrens. *Scientific Reports*, *9*(1), 1–9. https://doi.org/10.1038/s41598-019-51114-y

Samhouri, J. F., Andrews, K. S., Fay, G., Harvey, C. J., Hazen, E. L., Hennessey, S. M.,

Holsman, K., Hunsicker, M. E., Large, S. I., & Marshall, K. N. (2017). Defining ecosystem thresholds for human activities and environmental pressures in the California Current. *Ecosphere*, *8*(6), e01860.

Sanders, J. G., Beichman, A. C., Roman, J., Scott, J. J., Emerson, D., McCarthy, J. J., & Girguis, P. R. (2015). Baleen whales host a unique gut microbiome with similarities to both carnivores and herbivores. *Nature Communications*, *6*, 8285.

Sanford, E., Sones, J. L., García-Reyes, M., Goddard, J. H. R. R., & Largier, J. L. (2019). Widespread shifts in the coastal biota of northern California during the 2014–2016 marine heatwaves. *Scientific Reports*, *9*(1), 4216. https://doi.org/10.1038/s41598-019-40784-3

Santora, J. A., Mantua, N. J., Schroeder, I. D., Field, J. C., Hazen, E. L., Bograd, S. J., Sydeman, W. J., Wells, B. K., Calambokidis, J., Saez, L., Lawson, D., & Forney, K. A. (2020). Habitat compression and ecosystem shifts as potential links between marine heatwave and record whale entanglements. *Nature Communications*, *11*(1), 1–12. https://doi.org/10.1038/s41467-019-14215-w

Sato, Y., Miya, M., Fukunaga, T., Sado, T., & Iwasaki, W. (2018). MitoFish and MiFish pipeline: a mitochondrial genome database of fish with an analysis pipeline for environmental DNA metabarcoding. *Molecular Biology and Evolution*, *35*(6), 1553–1555.

Schenekar, T., Schletterer, M., Lecaudey, L. A., & Weiss, S. J. (2020). Reference databases, primer choice, and assay sensitivity for environmental metabarcoding: Lessons learnt from a re-evaluation of an eDNA fish assessment in the Volga headwaters. *River Research and Applications*.

Sherman, K. (1991). The large marine ecosystem concept: research and management strategy for living marine resources. *Ecological Applications*, *1*(4), 349–360.

Siegwald, L., Touzet, H., Lemoine, Y., Hot, D., Audebert, C., & Caboche, S. (2017). Assessment of Common and Emerging Bioinformatics Pipelines for Targeted Metagenomics. *PloS One*, *12*(1). https://doi.org/10.1371/journal.pone.0169563

Stat, M., Huggett, M. J., Bernasconi, R., DiBattista, J. D., Berry, T. E., Newman, S. J., Harvey, E. S., & Bunce, M. (2017). Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment. *Scientific Reports*, *7*(1), 12240. https://doi.org/10.1038/s41598-017-12501-5

Stoeckle, M. Y., Das Mishu, M., & Charlop-Powers, Z. (2020). Improved environmental DNA reference library detects overlooked marine fishes in New Jersey, United States. *Frontiers in Marine Science*, *7*, 226.

Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., Kulikova, T., Leinonen, R., Lin, Q., & Lombard, V. (2002). The EMBL nucleotide sequence database. *Nucleic Acids Research*, *30*(1), 21–26.

Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L. H. (2012). Environmental DNA. *Molecular Ecology*, *21*(8), 1789–1793. https://doi.org/10.1111/j.1365-294X.2012.05542.x

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*(8), 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Thompson, A. R., Chen, D. C., Guo, L. W., Hyde, J. R., & Watson, W. (2017). Larval abundances of rockfishes that were historically targeted by fishing increased over 16 years in association with a large marine protected area. *Royal Society Open Science*, *4*(9), 170639.

https://doi.org/10.1098/rsos.170639
Thomsen, P. F., Møller, P. R., Sigsgaard, E. E., Knudsen, S. W., Jørgensen, O. A., & Willerslev, E. (2016). Environmental DNA from seawater samples correlate with trawl catches of subarctic, deepwater fishes. *PLoS ONE*, *11*(11). https://doi.org/10.1371/journal.pone.0165252
Valsecchi, E., Bylemans, J., Goodman, S. J., Lombardi, R., Carr, I., Castellano, L., Galimberti, A., & Galli, P. (2019). Novel Universal Primers for Metabarcoding eDNA Surveys of Marine Mammals and Other Marine Vertebrates. *BioRxiv*, 759746.
Wadewitz, L. K. (2012). *The nature of borders: Salmon, boundaries, and bandits on the Salish Sea*. University of Washington Press.
Wakeling, M. N., Laver, T. W., Colclough, K., Parish, A., Ellard, S., & Baple, E. L. (2019). Misannotation of multiple-nucleotide variants risks misdiagnosis. *Wellcome Open Research*, *4*.
Walker, H. J., Hastings, P. A., Hyde, J. R., Lea, R. N., Snodgrass, O. E., & Bellquist, L. F. (2020). Unusual occurrences of fishes in the Southern California Current System during the warm water period of 2014–2018. *Estuarine, Coastal and Shelf Science*, *236*, 106634. https://doi.org/10.1016/j.ecss.2020.106634
Walsh, P. S., Metzger, D. A., & Higuchi, R. (1991). Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. *Biotechniques*, *10*(4), 506–513.
Ward, R. D., Hanner, R., & Hebert, P. D. N. (2009). The campaign to DNA barcode all fishes, FISH-BOL. *Journal of Fish Biology*, *74*(2), 329–356. https://doi.org/10.1111/j.1095-8649.2008.02080.x
Wells, R. J. D., Mohan, J. A., Dewar, H., Rooker, J. R., Tanaka, Y., Snodgrass, O. E., Kohin, S., Miller, N. R., & Ohshimo, S. (2020). Natal origin of Pacific bluefin tuna from the California Current Large Marine Ecosystem. *Biology Letters*, *16*(2), 20190878.
Williams, G. D., Levin, P. S., & Palsson, W. A. (2010). Rockfish in Puget Sound: An ecological history of exploitation. *Marine Policy*, *34*(5), 1010–1020. https://doi.org/https://doi.org/10.1016/j.marpol.2010.02.008
Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46. https://doi.org/10.1186/gb-2014-15-3-r46
Yamahara, K. M., Preston, C. M., Birch, J. M., Walz, K. R., Marin III, R., Jensen, S., Pargett, D., Roman, B., Zhang, Y., & Ryan, J. (2019). In-situ Autonomous Acquisition and Preservation of Marine Environmental DNA Using an Autonomous Underwater Vehicle. *Frontiers in Marine Science*, *6*, 373. https://doi.org/https://doi.org/10.3389/fmars.2019.00373
Yamamoto, S., Masuda, R., Sato, Y., Sado, T., Araki, H., Kondoh, M., Minamoto, T., & Miya, M. (2017). Environmental DNA metabarcoding reveals local fish communities in a species-rich coastal sea. *Scientific Reports*, *7*, 40368.

DATA ACCESSIBILITY

Reference databases and metabarcoding data are publicly available and stored on a Dryad repository (https://doi.org/10.5068/D1H963). All reference barcode sequences have been uploaded to GenBank (BioProject PRJNA731549). Additional supporting information is available at https://github.com/zjgold/FishCARD.

AUTHOR CONTRIBUTIONS

- Conceptualization ZG, ESC, DK, BF, RSB, KDG, ART, PHB, HJW
- Performed Research ZG, ESC, DK, BF, ART
- Funding Acquisition ZG, PHB, ART, KDG, DK, RSB
- Data Curation ZG, ESC, DK, BF, HJW
- Formal Analysis ZG, EEC
- Writing – Original Draft Preparation ZG, EEC, ESC, DK, BF, RSB, KDG, ART, PHB, HJW

CONFLICS OF INTEREST

The authors have no conflicts of interest to report.

TABLES

Table 1. Summary of Cross Validation Results. Comparison of performance metrics for taxonomic assignments using the global database as a reference to annotate sequences in the global database (global-global)[(test database-training database)], the regional database (global-regional), and using the regional database as a reference to annotate sequences in itself (regional-regional). Reporting metrics calculated using a taxonomic cutoff score of 60.

Metric	Global-Global	Global-Regional	Regional-Regional
Under-classification Rate	8.6%	11.8%	7.8%
Misclassification Rate	1.7%	1.8%	1.3%
Over-classification Rate	0.0%	0.0%	0.0%
Accuracy	89.7%	86.5%	90.9%
True Positive Rate	89.7%	86.5%	90.9%
Sensitivity	91.3%	88.0%	92.1%
Specificity	98.3%	98.2%	98.7%

Table 2. Summary of Seawater eDNA Metabarcoding Taxonomic Assignments for Tested Reference Databases.

Metric		Reference Database		
		CRUX-GenBank	**Global**	**Regional**
Database	Reference Barcode Origin	GenBank	GenBank + Generated	GenBank + Generated
	Species Included	**All**	**All**	**California Fishes**
Reads	Total Reads	330,877		
	Assigned to NA	81,014	81,002	81,006
	Assigned to Class Level	54,090	-	-
	Assigned to Order Level	727	-	-
	Assigned to Family Level	1,286	1,409	131
	Assigned to Genus Level	952	1,068	1,063
	Assigned to Species Level	192,808	247,398	248,677
ASVs	Total ASVs	341		
	Assigned to NA	172	169	170
	Assigned to Class Level	12	-	-
	Assigned to Order Level	3	-	-
	Assigned to Family Level	5	13	11
	Assigned to Genus Level	4	6	4
	Assigned to Species Level	145	153	156
Taxonomy	Unique Families Identified	31	28	27
	Unique Genera Identified	39	38	39
	Unique Species Identified	38	38	37
	CA Native Species	25	36	37
	Avg. ASVs Per Species	3.8	4.1	4.2

FIGURES

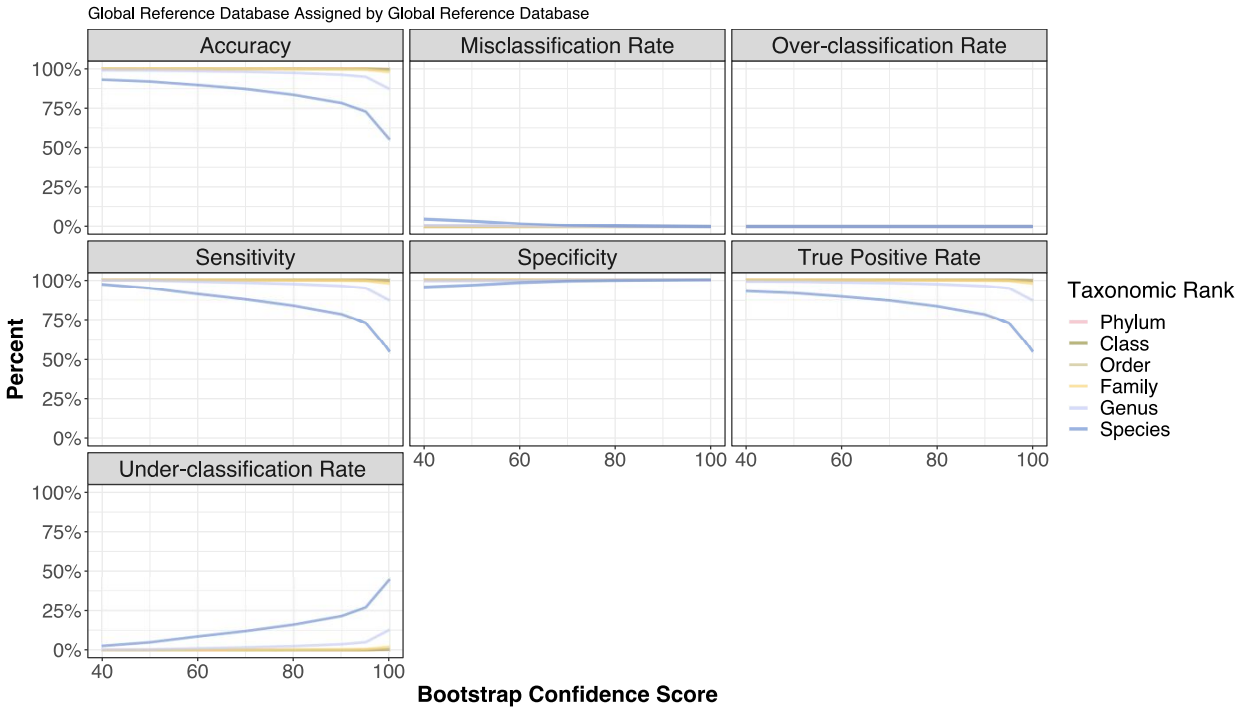
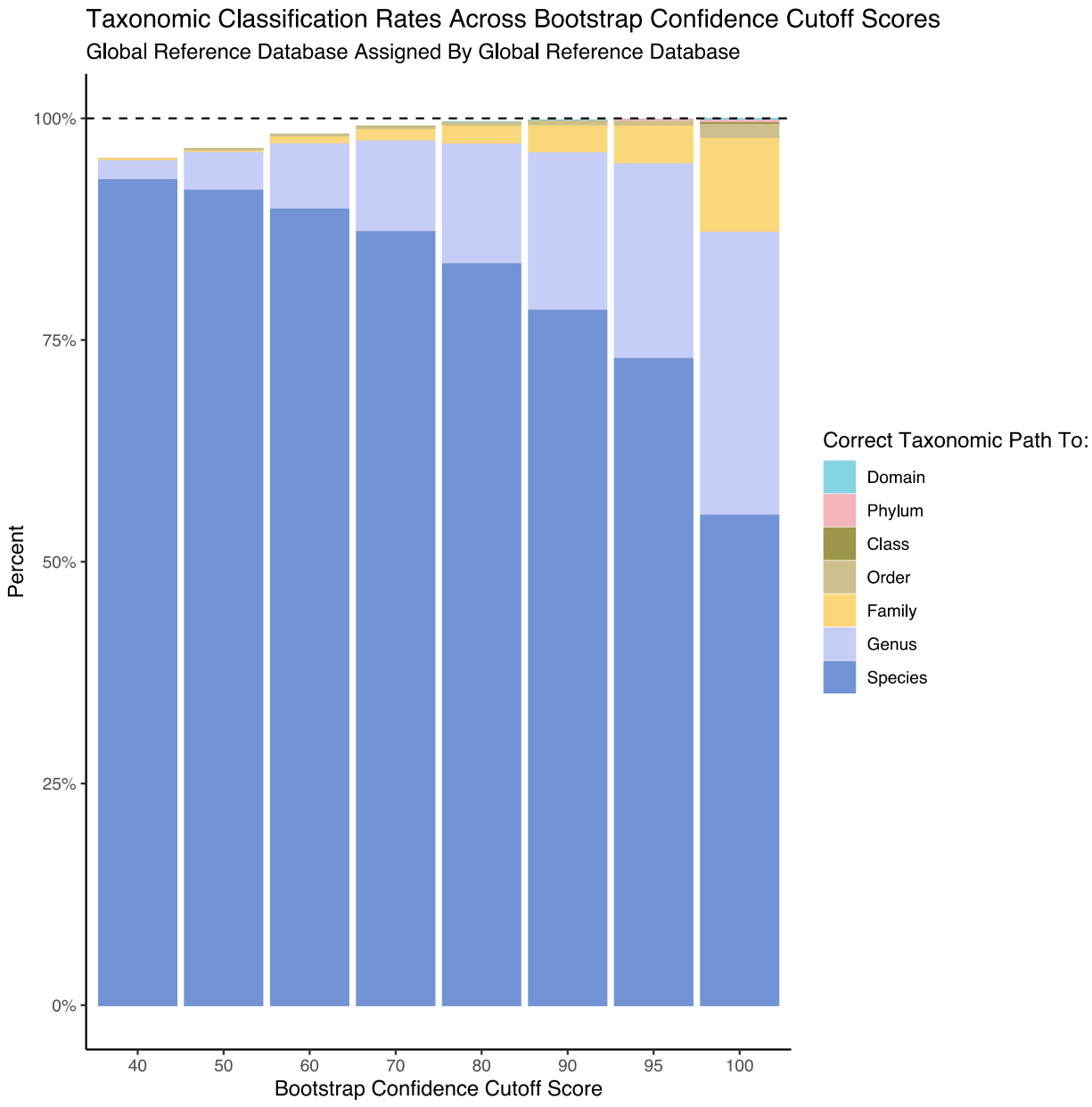


Figure 1. Effect of TAXXI Bootstrap Confidence Cutoff Scores on Taxonomic Assignment Metrics. Taxonomy cross-validation by identity (TAXXI) results for taxonomic assignments generated by using the global database as a reference to annotate the sequences in that same database. Accuracy, true positive rate, sensitivity, and misclassification increased with relaxed bootstrap confidence cutoff scores. Under-classification and specificity decreased with relaxed bootstrap confidence cutoff scores. Results for each taxonomic rank are colored.



1103

1104 **Figure 2. Taxonomic Classification Rates Across Bootstrap Confidence Cutoff Scores.**

1105 Results from taxonomy cross-validation by identity (TAXXI) using the global database as the reference to assign

1106 taxonomy to all sequences in that database. Correct species level matches increase with more relaxed bootstrap

1107 confidence cutoff scores. Correct taxonomic level matches are colored by the lowest common ancestor match.

1108 Dotted line indicates 100% and all mismatches were excluded.

1109

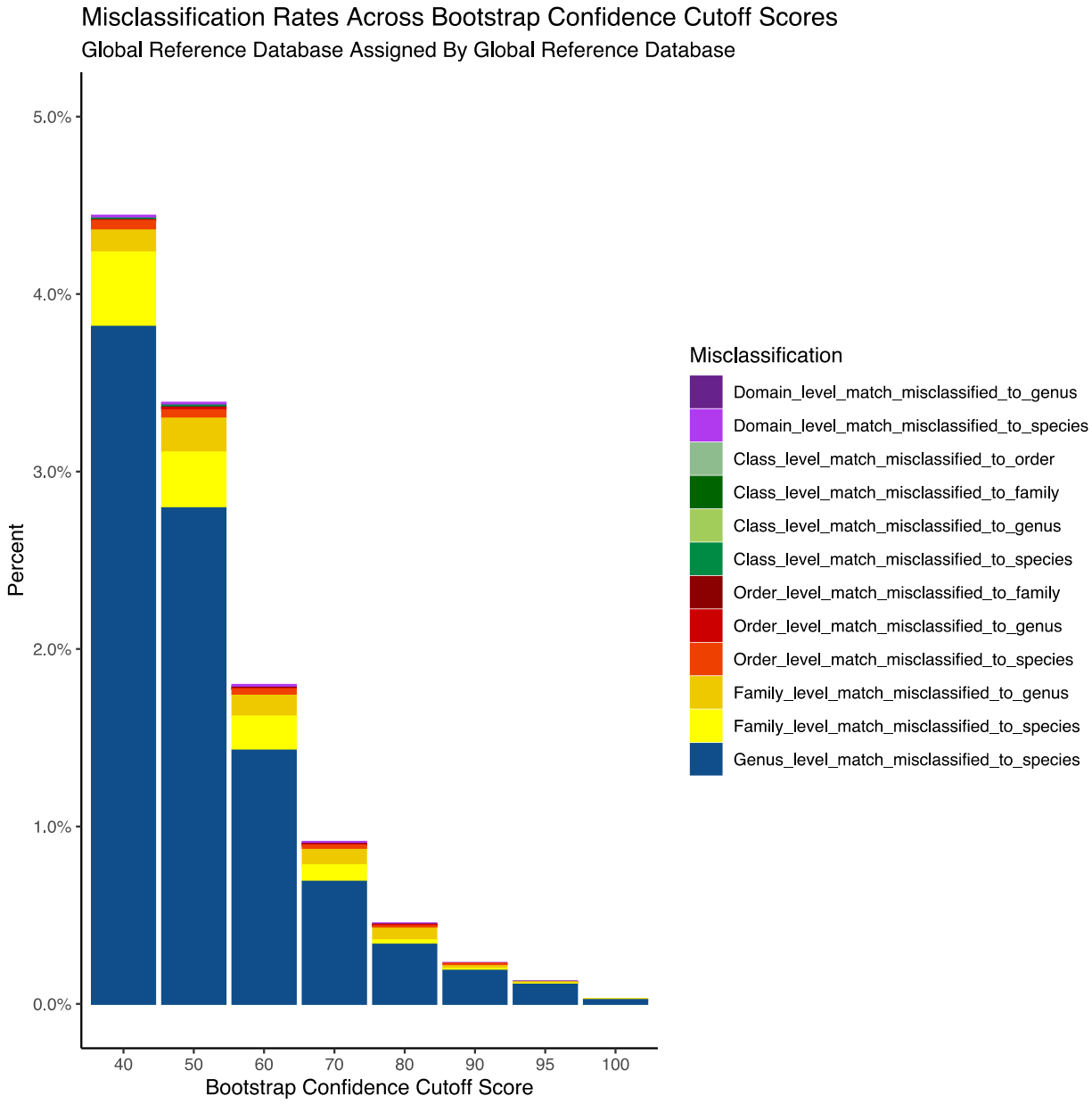


Figure 3. Misclassification Rates Across Bootstrap Confidence Cutoff Scores. Results from taxonomy cross-validation by identity (TAXXI) using the global reference database to assign taxonomy to all sequences in that database. Misclassification increased with relaxed bootstrap confidence cutoff scores. Misclassification types are colored.

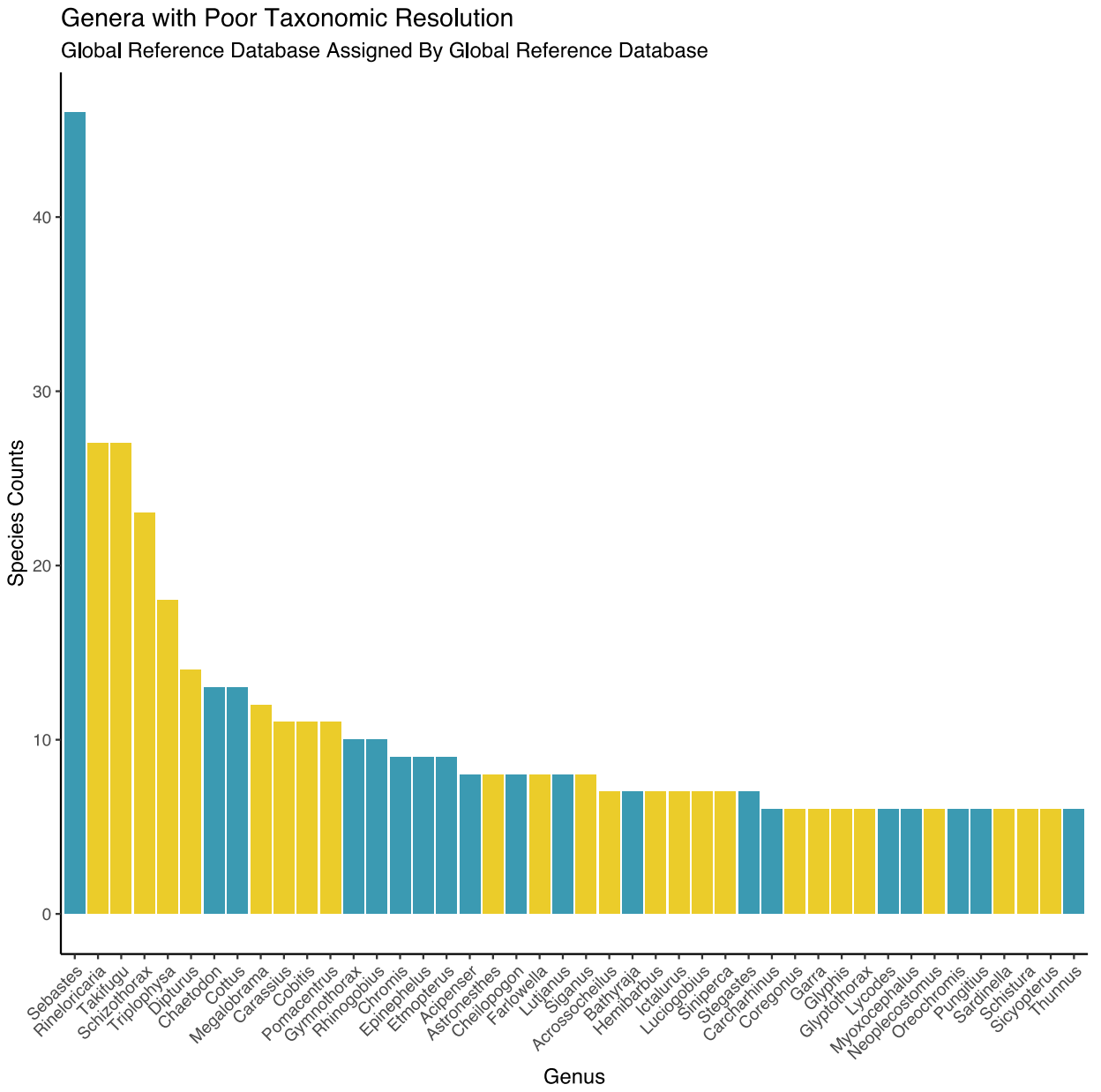


Figure 4. Genera with Poor Taxonomic Resolution. Genera poorly resolved to the species level by the MiFish *12S* barcode based on results from taxonomy cross-validation by identity (TAXXI) using the global

reference database to assign taxonomy to all sequences in that database using a bootstrap confidence cutoff of 60. Genera in blue occur in the California Current Large Marine Ecosystem.

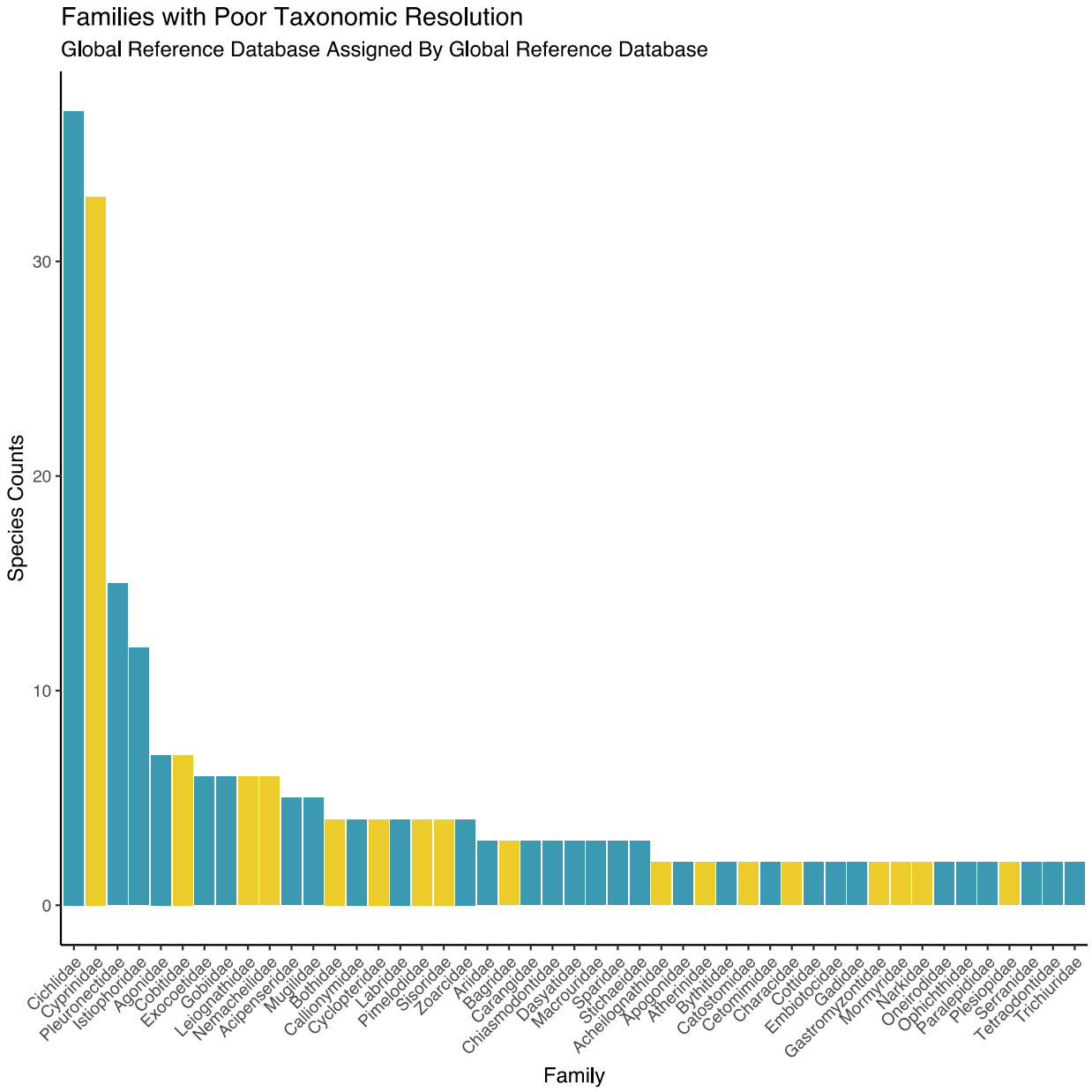


Figure 5. Families with Poor Taxonomic Resolution. Families poorly resolved to the species level by the MiFish *12S* barcode based on results from taxonomy cross-validation by identity (TAXXI) using the global

reference database to assign taxonomy to all sequences in that database using a bootstrap confidence cutoff (BCC) of 60. Families in blue that occur in the California Current Large Marine Ecosystem.

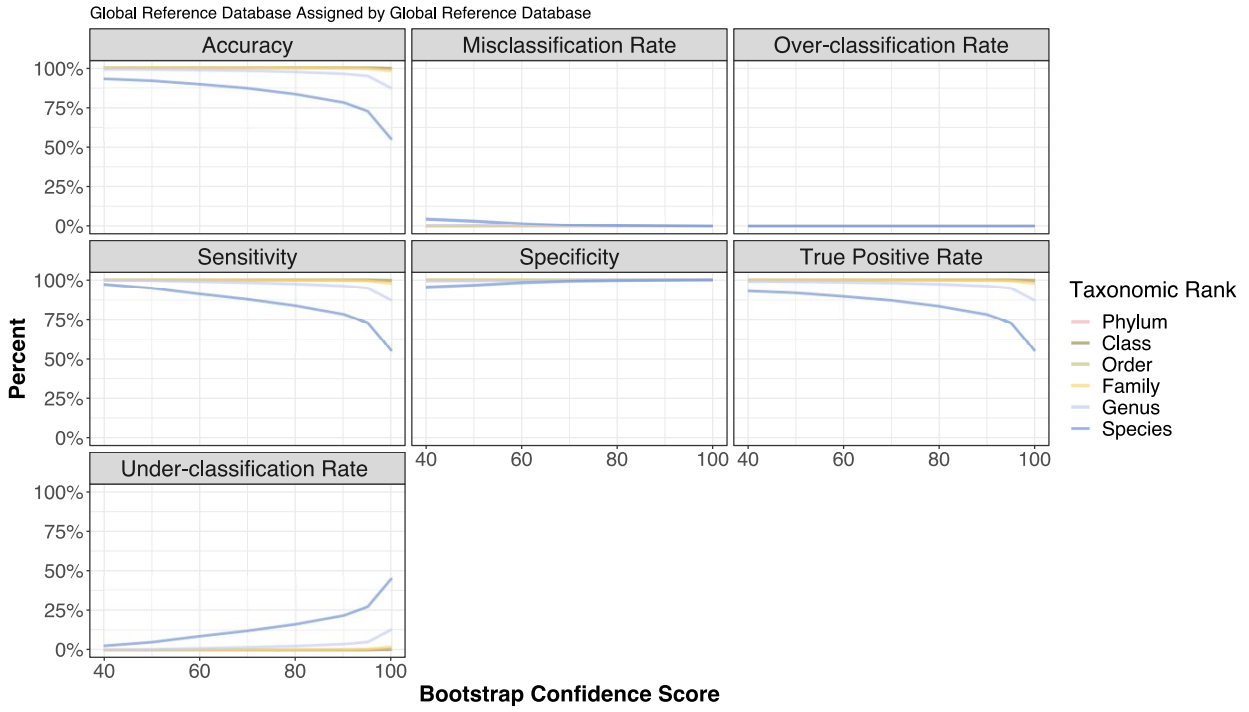


Figure 1. Effect of TAXXI Bootstrap Confidence Cutoff Scores on Taxonomic Assignment Metrics. Taxonomy cross-validation by identity (TAXXI) results for taxonomic assignments generated by using the global database as a reference to annotate the sequences in that same database. Accuracy, true positive rate, sensitivity, and misclassification increased with relaxed bootstrap confidence cutoff scores. Under-classification and specificity decreased with relaxed bootstrap confidence cutoff scores. Results for each taxonomic rank are colored.

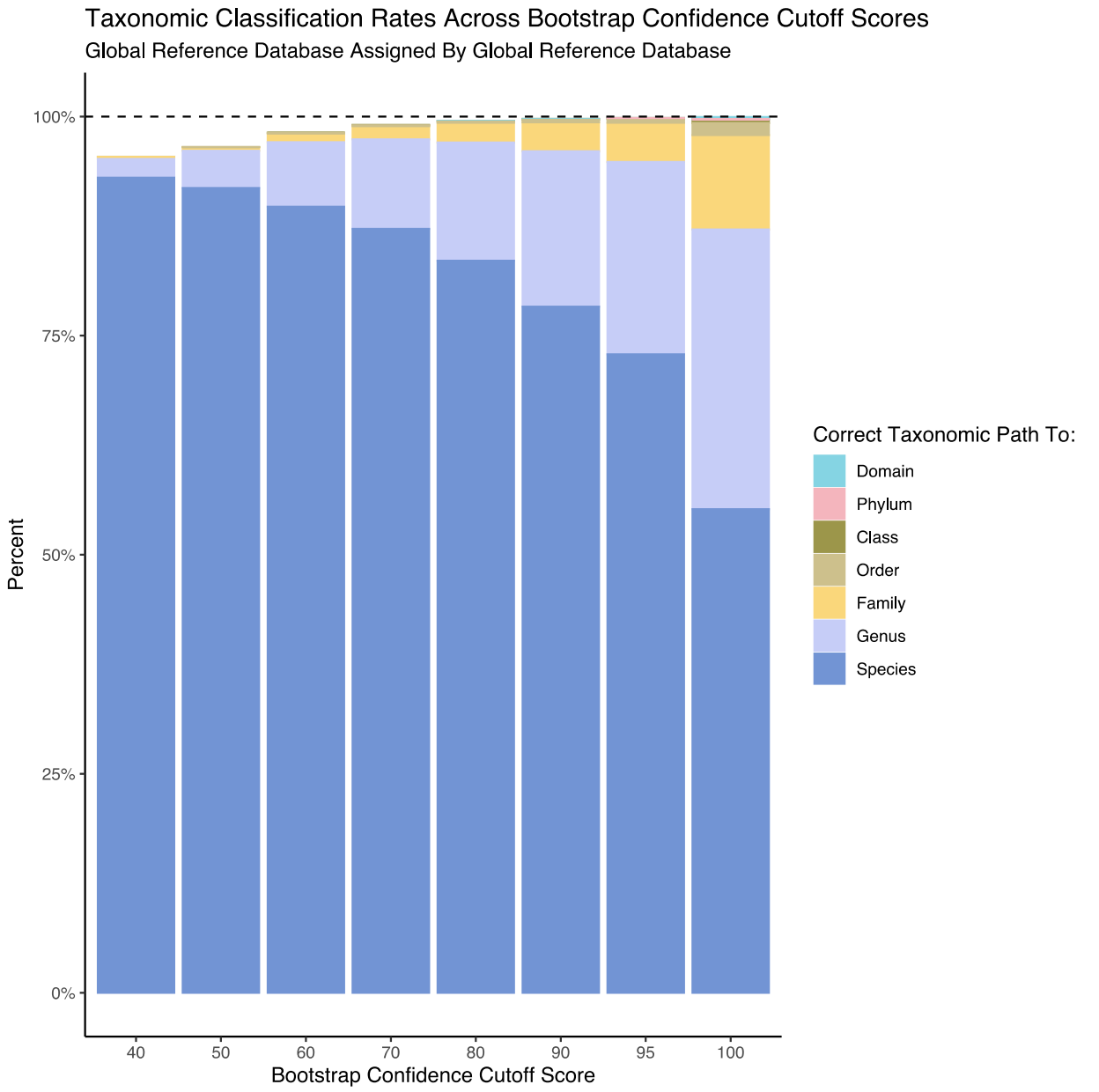


Figure 2. Taxonomic Classification Rates Across Bootstrap Confidence Cutoff Scores.

Results from taxonomy cross-validation by identity (TAXXI) using the global database as the reference to assign taxonomy to all sequences in that database. Correct species level matches increase with more relaxed bootstrap confidence cutoff scores. Correct taxonomic level matches are colored by the lowest common ancestor match. Dotted line indicates 100% and all mismatches were excluded.

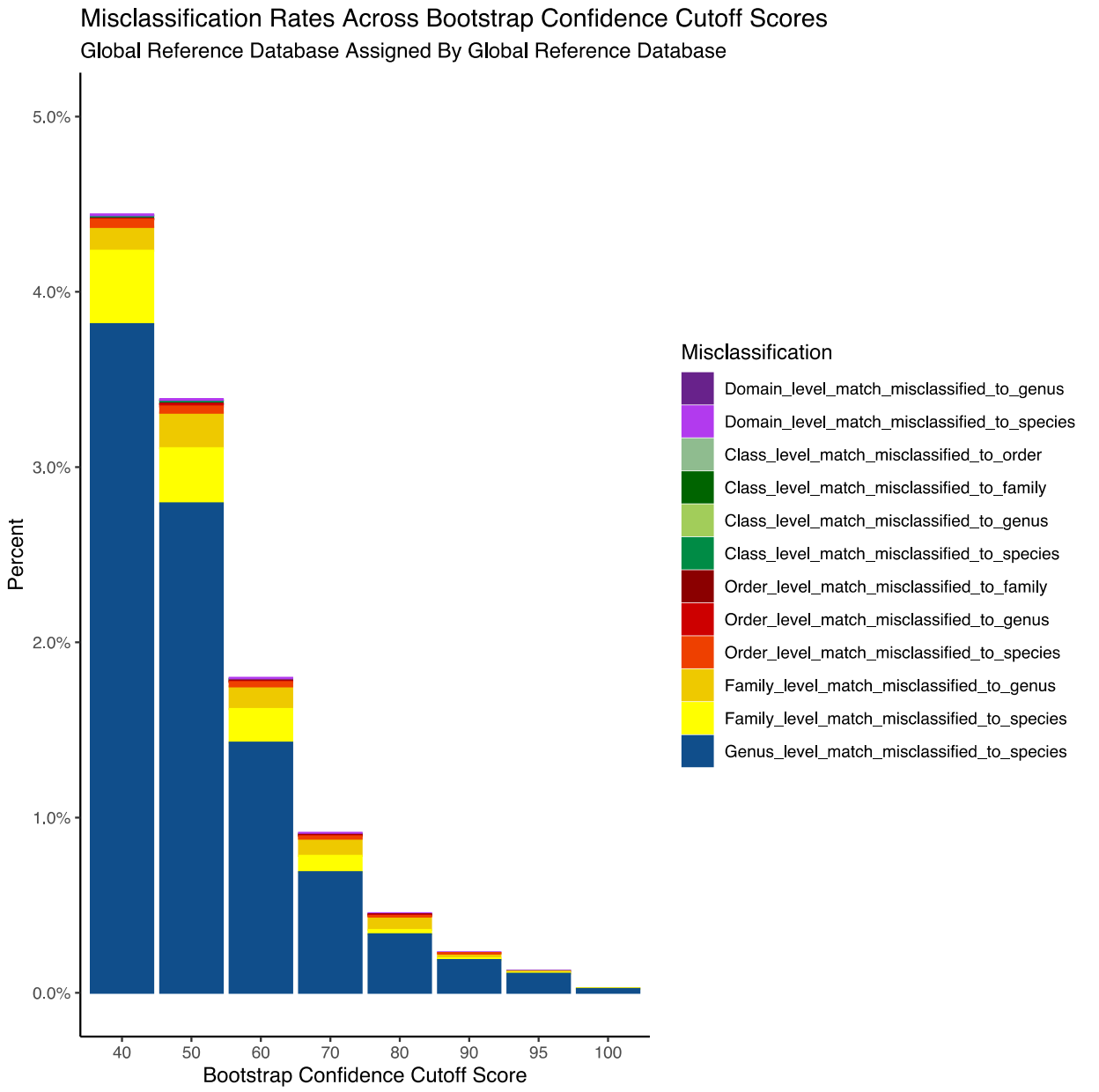


Figure 3. Misclassification Rates Across Bootstrap Confidence Cutoff Scores. Results from taxonomy cross-validation by identity (TAXXI) using the global reference database to assign taxonomy to all sequences in that database. Misclassification increased with relaxed bootstrap confidence cutoff scores. Misclassification types are colored.

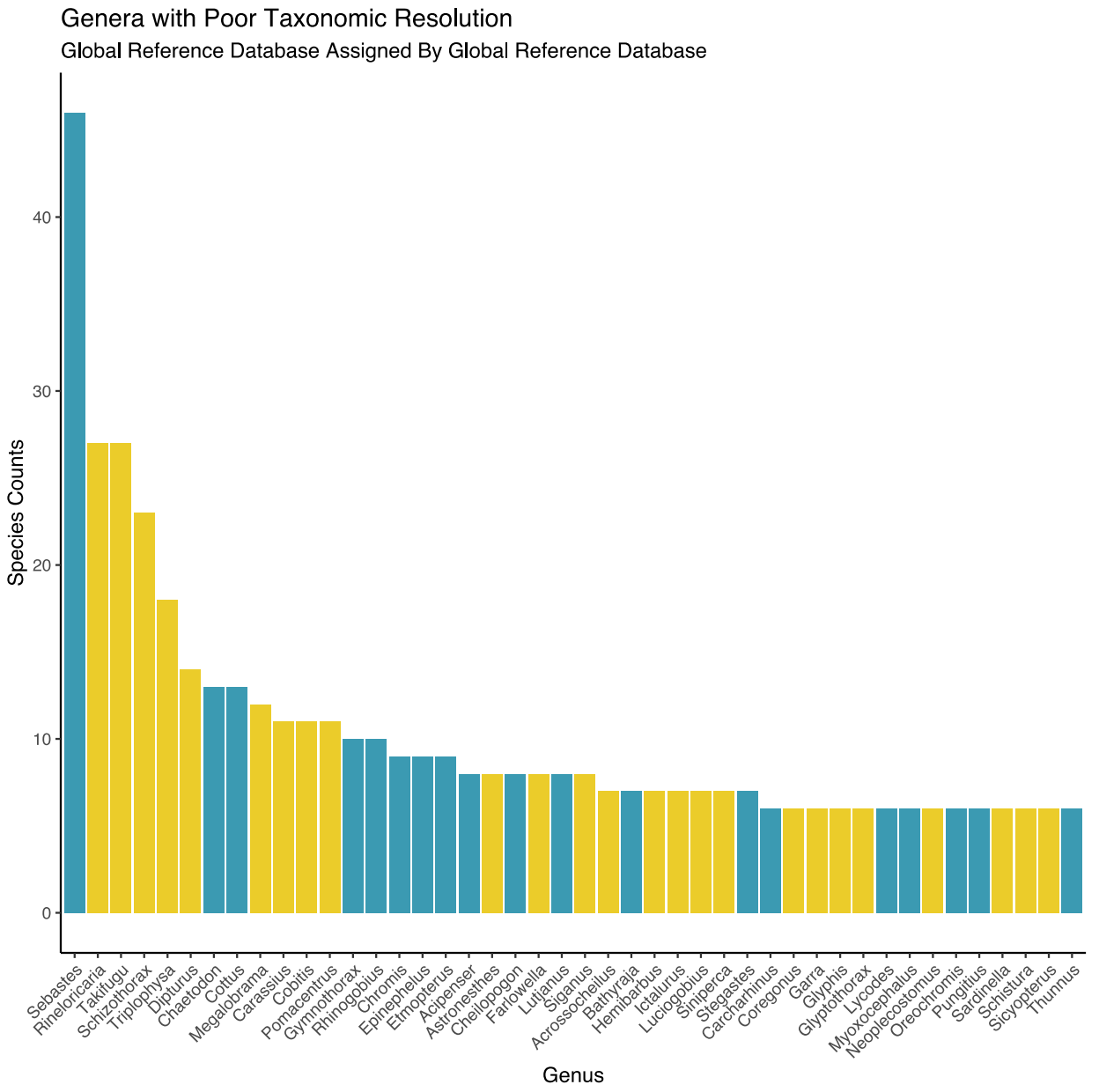


Figure 4. Genera with Poor Taxonomic Resolution. Genera poorly resolved to the species level by the MiFish *12S* barcode based on results from taxonomy cross-validation by identity (TAXXI) using the global reference database to assign taxonomy to all sequences in that database using a bootstrap confidence cutoff of 60. Genera in blue occur in the California Current Large Marine Ecosystem.

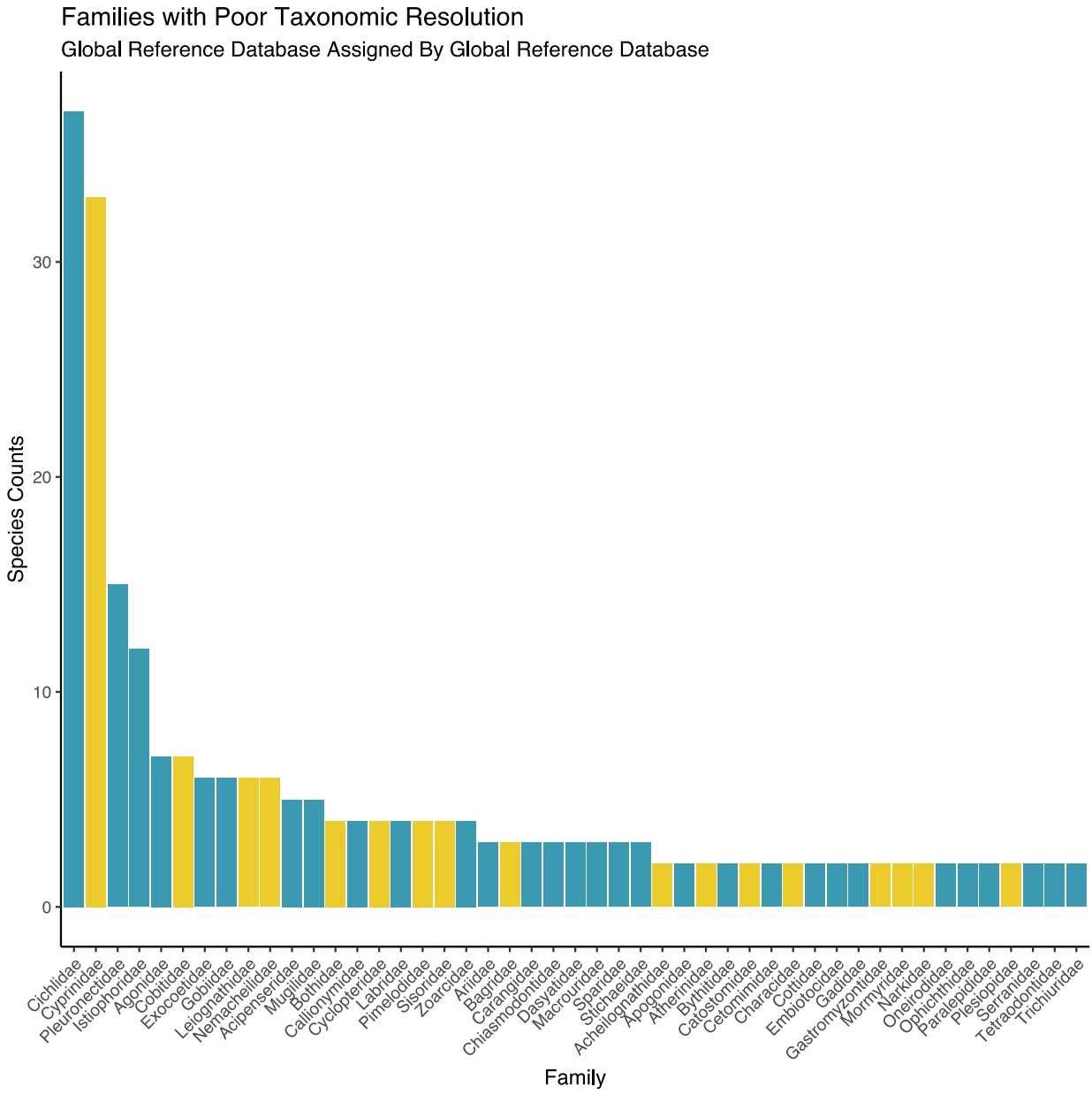


Figure 5. Families with Poor Taxonomic Resolution. Families poorly resolved to the species level by the MiFish *12S* barcode based on results from taxonomy cross-validation by identity (TAXXI) using the global reference database to assign taxonomy to all sequences in that database using a bootstrap confidence cutoff (BCC) of 60. Families in blue that occur in the California Current Large Marine Ecosystem.

Table 1. Summary of Cross Validation Results. Comparison of performance metrics for taxonomic assignments using the global database as a reference to annotate sequences in the global database (global-global)[(test database-training database)], the regional database (global-regional), and using the regional database as a reference to annotate sequences in itself (regional-regional). Reporting metrics calculated using a taxonomic cutoff score of 60.

Metric	Global-Global	Global-Regional	Regional-Regional
Under-classification Rate	8.6%	11.8%	7.8%
Misclassification Rate	1.7%	1.8%	1.3%
Over-classification Rate	0.0%	0.0%	0.0%
Accuracy	89.7%	86.5%	90.9%
True Positive Rate	89.7%	86.5%	90.9%
Sensitivity	91.3%	88.0%	92.1%
Specificity	98.3%	98.2%	98.7%

Table 2. Summary of Seawater eDNA Metabarcoding Taxonomic Assignments for Tested Reference Databases.

Metric		Reference Database		
		CRUX-GenBank	**Global**	**Regional**
Database	Reference Barcode Origin	GenBank	GenBank + Generated	GenBank + Generated
	Species Included	**All**	**All**	**California Fishes**
Reads	Total Reads	330,877		
	Assigned to NA	81,014	81,002	81,006
	Assigned to Class Level	54,090	-	-
	Assigned to Order Level	727	-	-
	Assigned to Family Level	1,286	1,409	131
	Assigned to Genus Level	952	1,068	1,063
	Assigned to Species Level	192,808	247,398	248,677
ASVs	Total ASVs	341		
	Assigned to NA	172	169	170
	Assigned to Class Level	12	-	-
	Assigned to Order Level	3	-	-
	Assigned to Family Level	5	13	11
	Assigned to Genus Level	4	6	4
	Assigned to Species Level	145	153	156
Taxonomy	Unique Families Identified	31	28	27
	Unique Genera Identified	39	38	39
	Unique Species Identified	38	38	37
	CA Native Species	25	36	37
	Avg. ASVs Per Species	3.8	4.1	4.2