



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

Mykle L. Hoban¹, Jonathan Whitney², Allen G. Collins³, Christopher Meyer⁴, Katherine R. Murphy⁵, Abigail J. Reft³ and Katherine E. Bemis³

¹ Hawai'i Institute of Marine Biology, University of Hawai'i at Mānoa, Kāne'ohe, Hawai'i, United States of America

² Pacific Islands Fisheries Science Center, National Oceanic and Atmospheric Administration, Honolulu, Hawai'i, United States of America

³ NOAA National Systematics Laboratory, Natural Museum of Natural History, Smithsonian Institution, Washington, D.C., United States of America

⁴ Department of Invertebrate Zoology, National Museum of Natural History, Smithsonian Institution, Washington, D.C., United States of America

⁵ Laboratories of Analytical Biology, National Museum of Natural History, Smithsonian Institution, Washington, D.C., United States of America

ABSTRACT

DNA barcoding is critical to conservation and biodiversity research, yet public reference databases are incomplete. Existing 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 separate PCR reactions, and limits resulting sequences to targeted regions. High-throughput sequencing techniques such as genome skimming enable assembly of complete mitogenomes, which contain the most commonly used barcoding loci (*e.g.*, COI, 12S, 16S), as well as nuclear ribosomal repeat regions (*e.g.*, ITS1&2, 18S). We evaluated the feasibility of genome skimming to generate barcode reference databases for marine fishes by assembling complete mitogenomes and nuclear ribosomal repeats. We tested genome skimming across a taxonomically diverse selection of 12 marine fish species from the collections of the National Museum of Natural History, Smithsonian Institution. We generated two sequencing libraries per species to test the impact of shearing method (enzymatic or mechanical), extraction method (kit-based or automated), and input DNA concentration. We produced complete mitogenomes for all non-chondrichthyans (11/12 species) and assembled nuclear ribosomal repeats (18S-ITS1-5.8S-ITS2-28S) for all taxa. The quality and completeness of mitogenome assemblies was not impacted by shearing method, extraction method or input DNA concentration. Our results reaffirm that genome skimming is an efficient and (at scale) cost-effective method to generate all mitochondrial and common nuclear DNA barcoding loci for multiple species simultaneously, which has great potential to scale

Submitted 3 May 2022
Accepted 5 July 2022
Published 5 August 2022

Corresponding author
Mykle L. Hoban,
mhoban@hawaii.edu,
mh@myklehoban.com

Academic editor
Xavier Pochon

Additional Information and
Declarations can be found on
page 15

DOI 10.7717/peerj.13790

© Copyright
2022 Hoban et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

for future projects and facilitate completing barcode reference databases for marine fishes.

Subjects Biodiversity, Genetics, Marine Biology

Keywords DNA barcoding, Genome skimming, Mitochondrial genomes, Fishes, Metabarcoding, Collections, Museum

INTRODUCTION

DNA barcoding has long been recognized as a critical component of biodiversity research (Hebert *et al.*, 2003; Ratnasingham & Hebert, 2013; DeSalle & Goldstein, 2019; Adamowicz *et al.*, 2019), but available barcode reference databases remain incomplete (Mugnai *et al.*, 2021). More comprehensive regional reference datasets in global databases better support research goals and applications such as discovering new species (Carpenter, Williams & Santos, 2017; Hoban & Williams, 2020), matching larval specimens to known adults (Johnson *et al.*, 2009; Hubert *et al.*, 2010), and authenticating seafood labeling (Marko, Nance & Guynn, 2011; Silva & Hellberg, 2021). Traditionally, DNA barcoding efforts relied on Sanger sequencing of single mitochondrial markers, particularly cytochrome oxidase subunit I (COI) for metazoans. However, there is increasing utility for other mitochondrial genes and noncoding regions (*e.g.*, 16S, 12S) as well as nuclear ribosomal genes that are present in tandem repeats (*e.g.*, 18S-ITS1-5.8S-ITS2-28S) (Pochon *et al.*, 2013; Berry *et al.*, 2017; Alexander *et al.*, 2020). To develop more complete DNA barcode databases, we evaluated a method of genome skimming that has potential to simultaneously recover multiple barcoding loci for many species.

DNA barcodes are essential resources, but the quality and utility of existing data is variable. For DNA barcodes to be of long-term value, they should be linked to physical voucher specimens in permanent natural history collections because voucher specimens allow for verification of identification and refinements in taxonomy (Schander & Willassen, 2005; Ward, Hanner & Hebert, 2009; but see Collins & Cruickshank, 2013). Another consideration for building barcode libraries stems from natural genetic variation in populations. For example, Hawaiian populations of widespread Indo-Pacific fishes are often genetically divergent and can comprise cryptic lineages (DiBattista *et al.*, 2010; DiBattista *et al.*, 2012; Bowen *et al.*, 2013). Thus, the most valuable barcode sequences are derived from voucher specimens associated with precise geospatial metadata (geotags), which are unfortunately missing for most archived genomic datasets (Toczydlowski *et al.*, 2021). Other attributes, such as color photographs of the specimen at the time of collection and detailed collection metadata, add to barcode value. Finally, to increase discoverability and data access, specimen and sequence metadata need to be linked through persistent digital identifiers across systems of record (Riginos *et al.*, 2020). These best practices in data stewardship are necessary to support cross-domain cyberinfrastructure to enable transdisciplinary research, discovery and reuse of material samples and their derived data (Davies *et al.*, 2021).

Efforts to characterize community biodiversity patterns through metabarcoding ([Leray & Knowlton, 2015](#); [Timmers et al., 2021](#)) and environmental DNA (eDNA) surveys ([Ficetola et al., 2008](#))—that rely on well-curated barcode databases to accurately assign sequences to taxonomy—have expanded dramatically ([Ruppert, Kline & Rahman, 2019](#)). In addition, approaches (such as eDNA) that are based on potentially fragmentary source material and/or those that target specific taxa are more precise with a multi-marker approach ([Stat et al., 2017](#); [West et al., 2020](#); [Casey et al., 2021](#)). Finally, targeting short hypervariable loci (e.g., [Riaz et al., 2011](#); [Miya et al., 2015](#)) can be more compatible with read lengths produced by high-throughput sequencing (HTS) platforms. The availability of multiple genetic markers associated with a single voucher specimen also makes species identifications more consistent across studies where researchers may use different loci.

As high-throughput sequencing has become more accessible and cost-effective, genome skimming, which uses low-pass, shallow shotgun sequencing of whole genomes, has become practical ([Trevisan et al., 2019](#)). Genome skimming does not enrich samples for specific target loci, yet it is successful at recovering high-copy regions such as mitochondrial and plastid genomes as well as nuclear or cytosolic sequences like ribosomal DNA ([Kane et al., 2012](#); [Straub et al., 2012](#); [Besnard et al., 2013](#); [Malé et al., 2014](#); [Ripma, Simpson & Hasenstab-Lehman, 2014](#); [Dodsworth, 2015](#); [Denver et al., 2016](#); [Grandjean et al., 2017](#); [Liu et al., 2020](#); [Raupach et al., 2022](#)). Genome skimming has great potential to fill DNA barcode reference databases because it generates sequence data for commonly used barcoding markers simultaneously ([Coissac et al., 2016](#)). This potential has been realized in a range of taxa from plants ([Alsos et al., 2020](#)) to arthropods ([Grandjean et al., 2017](#); [Raupach et al., 2022](#)). Our work complements [Therkildsen & Palumbi \(2017\)](#), who used a similar approach to examine genetic variation in Atlantic Silversides and [Margaryan et al. \(2021\)](#), who developed a mitogenome barcode database for vertebrates of Denmark, and extends these studies by showing that ribosomal barcoding loci are also readily accessible using genome skimming. Despite previous applications of genome skimming, it has yet to be tested broadly to capture specimen-backed DNA barcodes for marine fishes.

Natural history collections hold valuable materials to support regional and taxon-specific barcode database development, allowing gaps to be filled without the need to collect new specimens. While many institutions voucher tissue samples and/or DNA extractions alongside collected specimens, sequences are frequently published for only a limited number of loci (e.g., COI for metazoans, ITS for fungi ([Ratnasingham & Hebert, 2007](#))). In our study, which is part of an ongoing effort to complete the barcode reference database for Hawaiian marine fishes, we evaluated genome skimming as a method to rapidly and (when scaled up to massively parallel sequencing platforms) inexpensively capture all commonly used DNA barcoding loci for multiple samples and fish taxa simultaneously. Using genome skimming, we aimed to recover the complete mitochondrial genomes and ribosomal repeat regions of 12 taxonomically diverse species of marine fishes. For our test, we prepared and sequenced two libraries for each species (24 libraries total) from vouchered specimens in the National Museum of Natural History (NMNH) fish collection. We evaluated the quality of sequences and our ability to assemble complete mitogenomes and ribosomal repeats in the context of taxonomic diversity and shearing method, and across a range of

DNA extraction methods and input DNA concentrations. Here we report the results of our test and discuss how to adapt this method for large-scale generation of specimen-backed DNA barcodes.

MATERIALS & METHODS

Sample selection

We selected samples from 12 species across a broad taxonomic distribution of fishes, including one chondrichthyan and 11 teleosts (Fig. 1). This work is a component of an effort to generate specimen-backed barcodes for all species of Hawaiian marine fishes (~1,200 species; unpublished updated version of Mundy, 2005; Randall, 2007); thus, most specimens were Hawaiian species collected in Hawai'i (6/12) or species that occur in Hawai'i but that were collected elsewhere (3/12). We also included two western North Atlantic species: *Brosme brosme* (Cusk), which is a NOAA species of concern, and *Gymnura altavela* (Spiny Butterfly Ray), as a representative chondrichthyan. All samples were taken from existing DNA extracts in the National Museum of Natural History (NMNH) Biorepository, derived from specimens housed in the fish collection at NMNH (Table 1). Archived Biorepository DNA was originally extracted from tissues subsampled and preserved in the field at the time of specimen collection. Ten of the 12 specimens have live color photographs (Fig. 1). No mitogenomes or ribosomal repeats were available in GenBank for any of the species selected except *Gymnura altavela*, which was published during preparation of this manuscript (Kousteni et al., 2021). All selected Hawaiian species lacked regionally localized specimen-backed barcodes for at least one common fish 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

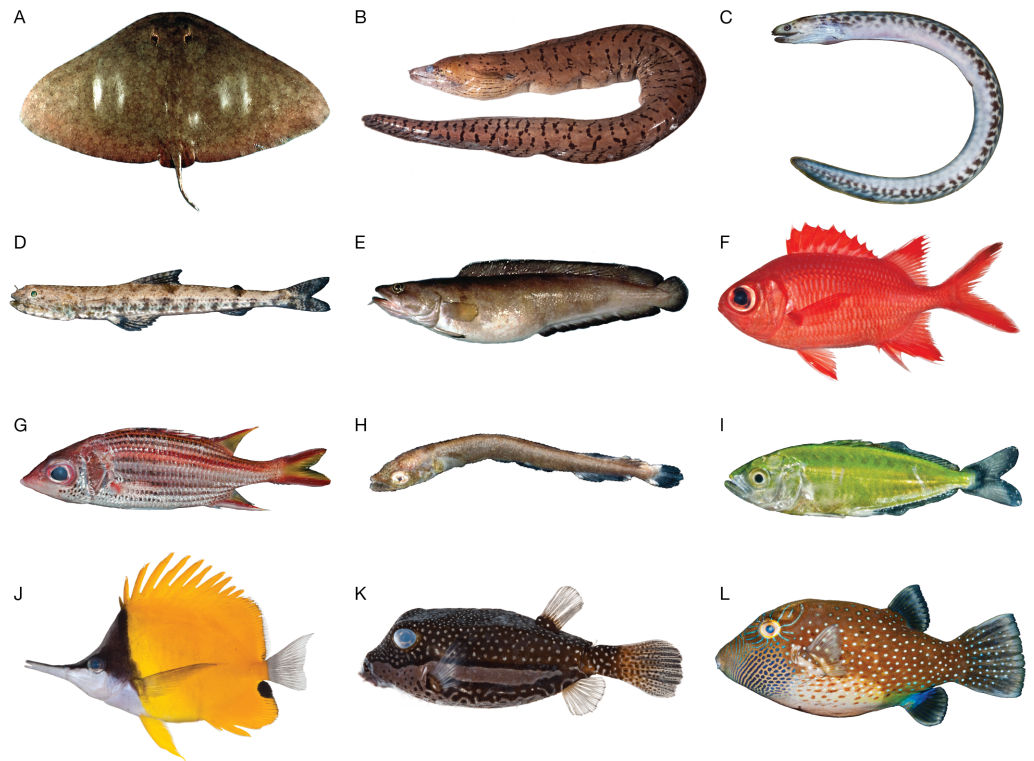


Figure 1 Species included in this MiSeq-based pilot study. (A) *Gymnura altavela*, Spiny Butterfly Ray, length unknown. (B) *Gymnothorax fimbriatus*, Fimbriated moray, USNM 395396, 850 mm TL. (C) *Gymnothorax undulatus*, Undulated moray, USNM 442319, 132 mm TL. (D) *Saurida nebulosa*, Clouded Lizardfish, 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.

Full-size  DOI: [10.7717/peerj.13790/fig-1](https://doi.org/10.7717/peerj.13790/fig-1)

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 min compared to those

Table 1 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
<i>Gymnura altavela</i> (Linnaeus, 1758)	Myliobatiformes	Gymnuridae	AutoGen	1.80 ^c	433343	11 Sep. 2006	MH378654
<i>Gymnothorax fimbriatus</i> (Bennett, 1832)	Anguilliformes	Muraenidae	BioSprint	2.31 ^b	395396	15 Oct. 2008	MK658634
<i>Gymnothorax undulatus</i> (Lacepède, 1803)	Anguilliformes	Muraenidae	AutoGen	2.31 ^b	442319	26 May 2017	MG816692
<i>Saurida nebulosa</i> Valenciennes, 1850	Aulopiformes	Synodontidae	AutoGen	1.53 ^b	442473	1 Jun. 2017	MG816726
<i>Tylosurus crocodilus</i> (Péron & Lesueur, 1821)	Beloniformes	Belonidae	AutoGen	1.00 ^a	442362	28 May 2017	MG816741
<i>Myripristis vittata</i> Valenciennes, 1831	Beryciformes	Holocentridae	BioSprint	0.90 ^b	411102	16 Oct. 2008	MZ598162
<i>Neoniphon sammara</i> (Forsskål, 1775)	Beryciformes	Holocentridae	AutoGen	0.80 ^a	442483	31 May 2017	MG816708
<i>Brosme brosme</i> (Ascanius, 1772)	Gadiformes	Lotidae	AutoGen	0.41 ^a	433199	20 Apr. 2008	MH378533
<i>Scomberoides lysan</i> (Forsskål, 1775)	Perciformes	Carangidae	AutoGen	0.73 ^b	442297	25 May 2017	MG816730
<i>Forcipiger flavissimus</i> Jordan & McGregor, 1898	Perciformes	Chaetodontidae	BioSprint	0.72 ^a	411089	17 Oct. 2008	MK657435
<i>Ostracion whitleyi</i> Fowler, 1931	Tetraodontiformes	Ostraciidae	BioSprint	0.98 ^b	411029	15 Oct. 2008	MK658705
<i>Canthigaster amboinensis</i> (Bleeker, 1864)	Tetraodontiformes	Tetraodontidae	AutoGen	0.41 ^b	442417	30 May 2017	MG816661

Notes.

^aGenome size estimates were available for this exact species on NCBI and/or genomesize.com.

^bGenome size estimates were calculated based on an average of available congeners or confamilials on NCBI and/or genomesize.com.

^cGenome size estimate for this species was based on an average of members of Batoidea available on NCBI and/or genomesize.com.

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 150 bp (paired end) to test scalability to higher-throughput platforms such as the Illumina NovaSeq 6000.

Assembly

We assessed two approaches to mitogenome assembly using Geneious Prime 2021.2.2 (<https://www.geneious.com>). First, we used the Map to Reference function and built-in Geneious mapper with the sensitivity set to “medium/low” and iterations set to “up to 10 times”, starting with published COI sequences (Table 1) for each of the 24 libraries. Resulting assemblies were inspected and trimmed at the ends (up to 50 bp) where coverage

was low (<5X). Consensus sequences were generated from the assembly results and used as subsequent reference seeds and the Map to Reference step repeated until the assemblies stopped increasing in size and identical stretches of sequences were detected at the 5' and 3' ends. The second approach used a complete mitogenome from either a congeneric or confamilial taxon as the reference sequence, and Map to Reference, using the same parameters for a single set of up to 10 iterations. Assemblies of ribosomal repeat regions were conducted similarly, with reiterations using the Map to Reference function in Geneious, using ribosomal sequences from closely related taxa published in GenBank (Table S2). In addition to assembling mitogenomes, we constructed nuclear genome preassemblies using SPAdes 3.15.3 (assembly module only) on paired forward and reverse read libraries (Prjibelski et al., 2020), and filtered out preassembly contigs shorter than 200 bp.

Genome sequencing coverage estimation

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

Annotation

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

Phylogenetic analyses

To assess relationships and validate taxonomic identities, we performed phylogenetic analyses including all mitogenomes generated in this study and confamilial taxa with published mitogenomes available in the MitoFish database (52 species; Table S5). Due to the large number of species with available mitogenomes in the family Carangidae, we only used species in *Seriola*, *Elegatis*, and *Decapterus*, the available genera most closely related to our taxon *Scomberoides lysan* (Damerou, Freese & Hanel, 2018; Rabosky et al., 2018). We used sequences of all protein-coding genes (PCGs) and two rRNAs. Each PCG or rRNA was individually aligned using MAFFT v7.505 (Kato & Standley, 2013) and then concatenated to a single final alignment. We used PartitionFinder2 to assess the partitioning of models of molecular evolution (Guindon et al., 2010; Lanfear et al., 2012; Lanfear et al., 2017). We partitioned the alignment by gene and, for PCGs, by codon position, then ran PartitionFinder2 using the “greedy” algorithm with branch lengths specified as unlinked to test the models supported in MrBayes. We conducted a Bayesian phylogenetic reconstruction using MrBayes v3.2.7 (Ronquist et al., 2012), running four independent searches of six chains for 22 million generations, saving trees every 1,000

Table 2 GenBank accession numbers for assembled mitogenomes and ribosomal repeat regions.

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

Notes.

^aBased on nearly-complete mitogenome assembly.

generation and discarding the first 15% as burn-in. We verified convergence of MCMC runs and model parameters using TRACER v1.7.2 (Rambaut *et al.*, 2018). We conducted a maximum likelihood (ML) phylogenetic reconstruction with the partitioned alignment using RAxML v8.2.12 (Stamatakis, 2014) and specified 1,000 bootstrap replicates to assess node support. Resulting trees were rooted to the two *Gymnura* species and plotted using ggtree (Yu *et al.*, 2017) and phytools (Revell, 2012) in R v4.1.2 (R Core Team, 2020).

Data availability

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

RESULTS**DNA concentration**

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

Table 3 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/ μ L)	Total raw reads	Calculated genome coverage	Reads mapped to mitogenome	Percent reads mapped	Avg. mitogenome coverage
<i>Gymnura altavela</i>	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
<i>Gymnothorax fimbriatus</i>	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
<i>Gymnothorax undulatus</i>	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
<i>Saurida nebulosa</i>	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
<i>Tylosurus crocodilus</i>	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
<i>Myripristis vittata</i>	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
<i>Neoniphon sammara</i>	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
<i>Brosme brosme</i>	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
<i>Scomberoides lysan</i>	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
<i>Forcipiger flavissimus</i>	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
<i>Ostracion whitleyi</i>	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
<i>Canthigaster amboinensis</i>	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

Sequence reads and genome coverage

We recovered 0.46 to 5.2 million reads (2.5 ± 1.1 million) per library. AutoGen and Qiagen extractions performed comparably (2.6 ± 1.3 million reads for AutoGen vs. 2.0 ± 0.4 million for Qiagen). Enzymatic shearing yielded more reads per library than mechanical shearing (2.9 ± 1.1 million reads for enzymatic vs. 1.8 ± 0.6 million reads for mechanical). Sequence duplication rates varied from 0.7–6.5% per sample. Based on estimated genome sizes, these read counts equate to $0.07\times$ to $1.04\times$ genome coverage, with enzymatic shearing ($0.50 \pm 0.30\times$) averaging higher than mechanical shearing ($0.30 \pm 0.19\times$). A summary of sequencing results across libraries is presented in [Table 3](#).

Assembly and sequence coverage

We readily assembled and annotated complete mitochondrial genomes for the 11 teleosts (see [Table 2](#) for assembled mitogenome accession numbers). Assembled sequences were identical whether we started from a small seed (COI) or mapped to a complete mitochondrial reference genome derived from a congeneric or confamilial taxon. We did not recover a complete mitogenome from *Gymnura altavela* (Spiny Butterfly Ray),

but assembled large sections of it (e.g., ~12,000 bp including COI; ~3,000 bp including 16S). During the course of this work, a complete mitochondrial genome was published for *G. altavela* (MT274571) based on a specimen from Greece (Kousteni et al., 2021). This allowed us to improve our assembly, resulting in a mitochondrial genome with a short gap in COI and a second gap in the D-loop. Fortunately, the gap spanned the published COI sequence for our specimen (USNM 433343; MH378654), allowing us to use 24 bases from that sequence to fill the missing space. As a result, we ultimately derived a nearly-complete mitochondrial genome for the Spiny Butterfly Ray (19,022 bp in our assembly as compared to 19,472 bp in MT274571).

Mitogenome coverage of the 22 successful assemblies ranged from $7\times$ to $108\times$ ($34 \pm 26\times$; Table 3). The *Gymnura altavela* libraries had a comparable number of reads to other species in our study, but coverage of the mitogenome was low for unknown reasons ($11.2\times$ with both libraries combined). Across all libraries, assembled mitogenome reads comprised 0.05% to 0.32% ($0.17 \pm 0.1\%$) of the total raw reads generated per specimen.

Using Geneious Map to Reference, we assembled and annotated ribosomal repeat regions (18S-ITS1-5.8S-ITS2-28S) for all 12 taxa by using 18S or 28S reference seeds (see Table 2 for assembled ribosomal repeat accession numbers).

Genome preassemblies generated by SPAdes (>200 bp) were uploaded to Zenodo (along with basic assembly statistics) and assigned persistent identifiers (Table 2). As expected, the preassemblies were limited, with a small fraction of contigs exceeding 1 kb in length. Nevertheless, preassembly contigs that correspond to the complete or nearly complete mitochondrial genomes and the ribosomal repeat regions were recovered for 7 and 8, respectively, of the 12 species in our study.

Mitogenome organization and structure

Mitogenomes for all species were arranged similarly, with some minor length variations, particularly in the control region (see Fig. 2 for example assembly of *Canthigaster amboinensis*; see Fig. S1 for all mitogenome assemblies). We detected no mitochondrial gene rearrangements among the 12 species we investigated. All species had 36 genes comprising 13 protein-coding genes (PCGs) and 23 tRNAs, with two rRNAs and the control region. In all cases, the majority strand encoded 12 PCGs, 15 tRNAs, both rRNAs, and the control region. The remaining eight tRNAs and a single PCG were encoded on the minority strand. GC content ranged from 43.1% (*Neoniphon sammara*) to 52.1% (*Gymnothorax fimbriatus*) (mean: $45.5 \pm 2.3\%$).

Phylogenetic analyses

Both ML and Bayesian methods produced identical topologies (Fig. 3), with the single exception of different branching order within the genus *Ostracion* (see Figs. S2 and S3 for raw ML and Bayesian trees respectively, including complete node support values). All specimens sequenced for this study were recovered within their respective taxonomic groups, and branching order among families matched that of the family-level backbone tree published in Rabosky et al. (2018). Node support in the ML tree (bootstrap value) was more variable than in the Bayesian tree (posterior probability). The ML tree had

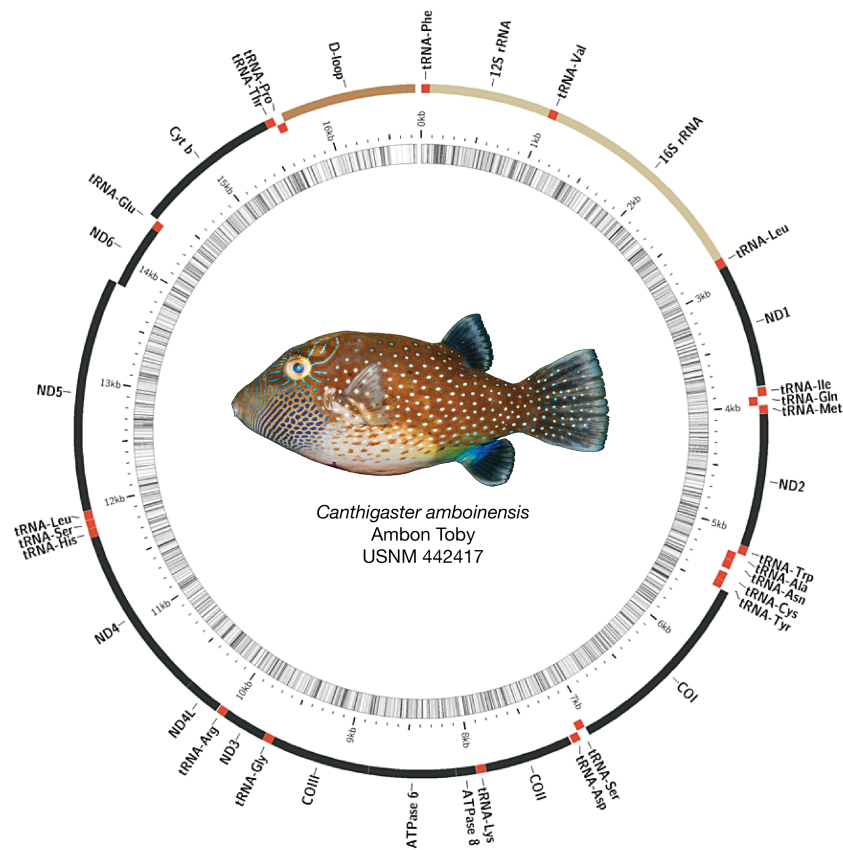


Figure 2 Assembled and annotated mitogenome of *Canthigaster amboinensis*, Ambon Toby, USNM 442417, 64 mm SL. Photograph by Diane Pitassy, NMNH.

Full-size DOI: 10.7717/peerj.13790/fig-2

many strongly supported nodes (70–100% bootstrap support), but two with weak support (<20%). The Bayesian tree was strongly supported throughout, with most nodes having >95% posterior probability.

DISCUSSION

Our results show that genome skimming by shallow shotgun sequencing is an efficient method for generating mitogenomes and ribosomal repeats of marine fishes. The methods are robust for a broad range of taxa, extraction types, shearing methods, and DNA concentrations. Both kit-based (Qiagen) and automated (AutoGen) extractions resulted in high quality sequence libraries, which indicates that this method can leverage existing DNA extractions housed in museum collections that were prepared for other purposes (e.g., single-marker Sanger sequencing).

As noted in Methods, our enzymatically sheared samples were held at 4 °C following the end of the ligation period for an additional 45 min compared to those mechanically sheared. This likely impacted their ligation efficiency and subsequent library yield. As a result, we cannot confirm that differences in final library yield nor differences in read counts resulted directly from the shearing method used. Although libraries were pooled in

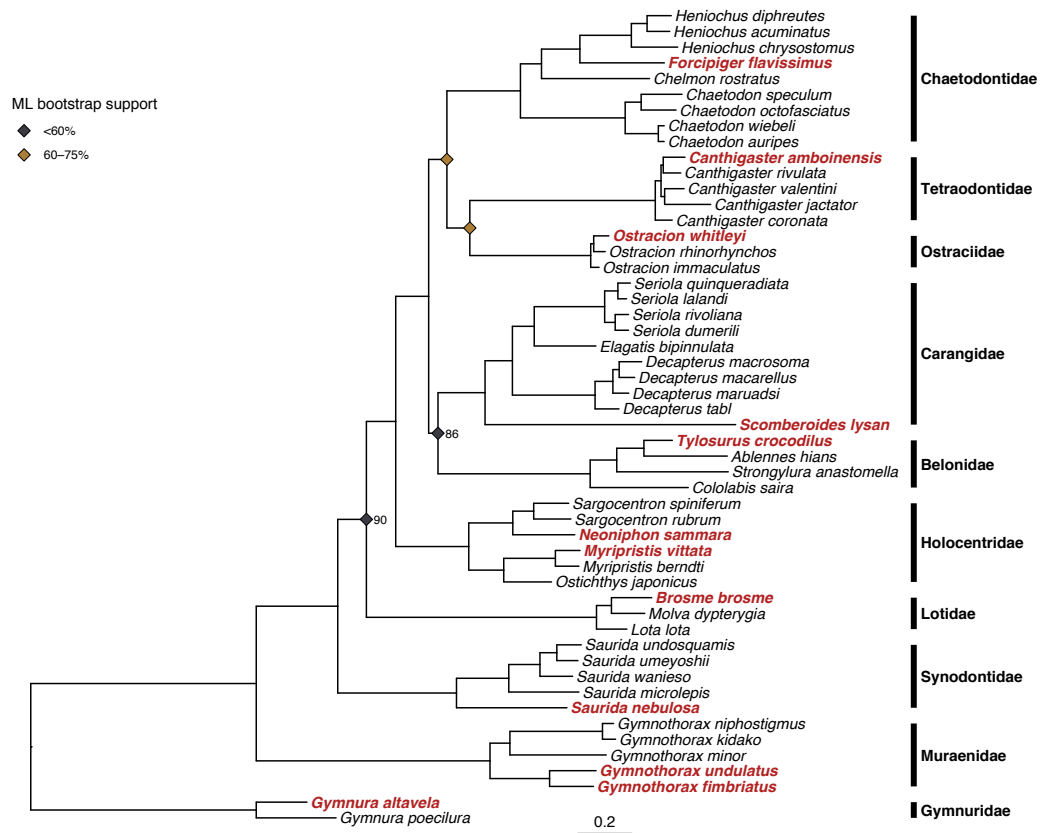


Figure 3 Results of phylogenetic analysis of 52 fish mitogenomes. Tree shown is the result of the Bayesian analysis confirming that 12 focal taxa (shown in red) are correctly placed among confamilials in corresponding families. Node support values <95% are shown for nodes at family- and genus-level splits. 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.

Full-size [DOI: 10.7717/peerj.13790/fig-3](https://doi.org/10.7717/peerj.13790/fig-3)

equimolar ratios, these values were calculated using total dsDNA (ng/μL) as measured by Qubit and TapeStation fragment size (bp). A more accurate method would be to quantify individual libraries with qPCR, thus measuring DNA that can be sequenced, rather than total DNA. Regardless of these differences, we demonstrated that enzymatic shearing can be an effective method for genome skimming; enzymatic shearing is also less expensive (~\$4 less/library; Tables S3 and S4), less labor intensive, and requires less specialized laboratory equipment.

We assembled mitogenomes with as few as half a million reads but had more consistent success with 2–3 million reads/library, which resulted in an average of 34 × coverage of the mitogenome.

Mitogenome assemblies used only 0.05% to 0.32% of the total raw sequence reads. The majority of unassembled reads were nuclear (e.g., chromosomal) and cytosolic (e.g., ribosomal RNA) sequences. The most common barcoding markers for fishes are mitochondrial: COI (Leray et al., 2013), 16S rRNA (Berry et al., 2017), and 12S rRNA (Miya et al., 2015). However, primer sets designed to amplify other taxa or communities

often target nuclear ribosomal loci such as the 18S rRNA and/or internal transcribed spacers (ITS1/2) (marine eukaryotes: [Pochon et al., 2013](#); scleractinian corals: [Alexander et al., 2020](#)). We successfully recovered complete ribosomal repeat regions (18S-ITS1-5.8S-ITS2-28S) from all of our sequence libraries, illustrating that our approach has applications beyond mitogenome assembly. Importantly, we recovered sequences for the most commonly used barcoding loci for all targeted taxa in a single pass. We provided raw sequence data in the NCBI Sequence Read Archive under BioProject [PRJNA720393](#) because there are likely additional sequences of interest to other researchers. In addition, we constructed genome preassemblies for each sample, which are also available ([Table 2](#)).

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

To test whether our methods are applicable across fish diversity, we included one chondrichthyan, the Spiny Butterfly Ray *Gymnura altavela*. Despite high success across teleosts, we did not recover a complete mitogenome for the chondrichthyan. The *G. altavela* libraries had read counts comparable to bony fish libraries, but mitogenome coverage was low and initial assemblies had gaps. Two potential causes of low coverage in *Gymnura* include exogenous (non-target) DNA and a greater number of shorter-than-expected DNA fragments in the sequencing libraries. During preparation of this manuscript, a complete mitochondrial genome was published from a specimen from Greece ([Kousteni et al., 2021](#)), and although it is ~3% diverged from our mitochondrial sequences, we used it to improve our assembly such that it included complete loci other than the D-loop. Gaps in the control region are relatively common in mitochondrial genome assemblies, particularly among rays ([Poortvliet et al., 2015](#); [Hinojosa-Alvarez et al., 2015](#)). This region often contains tandem repeats that present difficulty to bioinformatic assemblers ([White et al., 2018](#)) and have been attributed to heteroplasmy in other taxa ([Mundy, Winchell & Woodruff, 1996](#)). However, despite the D-loop gap in the complete mitogenome assembly of *G. altavela*, we still recovered targeted mitochondrial barcoding loci (COI, 12S, 16S). In future studies, we will sequence additional sharks, rays, and chimaeras to further explore laboratory and bioinformatic approaches for generating chondrichthyan mitogenomes.

We used the MiSeq platform to test shearing methods, compare extraction types and DNA concentrations, and to assess sequencing reads and coverage necessary to generate mitogenomes and ribosomal repeats across a broad taxonomic selection of fishes. To further our goal of completing barcode reference databases (for mitochondrial and ribosomal genes) for all species of Hawaiian fishes, we will sequence future genome skimming runs on an Illumina NovaSeq. The NovaSeq platform produces higher read output than MiSeq and therefore supports increased multiplexing of samples, allowing us to pool 384 samples

(species) in a single sequencing run. This will reduce sequencing costs from ~\$145 per sample on the MiSeq to ~\$16 on the NovaSeq, while also increasing the average yield from 2.5 million reads to 13 million per sample. The increased multiplexing capability of the NovaSeq brings the total cost (library preparation, quantitation, and sequencing) from ~\$161 per sample on the MiSeq to ~\$31 per sample, which will facilitate economical and rapid generation of complete mitogenomes and ribosomal repeats (encompassing all major barcoding loci) (see [Tables S3](#) and [S4](#)). Preliminary data (not reported here) from a NovaSeq run of 384 species show that our methods for mitogenome and ribosomal repeat recovery *via* genome skimming can be scaled to the higher-throughput platform. In this study, we employed manual assembly methods using Geneious Prime, whereas future assemblies will employ an automated bioinformatic pipeline to enable production of multilocus DNA barcode sequences at scale.

We enhanced the reference value of our derived genetic data through persistent digital identifiers. Raw reads and assembled sequences are linked through NCBI accessions (BioProject, BioSample, SRA, and nucleotide) to museum voucher specimens, as well as to derived tissues and DNA extracts at NMNH. Further, to ensure that data derived from and associated with these biomaterials can easily be accessed and reused, we cross linked NCBI and GEOME records through Archival Resource Key (ARK) identifiers ([Kunze, 2021](#)). Such best practices in data stewardship and the use of persistent identifiers across systems of record will facilitate cross-domain cyberinfrastructure and enable transdisciplinary research, discovery, and reuse of material samples and their derived data ([Davies et al., 2021](#)).

CONCLUSIONS

Our study shows that genome skimming is an efficient and cost effective method that will allow a shift in the DNA barcoding workflow from sequencing targeted loci in individual specimens to generating complete suites of barcode markers for many taxa in a single sequencing run. The methods we employed enable use of genetic samples housed in natural history collections to rapidly generate specimen-based, regionally localized DNA barcode reference data. This work has important implications for several large US-based initiatives: NOAA 'omics ([Goodwin et al., 2021](#)), NMNH Ocean DNA Initiative (<https://www.smithsonianmag.com/blogs/national-museum-of-natural-history/2021/07/07/meet-reef-expert-collecting-environmental-time-capsules/>), and the US Ocean Biocode ([Meyer et al., 2021](#)), each of which involves explicit aims to provide complete DNA barcode reference databases based on voucher specimens. Techniques and methods developed here are applicable to taxa and regions beyond marine fishes and the Hawaiian Islands. Comprehensive voucher-based reference databases are necessary to advance sequence-based detection, censusing, and monitoring of marine communities in the face of global change.

ACKNOWLEDGEMENTS

Samples from French Polynesia were acquired under a collaborative Centre de Recherche Insulaire et Observatoire de l'Environnement (CRIOBE) and Smithsonian Institution National Museum of Natural History (NMNH) initiative to survey the marine fishes of French Polynesia, including the Mo'orea Biocode Project. We thank NMNH and CRIOBE, in particular Jeffrey Williams (NMNH) and Serge Planes (CRIOBE), as the collectors and photographers of the samples. Specimens collected from Hawai'i were acquired under the MarineGEO Hawai'i 2017 project to survey the fishes of Kāne'ōhe Bay. We thank the Smithsonian Conservation Biology Institute, NMNH, and the Hawai'i Institute of Marine Biology, in particular Mary Hagedorn, Lynne R. Parenti, Diane Pitassy, Zeehan Jaafar, Kassi S. Cole, and Kiril Vinnikov, as the collectors of the samples. Photographs of Hawaiian fishes were provided by Diane Pitassy. The technical support and advanced computing resources from University of Hawai'i Information Technology Services – Cyberinfrastructure are gratefully acknowledged. Genetic benchwork and sequencing was completed at the Smithsonian NMNH Laboratories of Analytical Biology (LAB). At NMNH we thank Carole Baldwin, Daniel DiMichele, Chris Huddleston, Lynne R. Parenti, Diane Pitassy, Niamh Redmond, Makiri Sei, Lee Weigt, Jeff Williams, and Herman Wirshing for their support. We thank Cheryl Morrison, Andreas Zwick, and an anonymous reviewer for kind, thoughtful, and constructive comments on an earlier version of this paper. This is contribution #1893 from the Hawai'i Institute of Marine Biology, #11535 from the School of Ocean and Earth Science and Technology at the University of Hawai'i, and #2022_20 from NOAA's West Hawai'i Integrated Ecosystem Assessment Program.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received funding support from NOAA's Office of Science and Technology, NOAA's Pacific Islands Fisheries Science Center, The Cooperative Institute for Marine and Atmospheric Research, NOAA's West Hawai'i Integrated Ecosystem Assessment Program, and NMNH Laboratories of Analytical Biology. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:

NOAA's Office of Science and Technology.

NOAA's Pacific Islands Fisheries Science Center.

The Cooperative Institute for Marine and Atmospheric Research.

NOAA's West Hawai'i Integrated Ecosystem Assessment Program.

NMNH Laboratories of Analytical Biology.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Mykle L. Hoban conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Jonathan Whitney conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Allen G. Collins conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Christopher Meyer conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Katherine R. Murphy performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Abigail J. Reft analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Katherine E. Bemis conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

All sequences, assemblies, and short reads are available at GenBank: [PRJNA720393](https://www.ncbi.nlm.nih.gov/genbank/PRJNA720393).

All sample metadata is tracked in GEOME under the expedition ID NMFS_FISHES_MiSeqPilot_01 at <https://n2t.net/ark:/21547/EEV2>.

The code and alignments used to generate and plot phylogenetic trees are available at GitHub: https://github.com/hawaii-barcoding-initiative/mitogenome_tree.

Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj.13790#supplemental-information>.

REFERENCES

- Adamowicz SJ, Boatwright JS, Chain F, Fisher BL, Hogg ID, Leese F, Lijtmaer DA, Mwale M, Naaum AM, Pochon X, Steinke D, Wilson J-J, Wood S, Xu J, Xu S, Zhou X, Van der Bank M. 2019. Trends in DNA barcoding and metabarcoding. *Genome* 62:v–viii DOI 10.1139/gen-2019-0054.
- Alexander JB, Bunce M, White N, Wilkinson SP, Adam AAS, Berry T, Stat M, Thomas L, Newman SJ, Dugal L, Richards ZT. 2020. Development of a multi-assay approach for monitoring coral diversity using eDNA metabarcoding. *Coral Reefs* 39:159–171 DOI 10.1007/s00338-019-01875-9.
- Alsos IG, Lavergne S, Merkel MKF, Boleda M, Lammers Y, Alberti A, Pouchon C, Denoed F, Pitelkova I, Puşcaş M, Roquet C, Hurdu B-I, Thuiller W, Zimmermann

- NE, Hollingsworth PM, Coissac E. 2020. The treasure vault can be opened: large-scale genome skimming works well using herbarium and silica gel dried material. *Plants* 9:432 DOI 10.3390/plants9040432.
- Berry TE, Osterrieder SK, Murray DC, Coghlan ML, Richardson AJ, Grealy AK, Stat M, Bejder L, Bunce M. 2017. DNA metabarcoding for diet analysis and biodiversity: A case study using the endangered Australian sea lion (*Neophoca cinerea*). *Ecology and Evolution* 7:5435–5453 DOI 10.1002/ece3.3123.
- Besnard G, Christin P-A, Malé P-JG, Coissac E, Ralimanana H, Vorontsova MS. 2013. Phylogenomics and taxonomy of Lecomtelleae (Poaceae), an isolated panicoid lineage from Madagascar. *Annals of Botany* 112:1057–1066 DOI 10.1093/aob/mct174.
- Bowen BW, Rocha LA, Toonen RJ, Karl SA, Laboratory ToBo. 2013. The origins of tropical marine biodiversity. *Trends in Ecology & Evolution* 28:359–366 DOI 10.1016/j.tree.2013.01.018.
- Carpenter KE, Williams JT, Santos MD. 2017. *Acanthurus albimento*, a new species of surgeonfish (Acanthuriformes: Acanthuridae) from northeastern Luzon, Philippines, with comments on zoogeography. *Journal of the Ocean Science Foundation* 25:33–46 DOI 10.5281/zenodo.291792.
- Casey JM, Ransome E, Collins AG, Mahardini A, Kurniasih EM, Sembiring A, Schietekatte NMD, Cahyani NKD, Anggoro AWahyu, Moore M, Uehling A, Belcaid M, Barber PH, Geller JB, Meyer CP. 2021. DNA metabarcoding marker choice skews perception of marine eukaryotic biodiversity. *Environmental DNA* 3:1229–1246 DOI 10.1002/edn3.245.
- Coissac E, Hollingsworth PM, Lavergne S, Taberlet P. 2016. From barcodes to genomes: extending the concept of DNA barcoding. *Molecular Ecology* 25:1423–1428 DOI 10.1111/mec.13549.
- Collins RA, Cruickshank RH. 2013. The seven deadly sins of DNA barcoding. *Molecular Ecology Resources* 13:969–975 DOI 10.1111/1755-0998.12046.
- Damerau M, Freese M, Hanel R. 2018. Multi-gene phylogeny of jacks and pompanos (Carangidae), including placement of monotypic vadigo *Campogramma glaycos*. *Journal of Fish Biology* 92:190–202 DOI 10.1111/jfb.13509.
- Davies N, Deck J, Kansa EC, Kansa SW, Kunze J, Meyer C, Orrell T, Ramdeen S, Snyder R, Vieglais D, Walls RL, Lehnert K. 2021. Internet of samples (iSamples): toward an interdisciplinary cyberinfrastructure for material samples. *GigaScience* 10:giab028 DOI 10.1093/gigascience/giab028.
- Denver DR, Brown AMV, Howe DK, Peetz AB, Zasada IA. 2016. Genome skimming: a rapid approach to gaining diverse biological insights into multicellular pathogens. *PLOS Pathogens* 12:e1005713 DOI 10.1371/journal.ppat.1005713.
- DeSalle R, Goldstein P. 2019. Review and interpretation of trends in DNA barcoding. *Frontiers in Ecology and Evolution* 7:302 DOI 10.3389/fevo.2019.00302.
- DiBattista JD, Rocha LA, Craig MT, Feldheim KA, Bowen BW. 2012. Phylogeography of two closely related Indo-Pacific butterflyfishes reveals divergent evolutionary

histories and discordant results from mtDNA and microsatellites. *Journal of Heredity* **103**:617–629 DOI [10.1093/jhered/ess056](https://doi.org/10.1093/jhered/ess056).

- DiBattista JD, Wilcox C, Craig MT, Rocha LA, Bowen BW. 2010.** Phylogeography of the pacific blueline surgeonfish. *Acanthurus nigroris*, reveals high genetic connectivity and a cryptic endemic species in the hawaiian archipelago. *Journal of Marine Biology* **2011**:1–17 DOI [10.1155/2011/839134](https://doi.org/10.1155/2011/839134).
- Dodsworth S. 2015.** Genome skimming for next-generation biodiversity analysis. *Trends in Plant Science* **20**:525–527 DOI [10.1016/j.tplants.2015.06.012](https://doi.org/10.1016/j.tplants.2015.06.012).
- Ficetola GF, Miaud C, Pompanon F, Taberlet P. 2008.** Species detection using environmental DNA from water samples. *Biology Letters* **4**:423–425 DOI [10.1098/rsbl.2008.0118](https://doi.org/10.1098/rsbl.2008.0118).
- Glenn TC, Nilsen RA, Kieran TJ, Sanders JG, Bayona-Vásquez NJ, Finger JW, Pierson TW, Bentley KE, Hoffberg SL, Louha S, Leon FJG-D, Portilla MA del R, Reed KD, Anderson JL, Meece JK, Aggrey SE, Rekaya R, Alabady M, Belanger M, Winker K, Faircloth BC. 2019.** Adapterama I: universal stubs and primers for 384 unique dual-indexed or 147, 456 combinatorially-indexed Illumina libraries (iTru & iNext). *PeerJ* **7**:e7755 DOI [10.7717/peerj.7755](https://doi.org/10.7717/peerj.7755).
- Goodwin K, Egan K, Greig T, Philibotte J, Koss J, Larsen K, Layton D, Nichols K, O’Neil J, Parks D, Trtanj J, Werner C. 2021.** NOAA ‘omics strategic application of transformational tools strategic plan 2021–2025. Silver Spring: National Oceanic and Atmospheric Administration.
- Grandjean F, Tan MH, Gan HM, Lee YP, Kawai T, Distefano RJ, Blaha M, Roles AJ, Austin CM. 2017.** Rapid recovery of nuclear and mitochondrial genes by genome skimming from Northern Hemisphere freshwater crayfish. *Zoologica Scripta* **46**:718–728 DOI [10.1111/zsc.12247](https://doi.org/10.1111/zsc.12247).
- Gregory TR. 2021.** Animal genome size database. Available at <http://www.genomesize.com/> (accessed on 14 October 2021).
- Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010.** New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59**:307–321 DOI [10.1093/sysbio/syq010](https://doi.org/10.1093/sysbio/syq010).
- Hebert PDN, Cywinska A, Ball SL, DeWaard JR. 2003.** Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B* **270**:313–321 DOI [10.1098/rspb.2002.2218](https://doi.org/10.1098/rspb.2002.2218).
- Hinojosa-Alvarez S, Díaz-Jaimes P, Marcet-Houben M, Gabaldón T. 2015.** The complete mitochondrial genome of the Giant Manta ray, *Manta birostris*. *Mitochondrial DNA* **26**:787–788 DOI [10.3109/19401736.2013.855753](https://doi.org/10.3109/19401736.2013.855753).
- Hoban ML, Williams JT. 2020.** *Cirripectes matatakaro*, a new species of combtooth blenny from the Central Pacific, illuminates the origins of the Hawaiian fish fauna. *PeerJ* **8**:e8852 DOI [10.7717/peerj.8852](https://doi.org/10.7717/peerj.8852).

- Hubert N, Delrieu-Trottin E, Irisson J-O, Meyer C, Planes S. 2010.** Identifying coral reef fish larvae through DNA barcoding: a test case with the families Acanthuridae and Holocentridae. *Molecular Phylogenetics and Evolution* 55:1195–1203 DOI 10.1016/j.ympev.2010.02.023.
- Iwasaki W, Fukunaga T, Isagozawa R, Yamada K, Maeda Y, Satoh TP, Sado T, Mabuchi K, Takeshima H, Miya M, Nishida M. 2013.** MitoFish and MitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. *Molecular Biology and Evolution* 30:2531–2540 DOI 10.1093/molbev/mst141.
- Johnson GD, Paxton JR, Sutton TT, Satoh TP, Sado T, Nishida M, Miya M. 2009.** Deep-sea mystery solved: astonishing larval transformations and extreme sexual dimorphism unite three fish families. *Biology Letters* 5:235–239 DOI 10.1098/rsbl.2008.0722.
- Kane N, Sveinsson S, Dempewolf H, Yang JY, Zhang D, Engels JMM, Cronk Q. 2012.** Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *American Journal of Botany* 99:320–329 DOI 10.3732/ajb.1100570.
- Katoh K, Standley DM. 2013.** MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30:772–780 DOI 10.1093/molbev/mst010.
- Kousteni V, Mazzoleni S, Vasileiadou K, Rovatsos M. 2021.** Complete mitochondrial DNA genome of nine species of sharks and rays and their phylogenetic placement among modern elasmobranchs. *Genes* 12:324 DOI 10.3390/genes12030324.
- Kunze J. 2021.** ARK alliance. Available at <https://arks.org> (accessed on 14 December 2021).
- Lanfear R, Calcott B, Ho SYW, Guindon S. 2012.** PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution* 29:1695–1701 DOI 10.1093/molbev/mss020.
- Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2017.** PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution* 34:772–773 DOI 10.1093/molbev/msw260.
- 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 of the United States of America* 112:2076–2081 DOI 10.1073/pnas.1424997112.
- Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, Boehm JT, Machida RJ. 2013.** A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology* 10:34 DOI 10.1186/1742-9994-10-34.
- Liu B-B, Campbell CS, Hong D-Y, Wen J. 2020.** Phylogenetic relationships and chloroplast capture in the *Amelanchier-Malacomeles-Peraphyllum* clade (Maleae,

- Rosaceae): Evidence from chloroplast genome and nuclear ribosomal DNA data using genome skimming. *Molecular Phylogenetics and Evolution* **147**:106784 DOI [10.1016/j.ympev.2020.106784](https://doi.org/10.1016/j.ympev.2020.106784).
- Malé P-JG, Bardon L, Besnard G, Coissac E, Delsuc F, Engel J, Lhuillier E, Scotti-Saintagne C, Tinaut A, Chave J. 2014.** Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. *Molecular Ecology Resources* **14**:966–975 DOI [10.1111/1755-0998.12246](https://doi.org/10.1111/1755-0998.12246).
- Margaryan A, Noer CL, Richter SR, Restrup ME, Bülow-Hansen JL, Leerhøi F, Langkjær EMR, Gopalakrishnan S, Carøe C, Gilbert MTP, Bohmann K. 2021.** Mitochondrial genomes of Danish vertebrate species generated for the national DNA reference database, DNAmark. *Environmental DNA* **3**:472–480 DOI [10.1002/edn3.138](https://doi.org/10.1002/edn3.138).
- Marko PB, Nance HA, Guynn KD. 2011.** Genetic detection of mislabeled fish from a certified sustainable fishery. *Current Biology* **21**:R621–R622 DOI [10.1016/j.cub.2011.07.006](https://doi.org/10.1016/j.cub.2011.07.006).
- Meyer C, Duffy E, Collins A, Paulay G, Wetzer R. 2021.** The US ocean biocode. *Marine Technology Society Journal* **55**:140–141 DOI [10.4031/MTSJ.55.3.33](https://doi.org/10.4031/MTSJ.55.3.33).
- Miya M, Sato Y, Fukunaga T, Sado T, Poulsen JY, 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**:150088 DOI [10.1098/rsos.150088](https://doi.org/10.1098/rsos.150088).
- Mugnai F, Megléc E, Abbiati M, Bavestrello G, Bertasi F, Bo M, Capa M, Chenuil A, Colangelo MA, De Clerck O, Gutiérrez JM, Lattanzi L, Leduc M, Martin D, Matterson KO, Mikac B, Plaisance L, Ponti M, Riesgo A, Rossi V, Turicchia E, Waeschenbach A, Wangensteen OS, Costantini F. 2021.** Are well-studied marine biodiversity hotspots still blackspots for animal barcoding? *Global Ecology and Conservation* **32**:e01909 DOI [10.1016/j.gecco.2021.e01909](https://doi.org/10.1016/j.gecco.2021.e01909).
- Mundy BC. 2005.** Checklist of the fishes of the Hawaiian archipelago. *Bishop Museum Bulletins in Zoology* **6**:1–706.
- Mundy NI, Winchell CS, Woodruff DS. 1996.** Tandem repeats and heteroplasmy in the mitochondrial DNA control region of the Loggerhead Shrike (*Lanius ludovicianus*). *Journal of Heredity* **87**:21–26 DOI [10.1093/oxfordjournals.jhered.a022948](https://doi.org/10.1093/oxfordjournals.jhered.a022948).
- Pochon X, Bott NJ, Smith KF, Wood SA. 2013.** Evaluating detection limits of next-generation sequencing for the surveillance and monitoring of international marine pests. *PLOS ONE* **8**:e73935 DOI [10.1371/journal.pone.0073935](https://doi.org/10.1371/journal.pone.0073935).
- Poortvliet M, Olsen JL, Croll DA, Bernardi G, Newton K, Kollias S, O’Sullivan J, Fernando D, Stevens G, Galván Magaña F, Seret B, Wintner S, Hoarau G. 2015.** A dated molecular phylogeny of manta and devil rays (Mobulidae) based on mitogenome and nuclear sequences. *Molecular Phylogenetics and Evolution* **83**:72–85 DOI [10.1016/j.ympev.2014.10.012](https://doi.org/10.1016/j.ympev.2014.10.012).

- Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020.** Using SPAdes de novo assembler. *Current Protocols in Bioinformatics* **70**:e102 DOI [10.1002/cpbi.102](https://doi.org/10.1002/cpbi.102).
- R Core Team. 2020.** R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Available at <https://www.r-project.org>.
- Rabosky DL, Chang J, Title PO, Cowman PF, Sallan L, Friedman M, Kaschner K, Garilao C, Near TJ, Coll M, Alfaro ME. 2018.** An inverse latitudinal gradient in speciation rate for marine fishes. *Nature* **559**:392–395 DOI [10.1038/s41586-018-0273-1](https://doi.org/10.1038/s41586-018-0273-1).
- Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018.** Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology* **67**:901–904 DOI [10.1093/sysbio/syy032](https://doi.org/10.1093/sysbio/syy032).
- Randall JE. 2007.** *Reef and shore fishes of the Hawaiian Islands*. Honolulu, Hawai'i: University of Hawai'i Press.
- Ratnasingham S, Hebert PDN. 2007.** BOLD: the barcode of life data system (<http://www.barcodinglife.org>). *Molecular Ecology Notes* **7**:355–364 DOI [10.1111/j.1471-8286.2007.01678.x](https://doi.org/10.1111/j.1471-8286.2007.01678.x).
- Ratnasingham S, Hebert PDN. 2013.** A DNA-based registry for all animal species: the barcode index number (BIN) system. *PLOS ONE* **8**:e66213 DOI [10.1371/journal.pone.0066213](https://doi.org/10.1371/journal.pone.0066213).
- Raupach MJ, Deister F, Villastrigo A, Balke M. 2022.** The complete mitochondrial genomes of *Notiophilus quadripunctatus* Dejean, 1826 and *Omophron limbatum* (Fabricius, 1777): new insights into the mitogenome phylogeny of the Carabidae (Insecta, Coleoptera). *Insect Systematics & Evolution* **1**:1–22 DOI [10.1163/1876312X-bja10027](https://doi.org/10.1163/1876312X-bja10027).
- Revell LJ. 2012.** phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**:217–223 DOI [10.1111/j.2041-210X.2011.00169.x](https://doi.org/10.1111/j.2041-210X.2011.00169.x).
- Riaz T, Shehzad W, Viari A, Pompanon F, Taberlet P, Coissac E. 2011.** ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Research* **39**:e145 DOI [10.1093/nar/gkr732](https://doi.org/10.1093/nar/gkr732).
- Riginos C, Crandall ED, Liggins L, Gaither MR, Ewing RB, Meyer C, Andrews KR, Euclide PT, Titus BM, Therkildsen NO, Salces-Castellano A, Stewart LC, Toonen RJ, Deck J. 2020.** Building a global genomics observatory: using GEOME (the Genomic Observatories Metadatabase) to expedite and improve deposition and retrieval of genetic data and metadata for biodiversity research. *Molecular Ecology Resources* **20**:1458–1469 DOI [10.1111/1755-0998.13269](https://doi.org/10.1111/1755-0998.13269).
- Ripma LA, Simpson MG, Hasenstab-Lehman K. 2014.** Geneious! Simplified genome skimming methods for phylogenetic systematic studies: A case study in Oreocarya (Boraginaceae). *Applications in Plant Sciences* **2**:1400062 DOI [10.3732/apps.1400062](https://doi.org/10.3732/apps.1400062).
- Ronquist F, Teslenko M, Van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012.** MrBayes 3.2: efficient Bayesian

- phylogenetic inference and model choice across a large model space. *Systematic Biology* **61**:539–542 DOI [10.1093/sysbio/sys029](https://doi.org/10.1093/sysbio/sys029).
- Ruppert KM, Kline RJ, Rahman MS. 2019.** Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: a systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation* **17**:e00547 DOI [10.1016/j.gecco.2019.e00547](https://doi.org/10.1016/j.gecco.2019.e00547).
- Schander C, Willassen E. 2005.** What can biological barcoding do for marine biology? *Marine Biology Research* **1**:79–83 DOI [10.1080/17451000510018962](https://doi.org/10.1080/17451000510018962).
- Silva AJ, Hellberg RS. 2021.** Chapter six—DNA-based techniques for seafood species authentication. In: Toldrá F, ed. *Advances in food and nutrition research*. Cambridge, Massachusetts: Academic Press, 207–255 DOI [10.1016/bs.afnr.2020.09.001](https://doi.org/10.1016/bs.afnr.2020.09.001).
- Stamatakis A. 2014.** RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313 DOI [10.1093/bioinformatics/btu033](https://doi.org/10.1093/bioinformatics/btu033).
- Stat M, Huggett MJ, Bernasconi R, DiBattista JD, Berry TE, Newman SJ, Harvey ES, Bunce M. 2017.** Ecosystem biomonitoring with eDNA: Metabarcoding across the tree of life in a tropical marine environment. *Nature Publishing Group* **7**:12240 DOI [10.1038/s41598-017-12501-5](https://doi.org/10.1038/s41598-017-12501-5).
- Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A. 2012.** Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *American Journal of Botany* **99**:349–364 DOI [10.3732/ajb.1100335](https://doi.org/10.3732/ajb.1100335).
- Therkildsen NO, Palumbi SR. 2017.** Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular Ecology Resources* **17**:194–208 DOI [10.1111/1755-0998.12593](https://doi.org/10.1111/1755-0998.12593).
- Timmers MA, Jury CP, Vicente J, Bahr KD, Webb MK, Toonen RJ. 2021.** Biodiversity of coral reef cryptobiota shuffles but does not decline under the combined stressors of ocean warming and acidification. *Proceedings of the National Academy of Sciences of the United States of America* **118**:e2103275118 DOI [10.1073/pnas.2103275118](https://doi.org/10.1073/pnas.2103275118).
- Toczydlowski RH, Liggins L, Gaither MR, Anderson TJ, Barton RL, Berg JT, Beskid SG, Davis B, Delgado A, Farrell E, Ghoojaji M, Himmelsbach N, Holmes AE, Queeno SR, Trinh T, Weyand CA, Bradburd GS, Riginos C, Toonen RJ, Crandall ED. 2021.** Poor data stewardship will hinder global genetic diversity surveillance. *Proceedings of the National Academy of Sciences of the United States of America* **118**:e2107934118 DOI [10.1073/pnas.2107934118](https://doi.org/10.1073/pnas.2107934118).
- Trevisan B, Alcantara DMC, Machado DJ, Marques FPL, Lahr DJG. 2019.** Genome skimming is a low-cost and robust strategy to assemble complete mitochondrial genomes from ethanol preserved specimens in biodiversity studies. *PeerJ* **7**:e7543 DOI [10.7717/peerj.7543](https://doi.org/10.7717/peerj.7543).
- Ward RD, Hanner R, Hebert PDN. 2009.** The campaign to DNA barcode all fishes, FISH-BOL. *Journal of Fish Biology* **74**:329–356 DOI [10.1111/j.1095-8649.2008.02080.x](https://doi.org/10.1111/j.1095-8649.2008.02080.x).

- West KM, Stat M, Harvey ES, Skepper CL, DiBattista JD, Richards ZT, Travers MJ, Newman SJ, Bunce M. 2020.** eDNA metabarcoding survey reveals fine-scale coral reef community variation across a remote, tropical island ecosystem. *Molecular Ecology* **29(6)**:1069–1086 DOI [10.1111/mec.15382](https://doi.org/10.1111/mec.15382).
- White WT, Corrigan S, Yang L, Henderson AC, Bazinet AL, Swofford DL, Naylor GJP. 2018.** Phylogeny of the manta and devilrays (Chondrichthyes: Mobulidae), with an updated taxonomic arrangement for the family. *Zoological Journal of the Linnean Society* **182**:50–75 DOI [10.1093/zoolinnean/zlx018](https://doi.org/10.1093/zoolinnean/zlx018).
- Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. 2017.** ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **8**:28–36 DOI [10.1111/2041-210X.12628](https://doi.org/10.1111/2041-210X.12628).