Skimming for barcodes: rapid production of mitochondrial genome and nuclear ribosomal repeat reference markers through shallow shotgun sequencing

Mykle L. Hoban[1], Jonathan Whitney[2], Allen G. Collins[3], Christopher Meyer[4], Katherine R. Murphy[5], Abigail J. Reft[3], Katherine E. Bemis[3]

[1] Hawai'i Institute of Marine Biology, University of Hawai'i at Mānoa, Kāne'ohe, Hawai'i, USA

[2] Pacific Islands Fisheries Science Center, National Oceanic and Atmospheric Administration, Honolulu, HI, 96818, USA

[3] NOAA National Systematics Laboratory, National Museum of Natural History, Smithsonian Institution, Washington, D.C. 20560, USA

[4] Department of Invertebrate Zoology, National Museum of Natural History, Smithsonian Institution, Washington, D.C. 20560, USA

[5] Laboratories of Analytical Biology, National Museum of Natural History, Smithsonian Institution, Washington, D.C. 20560, USA

Corresponding author
Mykle Hoban[1]
46-007 Lilipuna Rd, Kāne'ohe, HI, USA
Email address: mh@myklehoban.com

Abstract

DNA barcoding is critical to conservation and biodiversity research, yet public reference databases are incomplete. Existing metazoan barcode databases are biased toward cytochrome oxidase subunit I (COI) and frequently lack associated voucher specimens or geospatial metadata, which can hinder reliable species assignments. The emergence of metabarcoding approaches such as environmental DNA (eDNA) has necessitated multiple marker techniques combined with barcode reference databases backed by voucher specimens. Reference barcodes have traditionally been generated by Sanger sequencing, however sequencing multiple markers is costly for large numbers of specimens, requires multiple PCR reactions, and limits resulting sequences to targeted regions. High-throughput sequencing techniques such as genome skimming enable assembly of mitochondrial genomes (mitogenomes), which contain the most

38 commonly used barcoding loci (e.g., COI, 12S, 16S), as well as nuclear ribosomal repeat regions
39 (e.g., ITS1&2, 18S). Here, we evaluated genome skimming as a method to sequence
40 mitogenomes and nuclear ribosomal repeats to build comprehensive barcode reference databases
41 for marine fishes. We selected a taxonomically diverse group of 12 marine fish species
42 vouchered in the National Museum of Natural History (NHNH), Smithsonian Institution. Using
43 DNA from the NMNH Biorepository, we generated two sequence libraries per extract to test the
44 effect of shearing method (enzymatic or mechanical) across a range of input DNA concentrations
45 and between the two DNA extraction methods most commonly used at NMNH (kit-based or
46 automated). Complete mitogenomes for all non-chondrichthyans (11/12 species) and nuclear
47 ribosomal repeats (18S-ITS1-5.8S-ITS2-28S) were assembled for all taxa. Despite differences in
48 raw read counts, the quality and completeness of mitogenome assemblies was not impacted by
49 shearing method and did not vary across extraction methods or input DNA concentrations. Our
50 results affirm that genome skimming is an efficient and (at scale) cost-effective method to
51 generate all mitochondrial and common nuclear DNA barcoding loci, which has great potential
52 to scale for future projects and facilitate completing barcode reference databases for marine
53 fishes.

Introduction

55 DNA barcoding has long been recognized as a critical component of biodiversity research
56 (Hebert et al., 2003; Ratnasingham & Hebert, 2013; DeSalle & Goldstein, 2019; Adamowicz et
57 al., 2019), but available barcode reference databases remain incomplete (Mugnai et al., 2021).
58 More comprehensive regional reference datasets in global databases better support research goals
59 and applications such as discovering new species (Carpenter, Williams & Santos, 2017; Hoban
60 & Williams, 2020), matching larval specimens to known adults (Johnson et al., 2009; Hubert et
61 al., 2010), and authenticating seafood labeling (Marko, Nance & Guynn, 2011; Silva & Hellberg,
62 2021). Traditionally, DNA barcoding efforts relied on Sanger sequencing of single mitochondrial
63 markers, particularly cytochrome oxidase subunit I (COI) for metazoans. However, there is
64 increasing utility for other mitochondrial genes and noncoding regions (e.g., 16S, 12S) as well as
65 nuclear ribosomal genes that are present in tandem repeats (e.g., 18S-ITS1-5.8S-ITS2-28S)
66 (Pochon et al., 2013; Berry et al., 2017; Alexander et al., 2020). To develop more complete DNA
67 barcode databases, we evaluated a method of genome skimming that has potential to
68 simultaneously recover multiple barcoding loci for many species.
69
70 DNA barcodes are essential resources, but the quality and utility of existing data is variable. For
71 DNA barcodes to be of long-term value, they should be linked to physical voucher specimens in
72 permanent natural history collections because voucher specimens allow for verification of
73 identification and refinements in taxonomy (Schander & Willassen, 2005; Ward, Hanner &
74 Hebert, 2009; but see Collins & Cruickshank, 2013). Another consideration for building barcode
75 libraries stems from natural genetic variation in populations. For example, Hawaiian populations
76 of widespread Indo-Pacific fishes are often genetically divergent and can comprise cryptic

77 lineages (DiBattista et al., 2010, 2012; Bowen et al., 2013). Thus, the most valuable barcode
78 sequences are derived from voucher specimens associated with precise geospatial metadata
79 (geotags), which are unfortunately missing for most archived genomic datasets (Toczydlowski et
80 al., 2021). Other attributes, such as color photographs of the specimen at the time of collection
81 and detailed collection metadata, add to barcode value. Finally, to increase discoverability and
82 data access, specimen and sequence metadata need to be linked through persistent digital
83 identifiers across systems of record (Riginos et al., 2020). These best practices in data
84 stewardship are necessary to support cross-domain cyberinfrastructure to enable transdisciplinary
85 research, discovery and reuse of material samples and their derived data (Davies et al., 2021).
86
87 Efforts to characterize community biodiversity patterns through metabarcoding (Leray &
88 Knowlton, 2015; Timmers et al., 2021) and environmental DNA (eDNA) surveys (Ficetola et al.,
89 2008)—that rely on well-curated barcode databases to accurately assign sequences to
90 taxonomy—have expanded dramatically (Ruppert, Kline & Rahman, 2019). In addition,
91 approaches (such as eDNA) that are based on potentially fragmentary source material and/or
92 those that target specific taxa are more precise with a multi-marker approach (Stat et al., 2017;
93 West et al., 2020; Casey et al., 2021). Finally, targeting short hypervariable loci (e.g., Riaz et al.,
94 2011; Miya et al., 2015) can be more compatible with read lengths produced by high-throughput
95 sequencing (HTS) platforms. The availability of multiple genetic markers associated with a
96 single voucher specimen also makes species identifications more consistent across studies where
97 researchers may use different loci.
98
99 As high-throughput sequencing has become more accessible and cost-effective, genome
100 skimming, which uses low-pass, shallow shotgun sequencing of whole genomes, has become
101 practical (Trevisan et al., 2019). Genome skimming does not enrich samples for specific target
102 loci, yet it is successful at recovering high-copy regions such as mitochondrial and plastid
103 genomes as well as nuclear or cytosolic sequences like ribosomal DNA (Kane et al., 2012;
104 Straub et al., 2012; Besnard et al., 2013; Malé et al., 2014; Ripma, Simpson & Hasenstab-
105 Lehman, 2014; Dodsworth, 2015; Denver et al., 2016; Grandjean et al., 2017; Liu et al., 2020;
106 Raupach et al., 2022). Genome skimming has great potential to fill DNA barcode reference
107 databases because it generates sequence data for commonly used barcoding markers
108 simultaneously (Coissac et al., 2016). This potential has been realized in a range of taxa from
109 plants (Alsos et al., 2020) to arthropods (Grandjean et al., 2017; Raupach et al., 2022). Our work
110 complements Therkildsen & Palumbi (2017), who used a similar approach to examine genetic
111 variation in Atlantic Silversides and Margaryan et al. (2021), who developed a mitogenome
112 barcode database for vertebrates of Denmark, and extends these studies by showing that
113 ribosomal barcoding loci are also readily accessible using genome skimming. Despite previous
114 applications of genome skimming, it has yet to be tested broadly to capture specimen-backed
115 DNA barcodes for marine fishes.
116

117
118 Natural history collections hold valuable materials to support regional and taxon-specific
119 barcode database development, allowing gaps to be filled without the need to collect new
120 specimens. While many institutions voucher tissue samples and/or DNA extractions alongside
121 collected specimens, sequences are frequently published for only a limited number of loci (e.g.,
122 COI for metazoans, ITS for fungi (Ratnasingham & Hebert, 2007)). In our study, which is part of
123 an ongoing effort to complete the barcode reference database for Hawaiian marine fishes, we
124 evaluated genome skimming as a method to rapidly and (when scaled up to massively parallel
125 sequencing platforms) inexpensively capture all commonly used DNA barcoding loci for
126 multiple samples and fish taxa simultaneously. Using genome skimming, we aimed to recover
127 the complete mitochondrial genomes and ribosomal repeat regions of 12 taxonomically diverse
128 species of marine fishes. For our test, we prepared and sequenced two libraries for each species
129 (24 libraries total) from vouchered specimens in the National Museum of Natural History
130 (NMNH) fish collection. We evaluated the quality of sequences and our ability to assemble
131 complete mitogenomes and ribosomal repeats in the context of taxonomic diversity and shearing
132 method, and across a range of DNA extraction methods and input DNA concentrations. Here we
133 report the results of our test and discuss how to adapt this method for large-scale generation of
134 specimen-backed DNA barcodes.

Materials & Methods

Sample selection

137 We selected samples from 12 species across a broad taxonomic distribution of fishes, including
138 one chondrichthyan and 11 teleosts (Fig. 1). This work is a component of an effort to generate
139 specimen-backed barcodes for all species of Hawaiian marine fishes (~1200 species; unpublished
140 updated version of Mundy, 2005; Randall, 2007); thus, most specimens were Hawaiian species
141 collected in Hawaiʻi (6/12) or species that occur in Hawaiʻi but that were collected elsewhere
142 (3/12). We also included two western North Atlantic species: *Brosme brosme* (Cusk), which is a
143 NOAA species of concern, and *Gymnura altavela* (Spiny Butterfly Ray), as a representative
144 chondrichthyan. All samples were taken from existing DNA extracts in the National Museum of
145 Natural History (NMNH) Biorepository, derived from specimens housed in the fish collection at
146 NMNH (Table 1). Archived Biorepository DNA was originally extracted from tissues
147 subsampled and preserved in the field at the time of specimen collection. Ten of the 12
148 specimens have live color photographs (Fig. 1). No mitogenomes or ribosomal repeats were
149 available in GenBank for any of the species selected except *Gymnura altavela*, which was
150 published during preparation of this manuscript (Kousteni et al., 2021). All selected Hawaiian
151 species lacked regionally localized specimen-backed barcodes for at least one common fish
152 barcoding locus (COI, 16S, 12S; Table S1).

DNA concentration and extractions

DNA extracts representing a range of concentrations (0.9–34.0 ng/μL) were retrieved from the NMNH Biorepository. We did not standardize concentrations prior to library preparation. To demonstrate that the two extraction methods commonly used at NMNH yield viable outcomes, we included four samples extracted with the Qiagen BioSprint DNA blood kit (Qiagen, Inc.; Venlo, Netherlands) and eight samples extracted by an AutoGenPrep 965 automated DNA extraction robot (AutoGen; Holliston, MA, USA) following the manufacturer's tissue protocols. These are standard DNA extraction technologies used for Sanger-based DNA barcoding, similar to those that have been used to generate the majority of available DNA extracts in existing collections.

Shearing method and library preparation

We prepared two libraries for each of the 12 fish species, one sheared enzymatically and the other sheared mechanically, for a total of 24 libraries. Input DNA for the mechanically sheared libraries was prepared using a Covaris ME220 sonicator (Covaris; Woburn, MA, USA), then libraries were constructed with the NEB Ultra II DNA library prep kit (New England Biolabs; Ipswich, MA, USA) according to the manufacturer's protocols (with the exception noted below). We prepared enzymatically sheared libraries using the NEB Ultra II FS DNA library prep kit (New England Biolabs), which incorporates enzymatic shearing as part of the kit workflow. We targeted an insert size of approximately 200 bp and amplified libraries using six cycles of PCR according to the kit manufacturer's chemistry and thermocycler settings. We used iTru y-yoke adapter stubs and iTru unique dual indices (Glenn et al., 2019) in place of NEB adapters and indices and tailored the amount of adapter based on DNA concentration following NEB guidelines. Individual libraries were quantified with a Qubit dsDNA HS assay (Thermo Fisher Scientific; Waltham, MA, USA) and run on a High Sensitivity D1000 ScreenTape (Agilent; Santa Clara, CA, USA) to assess library size in bp. Finally, libraries were pooled to equimolar amounts prior to sequencing.

During library preparation, our enzymatically sheared samples inadvertently sat at 4°C following the end of the ligation period for an additional 45 minutes compared to those mechanically sheared. This gave the enzymatically sheared samples more time to ligate and likely impacted their ligation efficiency and subsequent library yield.

Sequencing

Libraries were split into two pools, and each pool was sequenced in a single run on the Illumina MiSeq (Illumina Inc.; San Diego, CA, USA) using V3 chemistry at the Laboratories of Analytical Biology, NMNH. We limited the sequencing run length to 150bp (paired end) to test scalability to higher-throughput platforms such as the Illumina NovaSeq 6000.

189 **Assembly**

190 We assessed two approaches to mitogenome assembly using Geneious Prime 2021.2.2
191 (https://www.geneious.com). First, we used the Map to Reference function and built-in Geneious
192 mapper with the sensitivity set to "medium/low" and iterations set to "up to 10 times", starting
193 with published COI sequences (Table 1) for each of the 24 libraries. Resulting assemblies were
194 inspected and trimmed at the ends (up to 50 bp) where coverage was low (<5X). Consensus
195 sequences were generated from the assembly results and used as subsequent reference seeds and
196 the Map to Reference step repeated until the assemblies stopped increasing in size and identical
197 stretches of sequences were detected at the 5' and 3' ends. The second approach used a complete
198 mitogenome from either a congeneric or confamilial taxon as the reference sequence, and Map to
199 Reference, using the same parameters for a single set of up to 10 iterations. Assemblies of
200 ribosomal repeat regions were conducted similarly, with reiterations using the Map to Reference
201 function in Geneious, using ribosomal sequences from closely related taxa published in GenBank
202 (Table S2). In addition to assembling mitogenomes, we constructed nuclear genome
203 preassemblies using SPAdes 3.15.3 (assembly module only) on paired forward and reverse read
204 libraries (Prjibelski et al., 2020), and filtered out preassembly contigs shorter than 200 bp.

205 **Genome sequencing coverage estimation**

206 We estimated species genome sizes (Table 1) based on data available in GenBank or the Animal
207 Genome Size Database (Gregory, 2021). Where specific estimates were unavailable, we
208 calculated an average genome size of congeners or closely related confamilials. Since no
209 congener or confamilial genomes were available for *G. altavela*, we estimated genome size
210 based on the average genome size for Batoidea. We then estimated sequence coverage (*C*) for
211 each sample using the equation $C = LN/G$, where *L* was the sequencing read length, *N* was the
212 number of reads, and *G* was the estimated haploid genome length.

213 **Annotation**

214 We annotated assembled mitogenomes using the MitoAnnotator tool from the MitoFish
215 Mitochondrial Genome Database of Fish (Iwasaki et al., 2013). We manually annotated
216 ribosomal repeat regions by aligning to complete ribosomal repeat regions for fishes in GenBank
217 (Table 2). We did not annotate preassembly contigs.

218 **Phylogenetic analyses**

219 To assess relationships and validate taxonomic identities, we performed phylogenetic analyses
220 including all mitogenomes generated in this study and confamilial taxa with published
221 mitogenomes available in the MitoFish database (52 species; Table S5). Due to the large number
222 of species with available mitogenomes in the family Carangidae, we only used species in *Seriola,*
223 *Elegatis,* and *Decapterus,* the available genera most closely related to our taxon *Scomberoides*
224 *lysan* (Damerau, Freese & Hanel, 2018; Rabosky et al., 2018). We used sequences of all protein-

225 coding genes (PCGs) and two rRNAs. Each PCG or rRNA was individually aligned using
226 MAFFT v7.505 (Katoh & Standley, 2013) and then concatenated to a single final alignment. We
227 used PartitionFinder2 to assess the partitioning of models of molecular evolution (Guindon et al.,
228 2010; Lanfear et al., 2012, 2017). We partitioned the alignment by gene and, for PCGs, by codon
229 position, then ran PartitionFinder2 using the "greedy" algorithm with branch lengths specified as
230 unlinked to test the models supported in MrBayes. We conducted a Bayesian phylogenetic
231 reconstruction using MrBayes v3.2.7 (Ronquist et al., 2012), running four independent searches
232 of six chains for 22 million generations, saving trees every 1,000 generation and discarding the
233 first 15% as burn-in. We verified convergence of MCMC runs and model parameters using
234 TRACER v1.7.2 (Rambaut et al., 2018). We conducted a maximum likelihood (ML)
235 phylogenetic reconstruction with the partitioned alignment using RAxML v8.2.12 (Stamatakis,
236 2014) and specified 1,000 bootstrap replicates to assess node support. Resulting trees were
237 rooted to the two *Gymnura* species and plotted using ggtree (Yu et al., 2017) and phytools
238 (Revell, 2012) in R v4.1.2 (R Core Team, 2020).

Data availability

240 All voucher and material sample properties can be found in GEOME, the Genomic Observatories
241 Metadatabase (Riginos et al., 2020), under the expedition NMFS_FISHES_MiSeqPilot_01
242 (https://n2t.net/ark:/21547/EEV2). We deposited BioSample records, annotated mitogenome and
243 ribosomal repeat assemblies, and raw reads in GenBank (BioProject Accession: PRJNA720393).
244 Code and procedures used to perform phylogenetic analyses are available on GitHub
245 (https://github.com/hawaii-barcoding-initiative/mitogenome_tree).

Results

DNA Concentration

248 Total input DNA for library preparation ranged from 4.6 to 170 ng. Final libraries ranged from
249 0.16 to 3.34 ng/µL in concentration, with mechanically and enzymatically sheared libraries
250 averaging 0.71 ± 0.67 ng/µL (mean ± sd) and 1.72 ± 0.94 ng/µL, respectively. The average total
251 library size ranged from 318 to 392 bp, with mechanically- and enzymatically sheared libraries
252 averaging 345 ± 16 bp and 373 ± 18 bp, respectively. A summary of library quantification results
253 can be found in Table 3.

Sequence reads and genome coverage

255 We recovered 0.46 to 5.2 million reads (2.5 ± 1.1 million) per library. AutoGen and Qiagen
256 extractions performed comparably (2.6 ± 1.3 million reads for AutoGen vs. 2.0 ± 0.4 million for
257 Qiagen). Enzymatic shearing yielded more reads per library than mechanical shearing (2.9 ± 1.1
258 million reads for enzymatic vs. 1.8 ± 0.6 million reads for mechanical). Sequence duplication
259 rates varied from 0.7–6.5% per sample. Based on estimated genome sizes, these read counts

260 equate to 0.07× to 1.04× genome coverage, with enzymatic shearing (0.50 ± 0.30×) averaging
261 higher than mechanical shearing (0.30 ± 0.19×). A summary of sequencing results across
262 libraries is presented in Table 3.

263 **Assembly and sequence coverage**

264 We readily assembled and annotated complete mitochondrial genomes for the 11 teleosts (see
265 Table 2 for assembled mitogenome accession numbers). Assembled sequences were identical
266 whether we started from a small seed (COI) or mapped to a complete mitochondrial reference
267 genome derived from a congeneric or confamilial taxon. We did not recover a complete
268 mitogenome from *Gymnura altavela* (Spiny Butterfly Ray), but assembled large sections of it
269 (e.g., ~12,000 bp including COI; ~3,000 bp including 16S). During the course of this work, a
270 complete mitochondrial genome was published for *G. altavela* (MT274571) based on a specimen
271 from Greece (Kousteni et al., 2021). This allowed us to improve our assembly, resulting in a
272 mitochondrial genome with a short gap in COI and a second gap in the D-loop. Fortunately, the
273 gap spanned the published COI sequence for our specimen (USNM 433343; MH378654),
274 allowing us to use 24 bases from that sequence to fill the missing space. As a result, we
275 ultimately derived a nearly-complete mitochondrial genome for the Spiny Butterfly Ray (19,022
276 bp in our assembly as compared to 19,472 bp in MT274571).
277
278 Mitogenome coverage of the 22 successful assemblies ranged from 7× to 108× (34 ± 26×; Table
279 3). The *Gymnura altavela* libraries had a comparable number of reads to other species in our
280 study, but coverage of the mitogenome was low for unknown reasons (11.2× with both libraries
281 combined). Across all libraries, assembled mitogenome reads comprised 0.05% to 0.32% (0.17 ±
282 0.1%) of the total raw reads generated per specimen.
283
284 Using Geneious Map to Reference, we assembled and annotated ribosomal repeat regions (18S-
285 ITS1-5.8S-ITS2-28S) for all 12 taxa by using 18S or 28S reference seeds (see Table 2 for
286 assembled ribosomal repeat accession numbers).
287
288 Genome preassemblies generated by SPAdes (>200 bp) were uploaded to Zenodo (along with
289 basic assembly statistics) and assigned persistent identifiers (Table 2). As expected, the
290 preassemblies were limited, with a small fraction of contigs exceeding 1 kb in length.
291 Nevertheless, preassembly contigs that correspond to the complete or nearly complete
292 mitochondrial genomes and the ribosomal repeat regions were recovered for 7 and 8,
293 respectively, of the 12 species in our study.

294 **Mitogenome organization and structure**

295 Mitogenomes for all species were arranged similarly, with some minor length variations,
296 particularly in the control region (see Fig. 2 for example assembly of *Canthigaster amboinensis*;

297   see Fig. S1 for all mitogenome assemblies). We detected no mitochondrial gene rearrangements
298   among the 12 species we investigated. All species had 36 genes comprising 13 protein-coding
299   genes (PCGs) and 23 tRNAs, with two rRNAs and the control region. In all cases, the majority
300   strand encoded 12 PCGs, 15 tRNAs, both rRNAs, and the control region. The remaining eight
301   tRNAs and a single PCG were encoded on the minority strand. GC content ranged from 43.1%
302   (*Neoniphon sammara*) to 52.1% (*Gymnothorax fimbriatus*) (mean: 45.5 ± 2.3%).

303   **Phylogenetic analyses**

304   Both ML and Bayesian methods produced identical topologies (Fig. 3), with the single exception
305   of different branching order within the genus *Ostracion* (see Figs. S2 and S3 for raw ML and
306   Bayesian trees respectively, including complete node support values). All specimens sequenced
307   for this study were recovered within their respective taxonomic groups, and branching order
308   among families matched that of the family-level backbone tree published in Rabosky et al.
309   (2018). Node support in the ML tree (bootstrap value) was more variable than in the Bayesian
310   tree (posterior probability). The ML tree had many strongly supported nodes (70–100%
311   bootstrap support), but two with weak support (<20%). The Bayesian tree was strongly
312   supported throughout, with most nodes having >95% posterior probability.

313   **Discussion**

314   Our results show that genome skimming by shallow shotgun sequencing is an efficient method
315   for generating mitogenomes and ribosomal repeats of marine fishes. The methods are robust for
316   a broad range of taxa, extraction types, shearing methods, and DNA concentrations. Both kit-
317   based (Qiagen) and automated (AutoGen) extractions resulted in high quality sequence libraries,
318   which indicates that this method can leverage existing DNA extractions housed in museum
319   collections that were prepared for other purposes (e.g., single-marker Sanger sequencing).
320
321   As noted in Methods, our enzymatically sheared samples were held at 4°C following the end of
322   the ligation period for an additional 45 minutes compared to those mechanically sheared. This
323   likely impacted their ligation efficiency and subsequent library yield. As a result, we cannot
324   confirm that differences in final library yield nor differences in read counts resulted directly from
325   the shearing method used. Although libraries were pooled in equimolar ratios, these values were
326   calculated using total dsDNA (ng/µL) as measured by Qubit and TapeStation fragment size (bp).
327   A more accurate method would be to quantify individual libraries with qPCR, thus measuring
328   DNA that can be sequenced, rather than total DNA. Regardless of these differences, we
329   demonstrated that enzymatic shearing can be an effective method for genome skimming;
330   enzymatic shearing is also less expensive (~$4 less/library; Tables S3 and S4), less labor
331   intensive, and requires less specialized laboratory equipment.
332

333   We assembled mitogenomes with as few as half a million reads but had more consistent success
334   with 2–3 million reads/library, which resulted in an average of 34× coverage of the mitogenome.
335   Mitogenome assemblies used only 0.05% to 0.32% of the total raw sequence reads. The majority
336   of unassembled reads were nuclear (e.g., chromosomal) and cytosolic (e.g., ribosomal RNA)
337   sequences. The most common barcoding markers for fishes are mitochondrial: COI (Leray et al.,
338   2013), 16S rRNA (Berry et al., 2017), and 12S rRNA (Miya et al., 2015). However, primer sets
339   designed to amplify other taxa or communities often target nuclear ribosomal loci such as the
340   18S rRNA and/or internal transcribed spacers (ITS1/2) (marine eukaryotes: Pochon et al., 2013;
341   scleractinian corals: Alexander et al., 2020). We successfully recovered complete ribosomal
342   repeat regions (18S-ITS1-5.8S-ITS2-28S) from all of our sequence libraries, illustrating that our
343   approach has applications beyond mitogenome assembly. Importantly, we recovered sequences
344   for the most commonly used barcoding loci for all targeted taxa in a single pass. We provided
345   raw sequence data in the NCBI Sequence Read Archive under BioProject PRJNA720393
346   because there are likely additional sequences of interest to other researchers. In addition, we
347   constructed genome preassemblies for each sample, which are also available (Table 2).
348

349   Our phylogenetic analyses of concatenated protein-coding genes and rRNAs recovered a
350   topology consistent across tree-building methods (Fig. 3) and that comports with recent higher-
351   level fish phylogenies (e.g., Rabosky et al., 2018). Notably, in the combined tree the two nodes
352   most weakly supported by ML both had 100% Bayesian posterior probabilities. These
353   discrepancies may be due to how the two approaches partition molecular evolution models.
354   RAxML supports partitioning but only allows a single model across the alignment, whereas
355   MrBayes allows models to vary across partitions (e.g., genes & codon positions). Overall, this
356   phylogenetic analysis helps validate both the species identity of each voucher specimen and the
357   quality of genome-skimming derived mitogenome assemblies.
358

359   To test whether our methods are applicable across fish diversity, we included one
360   chondrichthyan, the Spiny Butterfly Ray *Gymnura altavela*. Despite high success across teleosts,
361   we did not recover a complete mitogenome for the chondrichthyan. The *G. altavela* libraries had
362   read counts comparable to bony fish libraries, but mitogenome coverage was low and initial
363   assemblies had gaps. Two potential causes of low coverage in *Gymnura* include exogenous (non-
364   target) DNA and a greater number of shorter-than-expected DNA fragments in the sequencing
365   libraries. During preparation of this manuscript, a complete mitochondrial genome was published
366   from a specimen from Greece (Kousteni et al., 2021), and although it is ~3% diverged from our
367   mitochondrial sequences, we used it to improve our assembly such that it included complete loci
368   other than the D-loop. Gaps in the control region are relatively common in mitochondrial
369   genome assemblies, particularly among rays (Poortvliet et al., 2015; Hinojosa-Alvarez et al.,
370   2015). This region often contains tandem repeats that present difficulty to bioinformatic
371   assemblers (White et al., 2018) and have been attributed to heteroplasmy in other taxa (Mundy,
372   Winchell & Woodruff, 1996). However, despite the D-loop gap in the complete mitogenome

373 assembly of *G. altavela*, we still recovered targeted mitochondrial barcoding loci (COI, 12S,
374 16S). In future studies, we will sequence additional sharks, rays, and chimaeras to further explore
375 laboratory and bioinformatic approaches for generating chondrichthyan mitogenomes.
376
377 We used the MiSeq platform to test shearing methods, compare extraction types and DNA
378 concentrations, and to assess sequencing reads and coverage necessary to generate mitogenomes
379 and ribosomal repeats across a broad taxonomic selection of fishes. To further our goal of
380 completing barcode reference databases (for mitochondrial and ribosomal genes) for all species
381 of Hawaiian fishes, we will sequence future genome skimming runs on an Illumina NovaSeq.
382 The NovaSeq platform produces higher read output than MiSeq and therefore supports increased
383 multiplexing of samples, allowing us to pool 384 samples (species) in a single sequencing run.
384 This will reduce sequencing costs from ~$145 per sample on the MiSeq to ~$16 on the NovaSeq,
385 while also increasing the average yield from 2.5 million reads to 13 million per sample. The
386 increased multiplexing capability of the NovaSeq brings the total cost (library preparation,
387 quantitation, and sequencing) from ~$161 per sample on the MiSeq to ~$31 per sample, which
388 will facilitate economical and rapid generation of complete mitogenomes and ribosomal repeats
389 (encompassing all major barcoding loci) (see Tables S3 and S4). Preliminary data (not reported
390 here) from a NovaSeq run of 384 species show that our methods for mitogenome and ribosomal
391 repeat recovery via genome skimming can be scaled to the higher-throughput platform. In this
392 study, we employed manual assembly methods using Geneious Prime, whereas future assemblies
393 will employ an automated bioinformatic pipeline to enable production of multilocus DNA
394 barcode sequences at scale.
395
396 We enhanced the reference value of our derived genetic data through persistent digital
397 identifiers. Raw reads and assembled sequences are linked through NCBI accessions (BioProject,
398 BioSample, SRA, and nucleotide) to museum voucher specimens, as well as to derived tissues
399 and DNA extracts at NMNH. Further, to ensure that data derived from and associated with these
400 biomaterials can easily be accessed and reused, we cross linked NCBI and GEOME records
401 through Archival Resource Key (ARK) identifiers (Kunze, 2021). Such best practices in data
402 stewardship and the use of persistent identifiers across systems of record will facilitate cross-
403 domain cyberinfrastructure and enable transdisciplinary research, discovery, and reuse of
404 material samples and their derived data (Davies et al., 2021).

405 **Conclusions**

406 Our study shows that genome skimming is an efficient and cost effective method that will allow
407 a shift in the DNA barcoding workflow from sequencing targeted loci in individual specimens to
408 generating complete suites of barcode markers for many taxa in a single sequencing run. The
409 methods we employed enable use of genetic samples housed in natural history collections to
410 rapidly generate specimen-based, regionally localized DNA barcode reference data. This work
411 has important implications for several large U.S.-based initiatives: NOAA 'omics (Goodwin et

412   al., 2021), NMNH Ocean DNA Initiative (https://www.smithsonianmag.com/blogs/national-
413   museum-of-natural-history/2021/07/07/meet-reef-expert-collecting-environmental-time-
414   capsules/), and the U.S. Ocean Biocode (Meyer et al., 2021), each of which involves explicit
415   aims to provide complete DNA barcode reference databases based on voucher specimens.
416   Techniques and methods developed here are applicable to taxa and regions beyond marine fishes
417   and the Hawaiian Islands. Comprehensive voucher-based reference databases are necessary to
418   advance sequence-based detection, censusing, and monitoring of marine communities in the face
419   of global change.

Acknowledgements

421   Samples from French Polynesia were acquired under a collaborative Centre de Recherche
422   Insulaire et Observatoire de l'Environnement (CRIOBE) and Smithsonian Institution National
423   Museum of Natural History (NMNH) initiative to survey the marine fishes of French Polynesia,
424   including the Moʻorea Biocode Project. We thank NMNH and CRIOBE, in particular Jeffrey
425   Williams (NMNH) and Serge Planes (CRIOBE), as the collectors and photographers of the
426   samples. Specimens collected from Hawaiʻi were acquired under the MarineGEO Hawaiʻi 2017
427   project to survey the fishes of Kāneʻohe Bay. We thank the Smithsonian Conservation Biology
428   Institute, NMNH, and the Hawaiʻi Institute of Marine Biology, in particular Mary Hagedorn,
429   Lynne R. Parenti, Diane Pitassy, Zeehan Jaafar, Kassi S. Cole, and Kiril Vinnikov, as the
430   collectors of the samples. Photographs of Hawaiian fishes were provided by Diane Pitassy.
431   Genetic benchwork and sequencing was completed at the Smithsonian NMNH Laboratories of
432   Analytical Biology (LAB). At NMNH we thank Carole Baldwin, Daniel DiMichele, Chris
433   Huddleston, Lynne R. Parenti, Diane Pitassy, Niamh Redmond, Makiri Sei, Lee Weigt, Jeff
434   Williams, and Herman Wirshing for their support. Funding support from NOAA's Office of
435   Science and Technology, NOAA's Pacific Islands Fisheries Science Center, The Cooperative
436   Institute for Marine and Atmospheric Research, NOAA's West Hawaiʻi Integrated Ecosystem
437   Assessment Program (contribution no. X), and NMNH Laboratories of Analytical Biology. This
438   is contribution #XXX from the Hawaiʻi Institute of Marine Biology and #XXX from the School
439   of Ocean and Earth Science and Technology at the University of Hawaiʻi.

References

441   Adamowicz SJ, Boatwright JS, Chain F, Fisher BL, Hogg ID, Leese F, Lijtmaer DA, Mwale M,

442   Naaum AM, Pochon X, Steinke D, Wilson J-J, Wood S, Xu J, Xu S, Zhou X, van der

443   Bank M. 2019. Trends in DNA barcoding and metabarcoding. *Genome* 62:v–viii. DOI:

444   10.1139/gen-2019-0054.

445 Alexander JB, Bunce M, White N, Wilkinson SP, Adam AAS, Berry T, Stat M, Thomas L,

446   Newman SJ, Dugal L, Richards ZT. 2020. Development of a multi-assay approach for

447   monitoring coral diversity using eDNA metabarcoding. *Coral Reefs* 39:159–171. DOI:

448   10.1007/s00338-019-01875-9.

449 Alsos IG, Lavergne S, Merkel MKF, Boleda M, Lammers Y, Alberti A, Pouchon C, Denoeud F,

450   Pitelkova I, Puşcaş M, Roquet C, Hurdu B-I, Thuiller W, Zimmermann NE,

451   Hollingsworth PM, Coissac E. 2020. The treasure vault can be opened: Large-scale

452   genome skimming works well using herbarium and silica gel dried material. *Plants*

453   9:432. DOI: 10.3390/plants9040432.

454 Berry TE, Osterrieder SK, Murray DC, Coghlan ML, Richardson AJ, Grealy AK, Stat M, Bejder

455   L, Bunce M. 2017. DNA metabarcoding for diet analysis and biodiversity: A case study

456   using the endangered Australian sea lion (*Neophoca cinerea*). *Ecology and Evolution*

457   7:5435–5453. DOI: 10.1002/ece3.3123.

458 Besnard G, Christin P-A, Malé P-JG, Coissac E, Ralimanana H, Vorontsova MS. 2013.

459   Phylogenomics and taxonomy of Lecomtelleae (Poaceae), an isolated panicoid lineage

460   from Madagascar. *Annals of Botany* 112:1057–1066. DOI: 10.1093/aob/mct174.

461 Bowen BW, Rocha LA, Toonen RJ, Karl SA, ToBo Laboratory. 2013. The origins of tropical

462   marine biodiversity. *Trends in Ecology & Evolution* 28:359–366. DOI:

463   10.1016/j.tree.2013.01.018.

464 Carpenter KE, Williams JT, Santos MD. 2017. *Acanthurus albimento*, a new species of

465   surgeonfish (Acanthuriformes: Acanthuridae) from northeastern Luzon, Philippines, with

466   comments on zoogeography. *Journal of the Ocean Science Foundation* 25. DOI:

467   10.5281/zenodo.291792.

468 Casey JM, Ransome E, Collins AG, Mahardini A, Kurniasih EM, Sembiring A, Schiettekatte
469   NMD, Cahyani NKD, Wahyu Anggoro A, Moore M, Uehling A, Belcaid M, Barber PH,
470   Geller JB, Meyer CP. 2021. DNA metabarcoding marker choice skews perception of
471   marine eukaryotic biodiversity. *Environmental DNA* 3:1229–1246. DOI:
472   10.1002/edn3.245.

473 Coissac E, Hollingsworth PM, Lavergne S, Taberlet P. 2016. From barcodes to genomes:
474   Extending the concept of DNA barcoding. *Molecular Ecology* 25:1423–1428. DOI:
475   10.1111/mec.13549.

476 Collins RA, Cruickshank RH. 2013. The seven deadly sins of DNA barcoding. *Molecular*
477   *Ecology Resources* 13:969–975. DOI: 10.1111/1755-0998.12046.

478 Damerau M, Freese M, Hanel R. 2018. Multi-gene phylogeny of jacks and pompanos
479   (Carangidae), including placement of monotypic vadigo *Campogramma glaycos*. *Journal*
480   *of Fish Biology* 92:190–202. DOI: 10.1111/jfb.13509.

481 Davies N, Deck J, Kansa EC, Kansa SW, Kunze J, Meyer C, Orrell T, Ramdeen S, Snyder R,
482   Vieglais D, Walls RL, Lehnert K. 2021. Internet of samples (iSamples): Toward an
483   interdisciplinary cyberinfrastructure for material samples. *GigaScience* 10:giab028. DOI:
484   10.1093/gigascience/giab028.

485 Denver DR, Brown AMV, Howe DK, Peetz AB, Zasada IA. 2016. Genome skimming: A rapid
486   approach to gaining diverse biological insights into multicellular pathogens. *PLOS*
487   *Pathogens* 12:e1005713. DOI: 10.1371/journal.ppat.1005713.

488 DeSalle R, Goldstein P. 2019. Review and interpretation of trends in DNA barcoding. *Frontiers*
489   *in Ecology and Evolution* 7.

490 DiBattista JD, Rocha LA, Craig MT, Feldheim KA, Bowen BW. 2012. Phylogeography of two

491   closely related Indo-Pacific butterflyfishes reveals divergent evolutionary histories and

492   discordant results from mtDNA and microsatellites. *Journal of Heredity* 103:617–629.

493   DOI: 10.1093/jhered/ess056.

494 DiBattista JD, Wilcox C, Craig MT, Rocha LA, Bowen BW. 2010. Phylogeography of the

495   pacific blueline surgeonfish, *Acanthurus nigroris*, reveals high genetic connectivity and a

496   cryptic endemic species in the hawaiian archipelago. *Journal of Marine Biology* 2011:1–

497   17. DOI: 10.1155/2011/839134.

498 Dodsworth S. 2015. Genome skimming for next-generation biodiversity analysis. *Trends in*

499   *Plant Science* 20:525–527. DOI: 10.1016/j.tplants.2015.06.012.

500 Ficetola GF, Miaud C, Pompanon F, Taberlet P. 2008. Species detection using environmental

501   DNA from water samples. *Biology Letters* 4:423–425. DOI: 10.1098/rsbl.2008.0118.

502 Glenn TC, Nilsen RA, Kieran TJ, Sanders JG, Bayona-Vásquez NJ, Finger JW, Pierson TW,

503   Bentley KE, Hoffberg SL, Louha S, Leon FJG-D, Portilla MA del R, Reed KD, Anderson

504   JL, Meece JK, Aggrey SE, Rekaya R, Alabady M, Belanger M, Winker K, Faircloth BC.

505   2019. Adapterama I: Universal stubs and primers for 384 unique dual-indexed or 147,456

506   combinatorially-indexed Illumina libraries (iTru & iNext). *PeerJ* 7:e7755. DOI:

507   10.7717/peerj.7755.

508 Goodwin K, Egan K, Greig T, Philibotte J, Koss J, Larsen K, Layton D, Nichols K, O'Neil J,

509   Parks D, Trtanj J, Werner C. 2021. *NOAA 'Omics Strategic Application of*

510   *Transformational Tools Strategic Plan 2021-2025*. Silver Spring, MD: National Oceanic

511   and Atmospheric Administration.

512 Grandjean F, Tan MH, Gan HM, Lee YP, Kawai T, Distefano RJ, Blaha M, Roles AJ, Austin

513   CM. 2017. Rapid recovery of nuclear and mitochondrial genes by genome skimming

514   from Northern Hemisphere freshwater crayfish. *Zoologica Scripta* 46:718–728. DOI:

515   10.1111/zsc.12247.

516 Gregory TR. 2021.Animal Genome Size Database. *Available at http://www.genomesize.com/*

517   (accessed October 14, 2021).

518 Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New

519   algorithms and methods to estimate maximum-likelihood phylogenies: assessing the

520   performance of PhyML 3.0. *Systematic Biology* 59:307–321. DOI:

521   10.1093/sysbio/syq010.

522 Hebert PDN, Cywinska A, Ball SL, deWaard JR. 2003. Biological identifications through DNA

523   barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*

524   270:313–321. DOI: 10.1098/rspb.2002.2218.

525 Hinojosa-Alvarez S, Díaz-Jaimes P, Marcet-Houben M, Gabaldón T. 2015. The complete

526   mitochondrial genome of the Giant Manta ray, *Manta birostris*. *Mitochondrial DNA*

527   26:787–788. DOI: 10.3109/19401736.2013.855753.

528 Hoban ML, Williams JT. 2020. *Cirripectes matatakaro*, a new species of combtooth blenny from

529   the Central Pacific, illuminates the origins of the Hawaiian fish fauna. *PeerJ* 8:e8852.

530   DOI: 10.7717/peerj.8852.

531 Hubert N, Delrieu-Trottin E, Irisson J-O, Meyer C, Planes S. 2010. Identifying coral reef fish

532   larvae through DNA barcoding: A test case with the families Acanthuridae and

533   Holocentridae. *Molecular Phylogenetics and Evolution* 55:1195–1203. DOI:

534   10.1016/j.ympev.2010.02.023.

535   Iwasaki W, Fukunaga T, Isagozawa R, Yamada K, Maeda Y, Satoh TP, Sado T, Mabuchi K,

536   Takeshima H, Miya M, Nishida M. 2013. MitoFish and MitoAnnotator: A mitochondrial

537   genome database of fish with an accurate and automatic annotation pipeline. *Molecular*

538   *Biology and Evolution* 30:2531–2540. DOI: 10.1093/molbev/mst141.

539   Johnson GD, Paxton JR, Sutton TT, Satoh TP, Sado T, Nishida M, Miya M. 2009. Deep-sea

540   mystery solved: Astonishing larval transformations and extreme sexual dimorphism unite

541   three fish families. *Biology Letters* 5:235–239. DOI: 10.1098/rsbl.2008.0722.

542   Kane N, Sveinsson S, Dempewolf H, Yang JY, Zhang D, Engels JMM, Cronk Q. 2012. Ultra-

543   barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and

544   nuclear ribosomal DNA. *American Journal of Botany* 99:320–329. DOI:

545   10.3732/ajb.1100570.

546   Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:

547   Improvements in performance and usability. *Molecular Biology and Evolution* 30:772–

548   780. DOI: 10.1093/molbev/mst010.

549   Kousteni V, Mazzoleni S, Vasileiadou K, Rovatsos M. 2021. Complete mitochondrial DNA

550   genome of nine species of sharks and rays and their phylogenetic placement among

551   modern elasmobranchs. *Genes* 12:324. DOI: 10.3390/genes12030324.

552   Kunze J. 2021.ARK Alliance. *Available at https://arks.org* (accessed December 14, 2021).

553   Lanfear R, Calcott B, Ho SYW, Guindon S. 2012. PartitionFinder: Combined selection of

554   partitioning schemes and substitution models for phylogenetic analyses. *Molecular*

555   *Biology and Evolution* 29:1695–1701. DOI: 10.1093/molbev/mss020.

556   Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2017. PartitionFinder 2: New

557   methods for selecting partitioned models of evolution for molecular and morphological

558   phylogenetic analyses. *Molecular Biology and Evolution* 34:772–773. DOI:

559   10.1093/molbev/msw260.

560   Leray M, Knowlton N. 2015. DNA barcoding and metabarcoding of standardized samples reveal

561   patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences*

562   112:2076–2081. DOI: 10.1073/pnas.1424997112.

563   Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, Boehm JT, Machida RJ. 2013.

564   A new versatile primer set targeting a short fragment of the mitochondrial COI region for

565   metabarcoding metazoan diversity: Application for characterizing coral reef fish gut

566   contents. *Frontiers in Zoology* 10:34. DOI: 10.1186/1742-9994-10-34.

567   Liu B-B, Campbell CS, Hong D-Y, Wen J. 2020. Phylogenetic relationships and chloroplast

568   capture in the *Amelanchier-Malacomeles-Peraphyllum* clade (Maleae, Rosaceae):

569   Evidence from chloroplast genome and nuclear ribosomal DNA data using genome

570   skimming. *Molecular Phylogenetics and Evolution* 147:106784. DOI:

571   10.1016/j.ympev.2020.106784.

572   Malé P-JG, Bardon L, Besnard G, Coissac E, Delsuc F, Engel J, Lhuillier E, Scotti-Saintagne C,

573   Tinaut A, Chave J. 2014. Genome skimming by shotgun sequencing helps resolve the

574   phylogeny of a pantropical tree family. *Molecular Ecology Resources* 14:966–975. DOI:

575   10.1111/1755-0998.12246.

576   Margaryan A, Noer CL, Richter SR, Restrup ME, Bülow-Hansen JL, Leerhøi F, Langkjær EMR,

577   Gopalakrishnan S, Carøe C, Gilbert MTP, Bohmann K. 2021. Mitochondrial genomes of

578   Danish vertebrate species generated for the national DNA reference database, DNAmark.

579   *Environmental DNA* 3:472–480. DOI: 10.1002/edn3.138.

580   Marko PB, Nance HA, Guynn KD. 2011. Genetic detection of mislabeled fish from a certified

581   sustainable fishery. *Current Biology* 21:R621–R622. DOI: 10.1016/j.cub.2011.07.006.

582   Meyer C, Duffy E, Collins A, Paulay G, Wetzer R. 2021. The U.S. ocean biocode. *Marine*

583   *Technology Society Journal* 55:140–141. DOI: 10.4031/MTSJ.55.3.33.

584   Miya M, Sato Y, Fukunaga T, Sado T, Poulsen JY, Sato K, Minamoto T, Yamamoto S,

585   Yamanaka H, Araki H, Kondoh M, Iwasaki W. 2015. MiFish, a set of universal PCR

586   primers for metabarcoding environmental DNA from fishes: detection of more than 230

587   subtropical marine species. *Royal Society Open Science* 2:150088. DOI:

588   10.1098/rsos.150088.

589   Mugnai F, Meglécz E, Abbiati M, Bavestrello G, Bertasi F, Bo M, Capa M, Chenuil A,

590   Colangelo MA, De Clerck O, Gutiérrez JM, Lattanzi L, Leduc M, Martin D, Matterson

591   KO, Mikac B, Plaisance L, Ponti M, Riesgo A, Rossi V, Turicchia E, Waeschenbach A,

592   Wangensteen OS, Costantini F. 2021. Are well-studied marine biodiversity hotspots still

593   blackspots for animal barcoding? *Global Ecology and Conservation* 32:e01909. DOI:

594   10.1016/j.gecco.2021.e01909.

595   Mundy BC. 2005. Checklist of the fishes of the Hawaiian archipelago. *Bishop Museum Bulletins*

596   *in Zoology* 6:1–706.

597   Mundy NI, Winchell CS, Woodruff DS. 1996. Tandem repeats and heteroplasmy in the

598   mitochondrial DNA control region of the Loggerhead Shrike (*Lanius ludovicianus*).

599   *Journal of Heredity* 87:21–26. DOI: 10.1093/oxfordjournals.jhered.a022948.

600   Pochon X, Bott NJ, Smith KF, Wood SA. 2013. Evaluating detection limits of next-generation

601   sequencing for the surveillance and monitoring of international marine pests. *PLOS ONE*

602   8:e73935. DOI: 10.1371/journal.pone.0073935.

603 Poortvliet M, Olsen JL, Croll DA, Bernardi G, Newton K, Kollias S, O'Sullivan J, Fernando D,

604 Stevens G, Galván Magaña F, Seret B, Wintner S, Hoarau G. 2015. A dated molecular

605 phylogeny of manta and devil rays (Mobulidae) based on mitogenome and nuclear

606 sequences. *Molecular Phylogenetics and Evolution* 83:72–85. DOI:

607 10.1016/j.ympev.2014.10.012.

608 Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020. Using SPAdes de

609 novo assembler. *Current Protocols in Bioinformatics* 70:e102. DOI: 10.1002/cpbi.102.

610 R Core Team. 2020. *R: A language and environment for statistical computing*. Vienna, Austria.

611 Rabosky DL, Chang J, Title PO, Cowman PF, Sallan L, Friedman M, Kaschner K, Garilao C,

612 Near TJ, Coll M, Alfaro ME. 2018. An inverse latitudinal gradient in speciation rate for

613 marine fishes. *Nature* 559:392–395. DOI: 10.1038/s41586-018-0273-1.

614 Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior summarization in

615 Bayesian phylogenetics using Tracer 1.7. *Systematic Biology* 67:901–904. DOI:

616 10.1093/sysbio/syy032.

617 Randall JE. 2007. *Reef and shore fishes of the Hawaiian Islands*. University of Hawai'i Press.

618 Ratnasingham S, Hebert PDN. 2007. BOLD: The Barcode of Life Data System

619 (http://www.barcodinglife.org). *Molecular Ecology Notes* 7:355–364. DOI:

620 10.1111/j.1471-8286.2007.01678.x.

621 Ratnasingham S, Hebert PDN. 2013. A DNA-based registry for all animal species: The barcode

622 index number (BIN) system. *PLOS ONE* 8:e66213. DOI: 10.1371/journal.pone.0066213.

623 Raupach MJ, Deister F, Villastrigo A, Balke M. 2022. The complete mitochondrial genomes of

624 *Notiophilus quadripunctatus* Dejean, 1826 and *Omophron limbatum* (Fabricius, 1777):

625   New insights into the mitogenome phylogeny of the Carabidae (Insecta, Coleoptera).

626   *Insect Systematics & Evolution* 1:1–22. DOI: 10.1163/1876312X-bja10027.

627   Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other

628   things). *Methods in Ecology and Evolution* 3:217–223. DOI: 10.1111/j.2041-

629   210X.2011.00169.x.

630   Riaz T, Shehzad W, Viari A, Pompanon F, Taberlet P, Coissac E. 2011. ecoPrimers: inference of

631   new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids*

632   *Research* 39:e145. DOI: 10.1093/nar/gkr732.

633   Riginos C, Crandall ED, Liggins L, Gaither MR, Ewing RB, Meyer C, Andrews KR, Euclide PT,

634   Titus BM, Therkildsen NO, Salces-Castellano A, Stewart LC, Toonen RJ, Deck J. 2020.

635   Building a global genomics observatory: Using GEOME (the Genomic Observatories

636   Metadatabase) to expedite and improve deposition and retrieval of genetic data and

637   metadata for biodiversity research. *Molecular Ecology Resources* 20:1458–1469. DOI:

638   10.1111/1755-0998.13269.

639   Ripma LA, Simpson MG, Hasenstab-Lehman K. 2014. Geneious! Simplified genome skimming

640   methods for phylogenetic systematic studies: A case study in Oreocarya (Boraginaceae).

641   *Applications in Plant Sciences* 2:1400062. DOI: 10.3732/apps.1400062.

642   Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L,

643   Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic

644   inference and model choice across a large model space. *Systematic Biology* 61:539–542.

645   DOI: 10.1093/sysbio/sys029.

646   Ruppert KM, Kline RJ, Rahman MS. 2019. Past, present, and future perspectives of

647   environmental DNA (eDNA) metabarcoding: A systematic review in methods,

648   monitoring, and applications of global eDNA. *Global Ecology and Conservation*

649   17:e00547. DOI: 10.1016/j.gecco.2019.e00547.

650   Schander C, Willassen E. 2005. What can biological barcoding do for marine biology? *Marine*

651   *Biology Research* 1:79–83. DOI: 10.1080/17451000510018962.

652   Silva AJ, Hellberg RS. 2021. Chapter Six - DNA-based techniques for seafood species

653   authentication. In: Toldrá F ed. *Advances in Food and Nutrition Research*. Academic

654   Press, 207–255. DOI: 10.1016/bs.afnr.2020.09.001.

655   Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of

656   large phylogenies. *Bioinformatics* 30:1312–1313. DOI: 10.1093/bioinformatics/btu033.

657   Stat M, Huggett MJ, Bernasconi R, DiBattista JD, Berry TE, Newman SJ, Harvey ES, Bunce M.

658   2017. Ecosystem biomonitoring with eDNA: Metabarcoding across the tree of life in a

659   tropical marine environment. *Nature Publishing Group* 7:12240. DOI: 10.1038/s41598-

660   017-12501-5.

661   Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A. 2012. Navigating the tip

662   of the genomic iceberg: Next-generation sequencing for plant systematics. *American*

663   *Journal of Botany* 99:349–364. DOI: 10.3732/ajb.1100335.

664   Therkildsen NO, Palumbi SR. 2017. Practical low-coverage genomewide sequencing of

665   hundreds of individually barcoded samples for population and evolutionary genomics in

666   nonmodel species. *Molecular Ecology Resources* 17:194–208. DOI: 10.1111/1755-

667   0998.12593.

668   Timmers MA, Jury CP, Vicente J, Bahr KD, Webb MK, Toonen RJ. 2021. Biodiversity of coral

669   reef cryptobiota shuffles but does not decline under the combined stressors of ocean

670   warming and acidification. *Proceedings of the National Academy of Sciences*

671   118:e2103275118. DOI: 10.1073/pnas.2103275118.

672   Toczydlowski RH, Liggins L, Gaither MR, Anderson TJ, Barton RL, Berg JT, Beskid SG, Davis

673   B, Delgado A, Farrell E, Ghoojaei M, Himmelsbach N, Holmes AE, Queeno SR, Trinh T,

674   Weyand CA, Bradburd GS, Riginos C, Toonen RJ, Crandall ED. 2021. Poor data

675   stewardship will hinder global genetic diversity surveillance. *Proceedings of the National*

676   *Academy of Sciences* 118:e2107934118. DOI: 10.1073/pnas.2107934118.

677   Trevisan B, Alcantara DMC, Machado DJ, Marques FPL, Lahr DJG. 2019. Genome skimming is

678   a low-cost and robust strategy to assemble complete mitochondrial genomes from ethanol

679   preserved specimens in biodiversity studies. *PeerJ* 7:e7543. DOI: 10.7717/peerj.7543.

680   Ward RD, Hanner R, Hebert PDN. 2009. The campaign to DNA barcode all fishes, FISH-BOL.

681   *Journal of Fish Biology* 74:329–356. DOI: 10.1111/j.1095-8649.2008.02080.x.

682   West KM, Stat M, Harvey ES, Skepper CL, DiBattista JD, Richards ZT, Travers MJ, Newman

683   SJ, Bunce M. 2020. eDNA metabarcoding survey reveals fine-scale coral reef community

684   variation across a remote, tropical island ecosystem. *Molecular Ecology* 30:246. DOI:

685   10.1111/mec.15382.

686   White WT, Corrigan S, Yang L, Henderson AC, Bazinet AL, Swofford DL, Naylor GJP. 2018.

687   Phylogeny of the manta and devilrays (Chondrichthyes: Mobulidae), with an updated

688   taxonomic arrangement for the family. *Zoological Journal of the Linnean Society*

689   182:50–75. DOI: 10.1093/zoolinnean/zlx018.

690   Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. 2017. ggtree: An R package for visualization and

691   annotation of phylogenetic trees with their covariates and other associated data. *Methods*

692   *in Ecology and Evolution* 8:28–36. DOI: 10.1111/2041-210X.12628.

693

Figure 1

Species included in this MiSeq-based pilot study.

(A) *Gymnura altavela*, Spiny Butterfly Ray, length unknown. (B) *Gymnothorax fimbriatus*, Fimbriated moray, USNM 395396, 850 mm TL. (C) *Gymnothorax undulatus*, Undulated moray, USNM 442319, 132 mm TL. (D) *Saurida nebulosa*, Clouded Lizardfish, USNM 442473, 56.2 mm SL. (E) *Brosme brosme*, Cusk, length unknown. (F) *Myripristis vittata*, Whitetip Soldierfish, USNM 411102, 120.1 mm SL. (G) *Neoniphon sammara*, Sammara Squirrelfish, USNM 442483, 130 mm SL. (H) *Tylosurus crocodilus*, Houndfish, USNM 442362, 13.6 mm SL. (I) *Scomberoides lysan*, Doublespotted Queenfish, USNM 442297, 22.3 mm SL. (J) *Forcipiger flavissimus*, Longnose Butterflyfish, USNM 411089, 129.1 mm SL. (K) *Ostracion whitleyi*, Whitley's Boxfish, USNM 411029, 81.2 mm SL. (L) *Canthigaster amboinensis*, Ambon Toby, USNM 442417, 64 mm SL. All photographs except A and E are the individuals for which we sequenced the mitogenome. Photographs A and E by Donald D. Flescher, NOAA; photographs B, F, J, and K by Jeff Williams, NMNH; and photographs C, D, G, H, I, and L by Diane Pitassy NMNH.

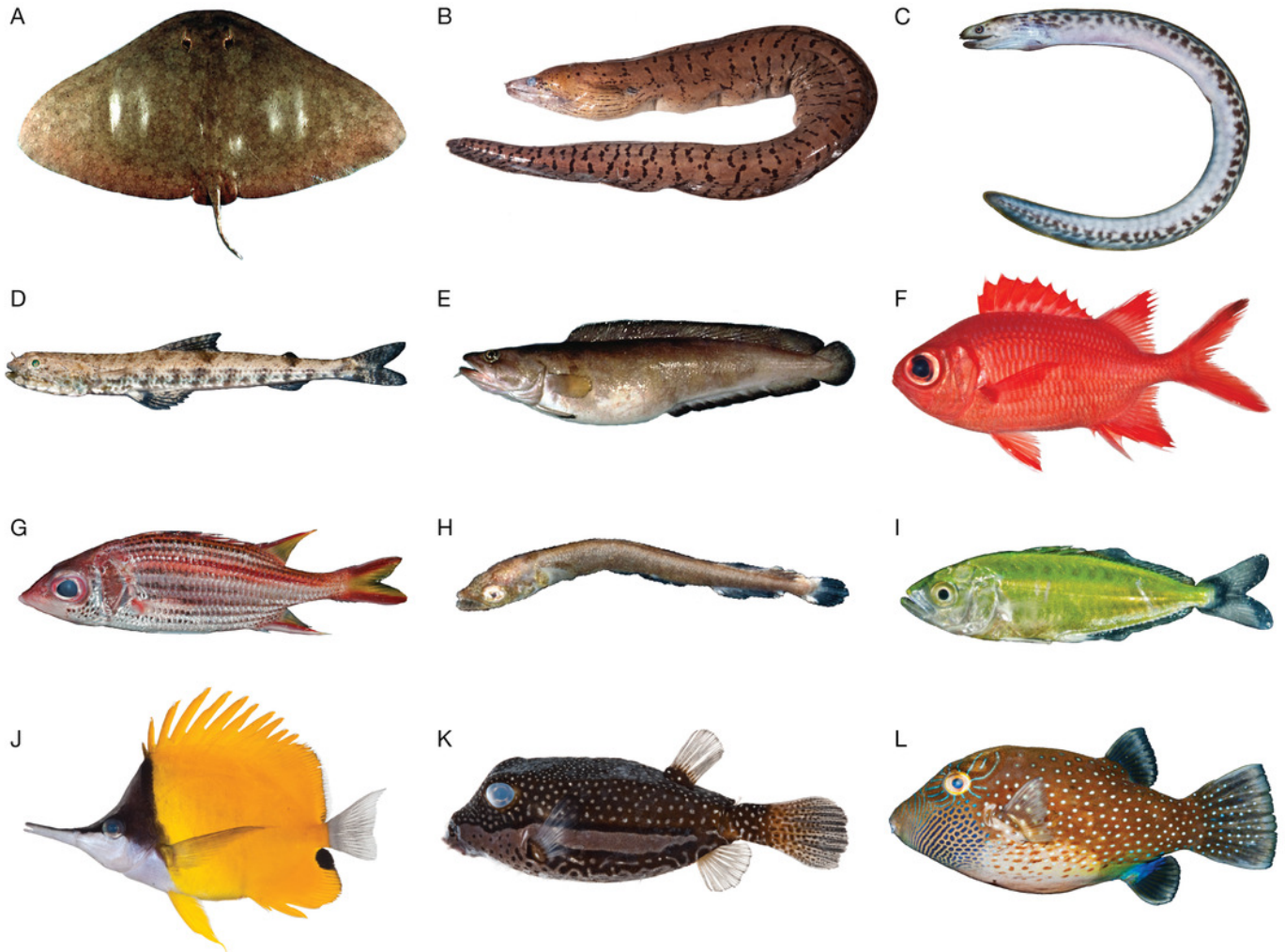



Figure 2

Assembled and annotated mitogenome of *Canthigaster amboinensis*, Ambon Toby, USNM 442417, 64 mm SL.

Photograph by Diane Pitassy, NMNH.

Canthigaster amboinensis
Ambon Toby
USNM 442417

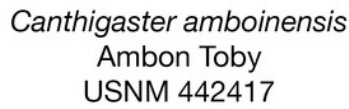
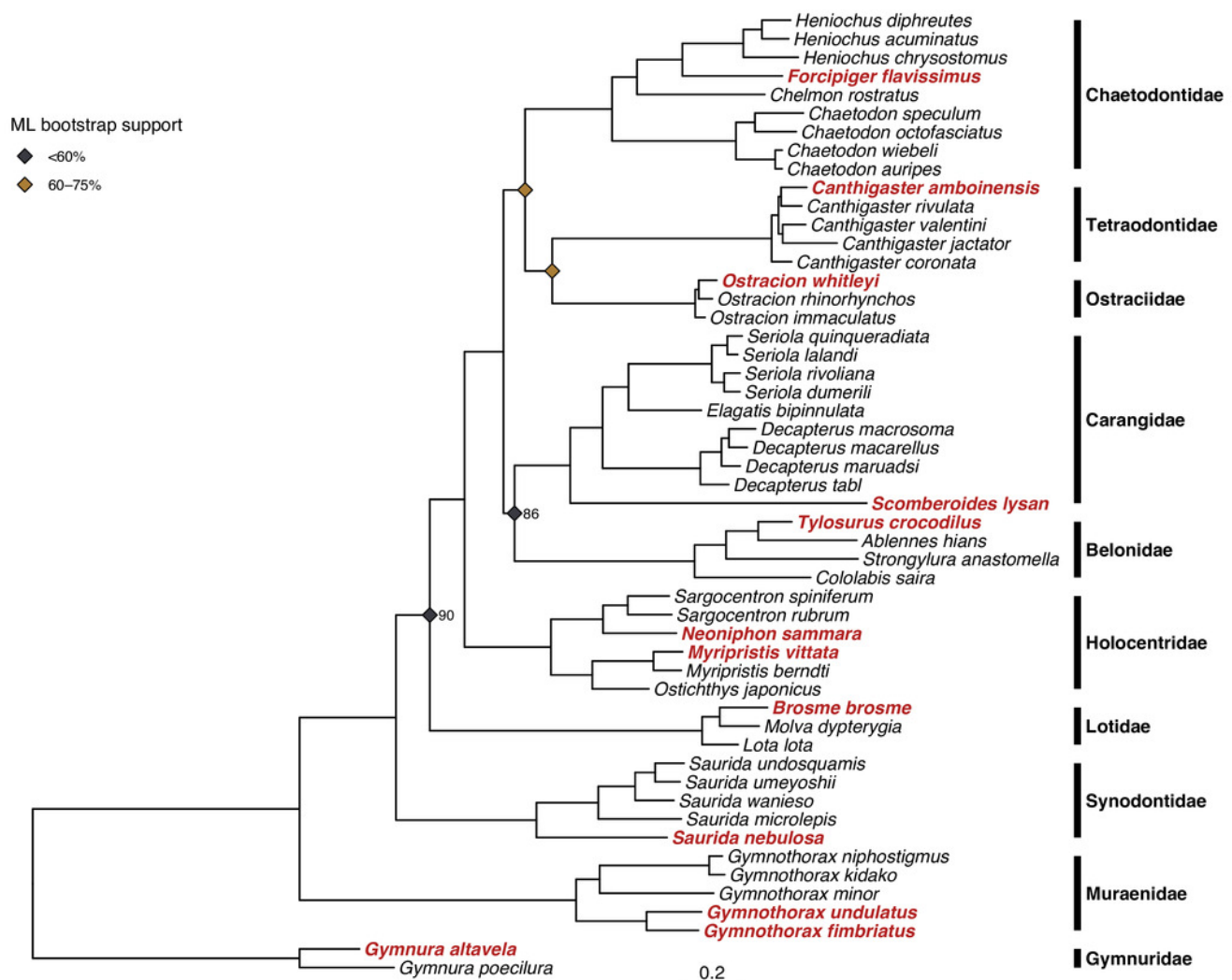



Figure 3

Results of phylogenetic analysis of 52 fish mitogenomes.

Tree shown is the result of the Bayesian analysis confirming that 12 focal taxa (shown in red) are correctly placed among confamilials in corresponding families. Node support values <95% are shown for nodes at family- and genus-level splits. Bayesian posterior probability is as labeled, and ML bootstrap support is indicated by the color of the node symbols. Unlabeled family- and genus-level nodes had 100% posterior probability and bootstrap support.

Summary of species and museum specimens included in this study.

Species in this and subsequent tables are arranged alphabetically by taxonomic order, family, and scientific name, with the chondrichthyan presented separately.

Scientific name	Order	Family	Extraction method	Estimated genome size (Gb)	USNM catalog number	Date Collected	COI reference accession
Gymnura altavela (Linnaeus, 1758)	Myliobatiformes	Gymnuridae	AutoGen	1.80[b]	433343	11 Sep. 2006	MH378654
Gymnothorax fimbriatus (Bennett, 1832)	Anguilliformes	Muraenidae	BioSprint	2.31[c]	395396	15 Oct. 2008	MK658634
Gymnothorax undulatus (Lacepède, 1803)	Anguilliformes	Muraenidae	AutoGen	2.31[c]	442319	26 May 2017	MG816692
Saurida nebulosa Valenciennes, 1850	Aulopiformes	Synodontidae	AutoGen	1.53[c]	442473	1 Jun. 2017	MG816726
Tylosurus crocodilus (Péron & Lesueur, 1821)	Beloniformes	Belonidae	AutoGen	1.00[s]	442362	28 May 2017	MG816741
Myripristis vittata Valenciennes, 1831	Beryciformes	Holocentridae	BioSprint	0.90[c]	411102	16 Oct. 2008	MZ598162
Neoniphon sammara (Forsskål, 1775)	Beryciformes	Holocentridae	AutoGen	0.80[s]	442483	31 May 2017	MG816708
Brosme brosme (Ascanius, 1772)	Gadiformes	Lotidae	AutoGen	0.41[s]	433199	20 Apr. 2008	MH378533
Scomberoides lysan (Forsskål, 1775)	Perciformes	Carangidae	AutoGen	0.73[c]	442297	25 May 2017	MG816730
Forcipiger flavissimus Jordan & McGregor, 1898	Perciformes	Chaetodontidae	BioSprint	0.72[s]	411089	17 Oct. 2008	MK657435

Ostracion whitleyi Fowler, 1931	Tetraodontiformes	Ostraciidae	BioSprint	0.98[c]	411029	15 Oct. 2008	MK658705
Canthigaster amboinensis (Bleeker, 1864)	Tetraodontiformes	Tetraodontidae	AutoGen	0.41[c]	442417	30 May 2017	MG816661

1   [s] Genome size estimates were available for this exact species on NCBI and/or genomesize.com

2   [c] Genome size estimates were calculated based on an average of available congeners or confamilials on NCBI and/or genomesize.com

3   [b] Genome size estimate for this species was based on an average of members of Batoidea available on NCBI and/or genomesize.com

GenBank accession numbers for assembled mitogenomes and ribosomal repeat regions.

Species	Accession Number (mitogenome)	Mitogeone length (bp)	Accession number (ribosomal repeat region)	DOI for Genome preassemblies and assembly statistics
Gymnura altavela	OK104094	19,022[a]	MZ286332	10.5281/zenodo.5507151
Gymnothorax fimbriatus	MZ297479	16,567	MZ286333	10.5281/zenodo.5507064
Gymnothorax undulatus	MZ329992	16,566	MZ286339	10.5281/zenodo.5507172
Saurida nebulosa	MZ329994	16,717	MZ286340	10.5281/zenodo.5507186
Tylosurus crocodilus	MZ329993	16,533	MZ286342	10.5281/zenodo.5507182
Myripristis vittata	MZ329989	16,520	MZ286336	10.5281/zenodo.5507128
Neoniphon sammara	MZ329995	16,743	MZ286341	10.5281/zenodo.5507201
Brosme brosme	MZ329990	16,483	MZ286337	10.5281/zenodo.5507143
Scomberoides lysan	MZ329991	16,767	MZ286338	10.5281/zenodo.5507164
Forcipiger flavissimus	MZ329988	16,600	MZ286335	10.5281/zenodo.5507111
Ostracion whitleyi	MZ297480	16,461	MZ286334	10.5281/zenodo.5507077
Canthigaster amboinensis	MZ188982	16,444	MZ188965	10.5281/zenodo.4753123

1 [a] based on nearly-complete mitogenome assembly

Library quantification and sequencing results; values shown are for both shearing methods (mechanical; enzymatic).

Species	Input DNA for library preparation (ng)	Average library size (bp)	Final library concentration (ng/µL)	Total raw reads	Calculated genome coverage	Reads mapped to mitogenome	Percent reads mapped	Avg. mitogenome coverage
Gymnura altavela	170	318; 326	2.50; 1.98	2,193,690; 2,224,022	0.18; 0.19	201; 1,141	0.01; 0.05	1.6; 8.9
Gymnothorax fimbriatus	78	353; 370	0.498; 1.31	1,522,912; 1,809,632	0.10; 0.12	2,336; 2,647	0.15; 0.15	20.7; 23.0
Gymnothorax undulatus	51	356; 379	0.984; 2.82	2,146,906; 5,168,856	0.14; 0.34	984; 2,245	0.05; 0.04	8.7; 19.5
Saurida nebulosa	27.6	353; 391	0.382; 1.87	2,120,606; 3,174,282	0.21; 0.31	5,290; 5,603	0.25; 0.18	47.1; 48.7
Tylosurus crocodilus	4.6	380; 390	0.156; 0.27	463,424; 2,451,640	0.07; 0.37	1,065; 5,507	0.23; 0.22	9.4; 48.6
Myripristis vittata	25.1	337; 354	0.352; 1.42	1,290,468; 2,342,102	0.21; 0.39	754; 1,615	0.06; 0.07	6.7; 13.8
Neoniphon sammara	17.1	352; 375	0.286; 0.876	2,276,566; 4,265,046	0.43; 0.80	2,169; 3,957	0.10; 0.09	19.3; 34.6
Brosme brosme	41	334; 392	0.366; 1.79	1,027,598; 1,635,836	0.37; 0.69	3,321; 5,148	0.32; 0.31	29.4; 45.1
Scomberoides lysan	33.9	340; 378	0.344; 1.30	2,621,818; 4,818,598	0.54; 0.99	7,249; 12,324	0.28; 0.26	64.2; 107.9
Forcipiger flavissimus	109	351; 378	1.06; 2.96	1,993,702; 2,116,356	0.41; 0.44	1,193; 1,311	0.06; 0.06	10.5; 11.1
Ostracion whitleyi	86.5	340; 371	1.32; 3.34	2,054,668; 2,473,712	0.31; 0.38	2,369; 3,069	0.12; 0.12	20.596; 27.089
Canthigaster amboinensis	19.1	331; 371	0.224; 0.678	1,880,384; 2,868,978	0.68; 1.04	6,070; 8,672	0.32; 0.30	53.132; 76.469