Skimming for barcodes: rapid production of mitochondrial genome and nuclear ribosomal repeat reference markers through shallow shotgun sequencing (#73270)

First submission

Guidance from your Editor

Please submit by **30 May 2022** for the benefit of the authors (and your $200 publishing discount) .

Structure and Criteria
Please read the 'Structure and Criteria' page for general guidance.

Custom checks
Make sure you include the custom checks shown below, in your review.

Raw data check
Review the raw data.

Image check
Check that figures and images have not been inappropriately manipulated.

Privacy reminder: If uploading an annotated PDF, remove identifiable information to remain anonymous.

Files

Download and review all files from the materials page.

3 Figure file(s)
7 Table file(s)

Custom checks

DNA data checks

- Have you checked the authors data deposition statement?
- Can you access the deposited data?
- Has the data been deposited correctly?
- Is the deposition information noted in the manuscript?

Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- Original primary research within Scope of the journal.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- Impact and novelty not assessed. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- All underlying data have been provided; they are robust, statistically sound, & controlled.
- Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

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

Mykle L Hoban [Corresp., 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, United States

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

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

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

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

Corresponding Author: Mykle L Hoban
Email address: mhoban@hawaii.edu

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

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 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 references ~~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 inpu 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 lo multiple species simultaneously, which has great potential to scale for future projects and facilitate completing barcode reference databases for marine fishes.

Better: "accurate regional reference data sets in global databases" - locally comprehensive, but in a global context to be able to recognise, e.g., new species (rather than mistaking new local occurrence records for new species as only a regional reference database was used).

Introduction

DNA barcoding is a critical component of modern biodiversity research (Hebert et al., 2003; Hebert & Gregory, 2005; Ratnasingham & Hebert, 2007; Hajibabaei et al., 2007), but available barcode reference databases remain incomplete. Thus, it is essential to develop accurate regional reference databases, which support research goals 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). Efforts to characterize community biodiversity patterns through metabarcoding (Leray & Knowlton, 2015; Timmers et al., 2021) and environmental DNA (eDNA) surveys (Ficetola et al., 2008)—which rely on well-curated barcode databases to accurately assign sequences to taxonomy—have expanded dramatically (Ruppert, Kline & Rahman, 2019). To develop more complete DNA barcode databases, we evaluated a method of genome skimming that has potential to rapidly recover multiple barcoding loci for many species simultaneously.

For DNA barcodes to be of long term value, they must be linked to physical (voucher) specimens in permanent natural history collections. This allows for verification of identification and refinements in taxonomy (Schander & Willassen, 2005; Ward, Hanner & Hebert, 2009; but see Collins & Cruickshank, 2013). Another consideration 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, 2012; Bowen et al., 2013). Thus, the most valuable barcode sequences are derived from voucher specimens associated with precise geospatial metadata (geotags), that are unfortunately missing for the majority of 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 must 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).

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). 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). 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 many barcoding markers associated with single voucher specimens also makes species identifications more broadly comparable across studies where researchers may employ different loci.

As high-throughput sequencing has become more accessible and cost-effective, techniques like genome skimming, which uses low-pass, shallow shotgun sequencing of whole genomes, have 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). This work follows and complements that of 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 in Denmark, and extends it by showing that ribosomal barcoding loci are also readily accessible with a genome skimming approach. Despite previous applications of this method, genome skimming has yet to be tested broadly as a method to capture specimen-backed DNA barcodes for marine fishes.

Natural history collections are reservoirs of massive genomic resources that have yet to be fully tapped. While many modern institutions voucher tissue samples and/or DNA extractions alongside collected specimens, they usually publish sequences solely for single barcoding loci

This is a very generalising statement (think of, e.g., target enrichment from collection specimens as counter examples). Any references / examples to support this generalisation?

Natural history collections hold valuable sources of material to support regional or taxon-specific barcode database development, allowing gaps to be filled without the need to collect new specimens. 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. In this process, 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. To evaluate how differences in specimen age and DNA quality affect resulting sequence assemblies, we assessed the quality of sequences and our ability to assemble complete mitogenomes and ribosomal repeats in the context of: (1) taxonomic diversity; (2) DNA extraction method; (3) input DNA concentration; and (4) shearing method. 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 Butterf Ray), as a representative chondrichthyan. All samples were derived from specimens housed the fish collection at NMNH (Table 1) and 10 of the 12 specimens have live color photogra (Fig. 1). No mitogenomes or ribosomal repeats were available in GenBank for any of the sp selected except *Gymnura altavela*, which was published during preparation of this manuscr (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).

It would be important to address in the sample selection section also the question of sample preservation, especially for fish specimens in natural history collections, which are often formalin-fixed. Are the samples in this study from frozen tissue samples preserved for DNA research or from specimens that were formalin-fixed and stored at room temperature for morphological studies? Related to this, how old were these samples? This is a relevant aspect to examine in this type of study, because a comprehensive reference collections requires also the sequencing of old specimens (think singletons), for which no dedicated tissue samples in biorepositories exist.

DNA concentration and extractions

DNA extracts representing a range of concentrations (0.9–34.0 ng/µL) were retrieved from NMNH Biorepository. We did not standardize concentrations prior to library preparation. T for differences in sequencing outcome between extraction methods, we included four sampl extracted with the Qiagen BioSprint DNA blood kit (Qiagen, Inc.; Venlo, Netherlands) and 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 follo 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 impact 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 Illur 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 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 Ger mapper with the sensitivity set to "medium/low" and iterations set to "up to 10 times", start with published COI sequences (Table 1) for each of the 24 libraries. Resulting assemblies w 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 seed

Please add information on the flow cell used - this is critical information to judge how many reads / sample can be expected (which relates to read coverage depth - for cost-saving genome skimming, the coverage has to be as low as possible while still recovering (near) complete mt-genomes for the majority of samples. Depth of coverage affects directly

As you are comparing coverage levels (number of mapped mt-genome reads), have you deduplicated your raw data to eliminate PCR duplicates? Or have you checked with, e.g., FASTQC that duplication levels are that low that they don't matter? Duplication rates are possibly differing between samples that had different quantities of starting material (i.e., higher duplication rates i the complexity of the starting material was low).

This "map to reference" approach only works well if coverage is sufficient for complete mt-genome assembly. This is only typical for samples that contain high-quality DNA (e.g., frozen tissue samples), but not for the majority of natural history collection specimens.

the Map to Reference step repeated until the assemblies stopped increasing in size and iden stretches of sequences were detected at the 5' and 3' ends. The second approach used a con mitogenome from either a congeneric or confamilial taxon as the reference sequence, and M 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.

This equation is rather simplistic and doesn't account for insert sizes (average 200bp) or off-target reads (exogenous DNA). PE150 reads could cover up to 300bp, but don't with 200bp inserts due to on average 100bp overlap. If this has an effect on the estimate depends on wether clusters (1 read = 150bp of 200bp) or read numbers (2 reads / cluster = 300bp assumed, but only 200bp available) are used in the estimate.

Genome sequencing coverage estimation

We estimated species genome sizes (Table 1) based on data available in GenBank or the Ar 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 calculated sequencing coverage estimates (C) for each sample using the equation $C = LN/G$, where L was the sequencing rea 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.

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_MiSeq_01 (https://n2t.net/ark:/21547/DyW2). We deposited BioSample records, annotated mitogenome and ribosomal repeat assemblies, and raw reads in GenBank (BioProject Accession: PRJNA720393).

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/µL in concentration, with mechanically-sheared and enzymatically-sheared libraries averaging 0.71 ± 0.67 ng/µL (mean ± sd) and 1.72 ± 0.94 ng/µL, respectively. The average total library size ranged from 318 to 392 bp, with mechanically-sheared 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.

Sequencing results

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 millio Qiagen). Enzymatic shearing yielded more reads per library than mechanical shearing (2.9 ± million reads for enzymatic vs. 1.8 ± 0.6 million reads for mechanical). Based on estimated genome sizes, these read counts equate to 0.07× to 1.04× genome coverage, with enzymatic shearing (0.50 ± 0.30×) averaging higher than mechanical shearing (0.30 ± 0.19×). A summ of sequencing results across libraries is presented in Table 3.

Why do you measure the performance of DNA extraction approaches via read counts rather than DNA yield and length? Using read counts introduces the additional variable of library building (e.g., mechanical vs enzymatic shearing, differences caused by bead clean-ups) which overlay the variable of DNA extraction approach. You are effectively changing multiple variables at the same time and compare them to the same outcome - this doesn't let you correlate each variable independently to the outcome, as done in the interpretation. If you want to investigate the effect of variables, always change only one variable at a time.

Assembly and sequencing 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 this 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 (19,022 bp in our assembly as compared to 19,472 bp in MT274571) for the Spiny Butterfly Ray.

Have you considered exogenous DNA and shorter DNA fragments?

Mitogenome coverage of the 22 successful assemblies ranged from 7× to 108× (34 ± 26×; 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× with both libraries combined). Across all libraries, assembled mitogenome reads comprised 0.05% to 0.32% (0.17 ± 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 (*Canthigaster amboinensis*), 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 ± 2.3%).

Discussion

Our results show that genome skimming is an efficient method for generating mitogenomes ribosomal repeats of marine fishes and that the methods are robust for a broad range of taxa extraction types, shearing methods, and DNA concentrations. Both manual (Qiagen) and automated (AutoGen) extraction methods resulted in high quality sequence libraries, which indicates that this method can leverage existing DNA extractions housed in museum collect 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 en the ligation period for an additional 45 minutes compared to those with mechanical shearing This likely impacted their ligation efficiency and subsequent library yield. As a result, we c confirm that differences in final library yield resulted directly from the shearing method use However, enzymatically sheared libraries yielded higher read counts than mechanically she with greater average mitogenome coverage, demonstrating that the method is effective for genome skimming. In addition, enzymatic shearing is 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

If mt-genomes / rRNA are the target of genome skimming, there is no striking difference between mechanical and enzymatic shearing as coverage is high and base compositional bias is similar across the mt-genome. If you want to investigate differences in coverage, you need to look at the nuclear data, where biases in enzymatic shearing patterns can have an effect. Thus, if genome skimming is carried out for low-coverage resequencing of nuclear genomes, the method of shearing might make a difference to the coverage or recovery.

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).

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. However, 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). Future studies will include additional sharks, rays, and chimaeras, as well as further exploration of laboratory and bioinformatic approaches.

We used the MiSeq platform to test extraction and shearing methods on a limited number of samples and to assess sequencing reads and coverage necessary to generate mitogenomes and ribosomal repeats across a broad phylogenetic sample 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 data yield from 2–4 million reads to 13 million per sample on average. 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 shows 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 use of 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 registered with 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 U.S. Ocean Biocode (Meyer et al., 2021), each of which involve explicit aims to provide complete DNA barcode reference databases based on voucher specimens housed in museum collections. Techniques and methods developed here are applicable to taxa and regions beyond marine fishes and the Hawaiian Islands. Taxonomically 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 (SI) initiative to survey the marine fishes of French Polynesia, including the Mo’orea Biocode Project. We thank CRIOBE and SI, in particular Jeffrey Williams (SI) 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‘ohe Bay. We thank the Smithsonian Conservation Biology Institute, the National Museum of Natural History, Smithsonian Institution, 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. 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. 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 (contribution no. X), and NMNH Laboratories of Analytical Biology. This is contribution #XXX from the Hawai'i Institute of Marine Biology and #XXX from the School of Ocean and Earth Science and Technology at the University of Hawai'i.

References

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, Denoeud F, Pitelkova I, Pușcaș 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

401 from Madagascar. *Annals of Botany* 112:1057–1066. DOI: 10.1093/aob/mct174.
402 Bowen BW, Rocha LA, Toonen RJ, Karl SA, ToBo Laboratory. 2013. The origins of tropical
403 marine biodiversity. *Trends in Ecology & Evolution* 28:359–366. DOI:
404 10.1016/j.tree.2013.01.018.
405 Carpenter KE, Williams JT, Santos MD. 2017. *Acanthurus albimento*, a new species of
406 surgeonfish (Acanthuriformes: Acanthuridae) from northeastern Luzon, Philippines, with
407 comments on zoogeography. *Journal of the Ocean Science Foundation* 25. DOI:
408 10.5281/zenodo.291792.
409 Coissac E, Hollingsworth PM, Lavergne S, Taberlet P. 2016. From barcodes to genomes:
410 extending the concept of DNA barcoding. *Molecular Ecology* 25:1423–1428. DOI:
411 10.1111/mec.13549.
412 Collins RA, Cruickshank RH. 2013. The seven deadly sins of DNA barcoding. *Molecular*
413 *Ecology Resources* 13:969–975. DOI: 10.1111/1755-0998.12046.
414 Davies N, Deck J, Kansa EC, Kansa SW, Kunze J, Meyer C, Orrell T, Ramdeen S, Snyder R,
415 Vieglais D, Walls RL, Lehnert K. 2021. Internet of Samples (iSamples): Toward an
416 interdisciplinary cyberinfrastructure for material samples. *GigaScience* 10:giab028. DOI:
417 10.1093/gigascience/giab028.
418 Denver DR, Brown AMV, Howe DK, Peetz AB, Zasada IA. 2016. Genome skimming: A rapid
419 approach to gaining diverse biological insights into multicellular pathogens. *PLOS*
420 *Pathogens* 12:e1005713. DOI: 10.1371/journal.ppat.1005713.
421 DiBattista JD, Rocha LA, Craig MT, Feldheim KA, Bowen BW. 2012. Phylogeography of two
422 closely related Indo-Pacific butterflyfishes reveals divergent evolutionary histories and
423 discordant results from mtDNA and microsatellites. *Journal of Heredity* 103:617–629.

DOI: 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.

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.

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.

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.

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, MD: 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.

447 Gregory TR. 2021.Animal Genome Size Database. *Available at http://www.genomesize.com/*
448 (accessed October 14, 2021).
449 Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA. 2007. DNA barcoding: how it
450 complements taxonomy, molecular phylogenetics and population genetics. *Trends in*
451 *Genetics* 23:167–172. DOI: 10.1016/j.tig.2007.02.001.
452 Hebert PDN, Cywinska A, Ball SL, deWaard JR. 2003. Biological identifications through DNA
453 barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*
454 270:313–321. DOI: 10.1098/rspb.2002.2218.
455 Hebert PDN, Gregory TR. 2005. The promise of DNA barcoding for taxonomy. *Systematic*
456 *Biology* 54:852–859. DOI: 10.1080/10635150500354886.
457 Hinojosa-Alvarez S, Díaz-Jaimes P, Marcet-Houben M, Gabaldón T. 2015. The complete
458 mitochondrial genome of the Giant Manta ray, *Manta birostris*. *Mitochondrial DNA*
459 26:787–788. DOI: 10.3109/19401736.2013.855753.
460 Hoban ML, Williams JT. 2020. *Cirripectes matatakaro*, a new species of combtooth blenny from
461 the Central Pacific, illuminates the origins of the Hawaiian fish fauna. *PeerJ* 8:e8852.
462 DOI: 10.7717/peerj.8852.
463 Hubert N, Delrieu-Trottin E, Irisson J-O, Meyer C, Planes S. 2010. Identifying coral reef fish
464 larvae through DNA barcoding: A test case with the families Acanthuridae and
465 Holocentridae. *Molecular Phylogenetics and Evolution* 55:1195–1203. DOI:
466 10.1016/j.ympev.2010.02.023.
467 Iwasaki W, Fukunaga T, Isagozawa R, Yamada K, Maeda Y, Satoh TP, Sado T, Mabuchi K,
468 Takeshima H, Miya M, Nishida M. 2013. MitoFish and MitoAnnotator: a mitochondrial
469 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.

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 December 14, 2021).

Leray M, Knowlton N. 2015. DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences* 112:2076–2081. 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.

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.

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.

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.

Meyer C, Duffy E, Collins A, Paulay G, Wetzer R. 2021. The U.S. Ocean Biocode. *Marine Technology Society Journal* 55:140–141. DOI: 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.

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.

516 Pochon X, Bott NJ, Smith KF, Wood SA. 2013. Evaluating detection limits of next-generation
517 sequencing for the surveillance and monitoring of international marine pests. *PLOS ONE*
518 8:e73935. DOI: 10.1371/journal.pone.0073935.
519 Poortvliet M, Olsen JL, Croll DA, Bernardi G, Newton K, Kollias S, O'Sullivan J, Fernando D,
520 Stevens G, Galván Magaña F, Seret B, Wintner S, Hoarau G. 2015. A dated molecular
521 phylogeny of manta and devil rays (Mobulidae) based on mitogenome and nuclear
522 sequences. *Molecular Phylogenetics and Evolution* 83:72–85. DOI:
523 10.1016/j.ympev.2014.10.012.
524 Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020. Using SPAdes de
525 novo assembler. *Current Protocols in Bioinformatics* 70:e102. DOI: 10.1002/cpbi.102.
526 Randall JE. 2007. *Reef and shore fishes of the Hawaiian Islands*. University of Hawai'i Press.
527 Ratnasingham S, Hebert PDN. 2007. BOLD: The Barcode of Life Data System
528 (http://www.barcodinglife.org). *Molecular Ecology Notes* 7:355–364. DOI:
529 10.1111/j.1471-8286.2007.01678.x.
530 Raupach MJ, Deister F, Villastrigo A, Balke M. 2022. The complete mitochondrial genomes of
531 *Notiophilus quadripunctatus* Dejean, 1826 and *Omophron limbatum* (Fabricius, 1777):
532 New insights into the mitogenome phylogeny of the Carabidae (Insecta, Coleoptera).
533 *Insect Systematics & Evolution* 1:1–22. DOI: 10.1163/1876312X-bja10027.
534 Riaz T, Shehzad W, Viari A, Pompanon F, Taberlet P, Coissac E. 2011. ecoPrimers: inference of
535 new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids*
536 *Research* 39:e145. DOI: 10.1093/nar/gkr732.
537 Riginos C, Crandall ED, Liggins L, Gaither MR, Ewing RB, Meyer C, Andrews KR, Euclide PT,
538 Titus BM, Therkildsen NO, Salces-Castellano A, Stewart LC, Toonen RJ, Deck J. 2020.

539 Building a global genomics observatory: Using GEOME (the Genomic Observatories
540 Metadatabase) to expedite and improve deposition and retrieval of genetic data and
541 metadata for biodiversity research. *Molecular Ecology Resources* 20:1458–1469. DOI:
542 10.1111/1755-0998.13269.
543 Ripma LA, Simpson MG, Hasenstab-Lehman K. 2014. Geneious! Simplified genome skimming
544 methods for phylogenetic systematic studies: A case study in Oreocarya (Boraginaceae).
545 *Applications in Plant Sciences* 2:1400062. DOI: 10.3732/apps.1400062.
546 Ruppert KM, Kline RJ, Rahman MS. 2019. Past, present, and future perspectives of
547 environmental DNA (eDNA) metabarcoding: A systematic review in methods,
548 monitoring, and applications of global eDNA. *Global Ecology and Conservation*
549 17:e00547. DOI: 10.1016/j.gecco.2019.e00547.
550 Schander C, Willassen E. 2005. What can biological barcoding do for marine biology? *Marine*
551 *Biology Research* 1:79–83. DOI: 10.1080/17451000510018962.
552 Silva AJ, Hellberg RS. 2021. Chapter Six - DNA-based techniques for seafood species
553 authentication. In: Toldrá F ed. *Advances in Food and Nutrition Research*. Academic
554 Press, 207–255. DOI: 10.1016/bs.afnr.2020.09.001.
555 Stat M, Huggett MJ, Bernasconi R, DiBattista JD, Berry TE, Newman SJ, Harvey ES, Bunce M.
556 2017. Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a
557 tropical marine environment. *Nature Publishing Group* 7:12240. DOI: 10.1038/s41598-
558 017-12501-5.
559 Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A. 2012. Navigating the tip
560 of the genomic iceberg: Next-generation sequencing for plant systematics. *American*
561 *Journal of Botany* 99:349–364. DOI: 10.3732/ajb.1100335.

562 Therkildsen NO, Palumbi SR. 2017. Practical low-coverage genomewide sequencing of
563 hundreds of individually barcoded samples for population and evolutionary genomics in
564 nonmodel species. *Molecular Ecology Resources* 17:194–208. DOI: 10.1111/1755-
565 0998.12593.
566 Timmers MA, Jury CP, Vicente J, Bahr KD, Webb MK, Toonen RJ. 2021. Biodiversity of coral
567 reef cryptobiota shuffles but does not decline under the combined stressors of ocean
568 warming and acidification. *Proceedings of the National Academy of Sciences*
569 118:e2103275118. DOI: 10.1073/pnas.2103275118.
570 Toczydlowski RH, Liggins L, Gaither MR, Anderson TJ, Barton RL, Berg JT, Beskid SG, Davis
571 B, Delgado A, Farrell E, Ghoojaei M, Himmelsbach N, Holmes AE, Queeno SR, Trinh T,
572 Weyand CA, Bradburd GS, Riginos C, Toonen RJ, Crandall ED. 2021. Poor data
573 stewardship will hinder global genetic diversity surveillance. *Proceedings of the National*
574 *Academy of Sciences* 118:e2107934118. DOI: 10.1073/pnas.2107934118.
575 Trevisan B, Alcantara DMC, Machado DJ, Marques FPL, Lahr DJG. 2019. Genome skimming is
576 a low-cost and robust strategy to assemble complete mitochondrial genomes from ethanol
577 preserved specimens in biodiversity studies. *PeerJ* 7:e7543. DOI: 10.7717/peerj.7543.
578 Ward RD, Hanner R, Hebert PDN. 2009. The campaign to DNA barcode all fishes, FISH-BOL.
579 *Journal of Fish Biology* 74:329–356. DOI: 10.1111/j.1095-8649.2008.02080.x.
580 West KM, Stat M, Harvey ES, Skepper CL, DiBattista JD, Richards ZT, Travers MJ, Newman
581 SJ, Bunce M. 2020. eDNA metabarcoding survey reveals fine-scale coral reef community
582 variation across a remote, tropical island ecosystem. *Molecular Ecology* 30:246. DOI:
583 10.1111/mec.15382.
584 White WT, Corrigan S, Yang L, Henderson AC, Bazinet AL, Swofford DL, Naylor GJP. 2018.

585 Phylogeny of the manta and devilrays (Chondrichthyes: mobulidae), with an updated
586 taxonomic arrangement for the family. *Zoological Journal of the Linnean Society*
587 182:50–75. DOI: 10.1093/zoolinnean/zlx018.

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.

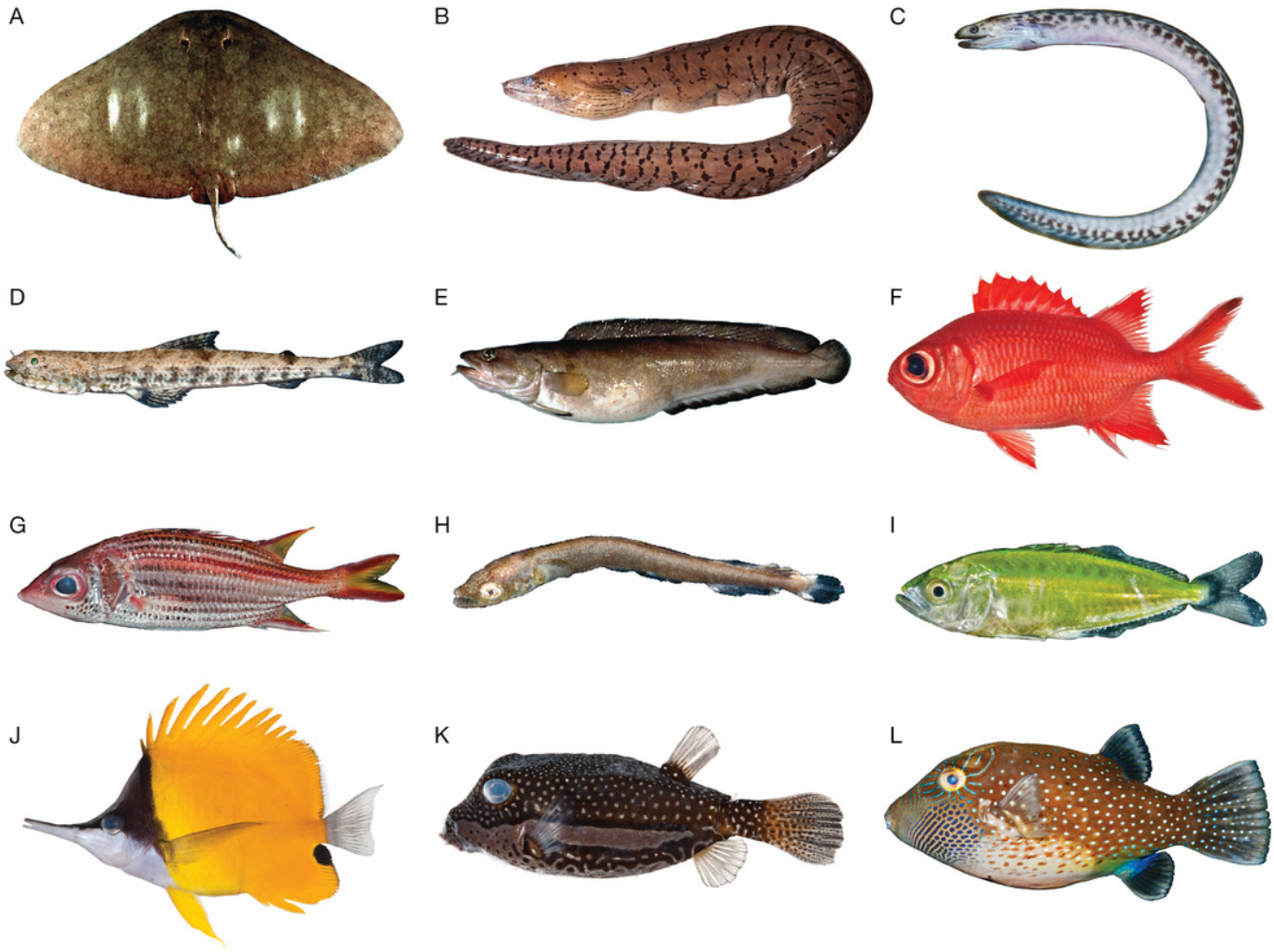
A
B
C
D
E
F
G
H
I
J
K
L

Figure 2

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

Photograph by Diane Pitassy, NMNH.

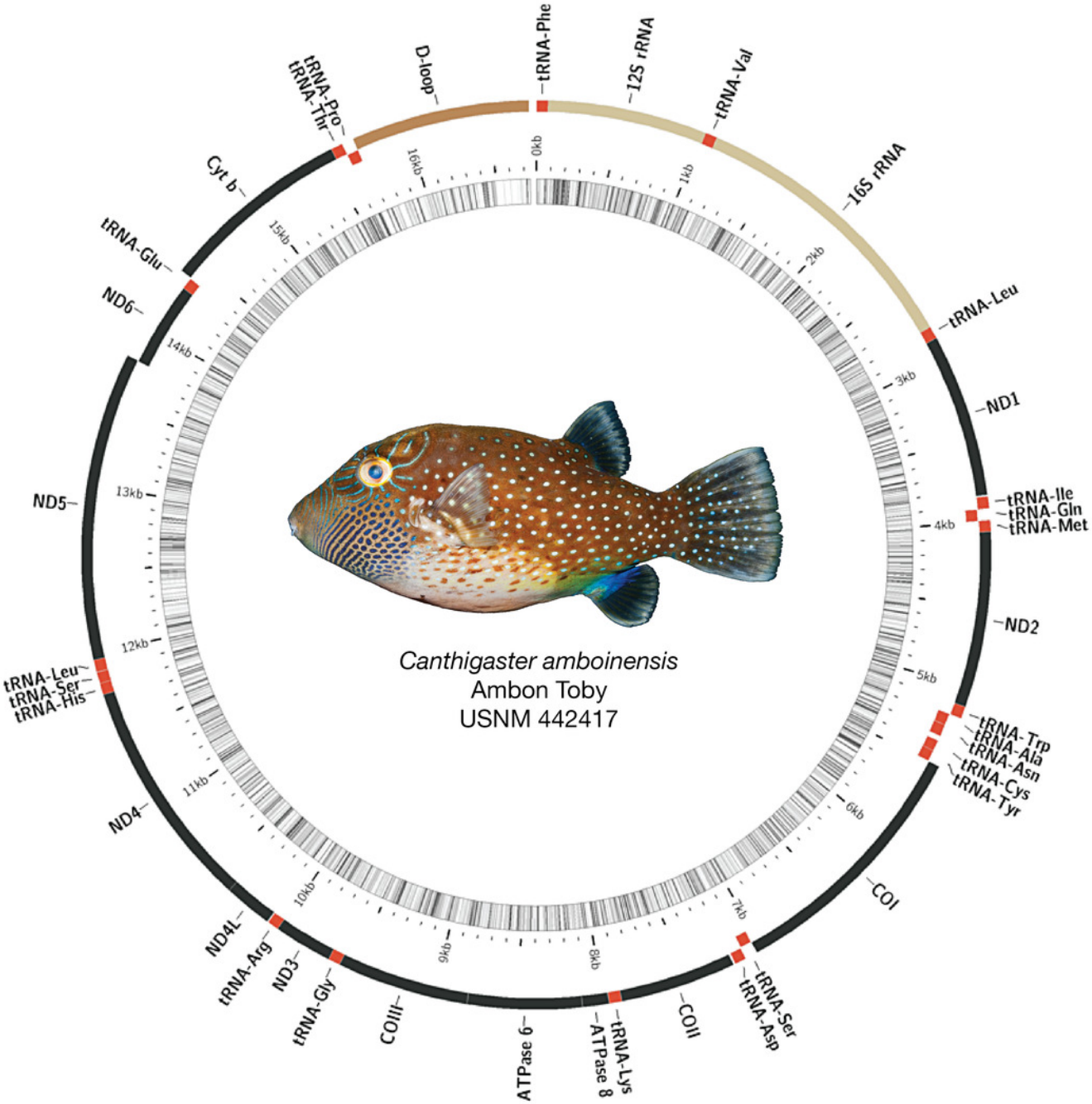
tRNA-Phe
12S rRNA
tRNA-Val
16S rRNA
tRNA-Leu
ND1
tRNA-Ile
tRNA-Gln
tRNA-Met
ND2
tRNA-Trp
tRNA-Ala
tRNA-Asn
tRNA-Cys
tRNA-Tyr
COI
tRNA-Ser
tRNA-Asp
COII
tRNA-Lys
ATPase 8
ATPase 6
COIII
tRNA-Gly
ND3
tRNA-Arg
ND4L
ND4
tRNA-His
tRNA-Ser
tRNA-Leu
ND5
ND6
tRNA-Glu
Cyt b
tRNA-Thr
tRNA-Pro
D-loop
0kb
1kb
2kb
3kb
4kb
5kb
6kb
7kb
8kb
9kb
10kb
11kb
12kb
13kb
14kb
15kb
16kb
Canthigaster amboinensis
Ambon Toby
USNM 442417

Table 1(on next page)

Summary of species and museum specimens included in this study. Species in this and subsequent tables are arranged by taxonomic order, family, and scientific name, with the chondrichthyan presented separately.

~~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	COI reference accession
Gymnura altavela (Linnaeus, 1758)	Myliobatiformes	Gymnuridae	AutoGen	1.80[b]	433343	MH378654
Gymnothorax fimbriatus (Bennett, 1832)	Anguilliformes	Muraenidae	BioSprint	2.31[c]	395396	MK658634
Gymnothorax undulatus (Lacepède, 1803)	Anguilliformes	Muraenidae	AutoGen	2.31[c]	442319	MG816692
Saurida nebulosa Valenciennes, 1850	Aulopiformes	Synodontidae	AutoGen	1.53[c]	442473	MG816726
Tylosurus crocodilus (Péron & Lesueur, 1821)	Beloniformes	Belonidae	AutoGen	1.00[s]	442362	MG816741
Myripristis vittata Valenciennes, 1831	Beryciformes	Holocentridae	BioSprint	0.90[c]	411102	MZ598162
Neoniphon sammara (Forsskål, 1775)	Beryciformes	Holocentridae	AutoGen	0.80[s]	442483	MG816708
Brosme brosme (Ascanius, 1772)	Gadiformes	Lotidae	AutoGen	0.41[s]	433199	MH378533
Scomberoides lysan (Forsskål, 1775)	Perciformes	Carangidae	AutoGen	0.73[c]	442297	MG816730
Forcipiger flavissimus Jordan & McGregor, 1898	Perciformes	Chaetodontidae	BioSprint	0.72[s]	411089	MK657435
Ostracion whitleyi Fowler, 1931	Tetraodontiformes	Ostraciidae	BioSprint	0.98[c]	411029	MK658705

Canthigaster amboinensis (Bleeker, 1864)	Tetraodontiformes	Tetraodontidae	AutoGen	0.41[c]	442417	MG816661

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

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

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

Table 2(on next page)

Accession numbers of 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

[a] based on nearly-complete mitogenome assembly

Table 3(on next page)

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

1