scientific data

OPEN

DATA DESCRIPTOR

Australia's marine fishes DNA barcode reference library for integrated taxonomy, metabarcoding & eDNA research

Sharon A. Appleyard[1]✉, Robert D. Ward[1], John J. Pogonoski[1], Alastair Graham[1], Peter R. Last[1], Bruce E. Deagle[1], Bronwyn Holmes[2], Martin F. Gomon[3], Dianne J. Bray[3], Jeffrey W. Johnson[4], Amanda C. Hay[5], Glenn I. Moore[6,7], Michael P. Hammer[8], Barry Russell[8] & Ken J. Graham[5]

Over 15 000 species of fishes are found globally in the marine environment and DNA barcodes are used extensively to describe, catalogue, understand and manage this diversity. The dataset outlined here represents a DNA barcode reference library of the mitochondrial cytochrome c oxidase subunit 1 gene (COI) from 9767 voucher specimens (representing at least 2220 species and 288 families) of marine fishes. This publicly available dataset in the Barcode of Life Data System (BOLD) represents 17 years (2005–2022) of barcoding of marine fishes identified from Australian territorial waters. Tissues targeted for sequencing with their matching physical specimens (and extracted DNA), obtained via a multi-agency sampling effort, are mostly maintained and curated by the CSIRO Australian National Fish Collection (ANFC) in Hobart, Australia. Species-level integrated taxonomy (assigned after combined morphological and genetic assessment) has been determined for 91% of the dataset. The library represents the most complete COI barcode reference dataset for marine fishes from Australian waters and is currently utilised for integrated taxonomy, (meta)barcoding and eDNA studies.

Background & Summary

In 2003, Hebert *et al.* proposed the use of the mitochondrial DNA gene cytochrome c oxidase subunit 1 (COI) as a global bio-identification (barcode) system, with a 650 base pair region of the COI nominated for species identifications in varied animal lineages[1,2]. More recently, incorrect species identification in public sequence databases such as GenBank is a growing problem[3–5]. The success therefore of DNA barcoding relies on robust and actively curated barcode reference libraries (and matching vouchers) that enable source DNA sequences to be compared to previously identified taxa (i.e. with known *a priori* taxonomic Linnean structure)[6–12]. Furthermore, taxonomic verification of published sequence data by referencing voucher specimens is essential, and DNA barcoding provides a good opportunity to check the accuracy of species identification[13]. In fishes, COI DNA barcoding is used extensively by taxonomists, geneticists and fisheries scientists for identification and discovery of taxa, independent testing of taxonomic systems, delineation of species boundaries, assessment of genetic diversity, and for forensic identification of samples[6,7,13–18]. While sequencing capacity enables detailed insights from whole genome initiatives (e.g. Earth Biogenome Project; Genome 10K), DNA barcoding libraries still remain highly relevant due to their scalability and relative simplicity for species-level identifications[11,12]. In fishes at least, COI is an excellent species level marker, which may resolve more than 97% of species boundaries[17]. There has also

[1]CSIRO Australian National Fish Collection, National Research Collections Australia, Hobart, TAS, 7000, Australia. [2]CSIRO Environment, Hobart, TAS, 7000, Australia. [3]Museum Victoria, Melbourne, VIC, 3000, Australia. [4]Queensland Museum, Collections and Research Centre, Hendra, QLD, 4011, Australia. [5]Australia Museum, Ichthyology, Sydney, NSW, 2010, Australia. [6]Collections and Research, Western Australian Museum, Welshpool, WA, 6106, Australia. [7]School of Biological Sciences, University of Western Australia, Nedlands, WA, 6009, Australia. [8]Museum and Art Gallery of the Northern Territory, Darwin, NT, 0810, Australia. ✉e-mail: Sharon.Appleyard@csiro.au

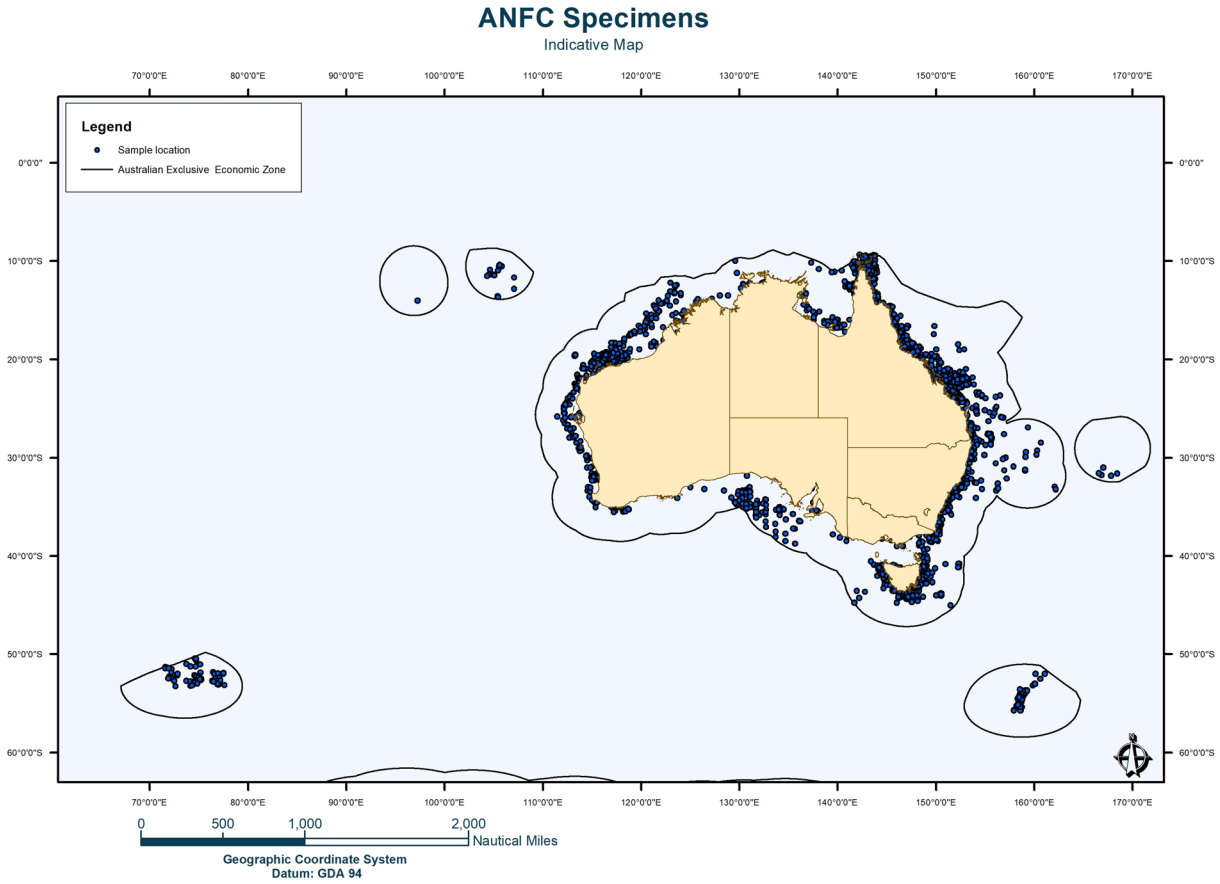

Fig. 1 Locations across the Australian Exclusive Economic Zone territorial waters from which ANFC specimens with latitudinal and longitudinal metadata were sampled.

been a rapid rise in the range of applications utilising DNA barcoding, with an increase in environmental DNA (eDNA) analyses and bulked tissue or individual typing via metabarcoding using short read sequencing[9–11,13].

We present here our reference sequence library of COI barcodes with sampling and location metadata for Australian marine sourced fishes (classes Actinopterygii, Elasmobranchii, Holocephalii, Myxini, Cephalaspidomorphi); the dataset is deposited in the Barcode of Life Data System (BOLD). With over 5000 fish species known in the Australian marine estate (and approximately 25% of these are endemic)[19] this reference dataset builds on an earlier paper[17], that presented 754 COI sequences from 207 species of mostly Australian marine fishes and complements an earlier Lizard Island paper[20] on Australian coral reef fishes. The current reference dataset (with >9700 sequences and 2200 species) incorporates work undertaken over the last 17 years (2005–2022) and represents the most extensive collection of Australian vouchered and barcoded marine fishes. The COI barcodes in this reference sequence library are linked with vouchered specimens and images that have been expertly identified by Australian ichthyologists, taxonomic experts and collaborators from the CSIRO Australian National Fish Collection (ANFC), the Australian Museum, Museums Victoria, Queensland Museum, Museum and Art Gallery of the Northern Territory, the Western Australian Museum and the South Australian Museum. Importantly, the use of paired sequenced tissues and matching vouchers allows a feedback loop to verify and inform identification and acts as a strong quality assurance and control mechanism. As a national Australian based fish collection, the ANFC developed the COI reference sequence libraries for Australian and regional marine fishes (with responsibility for the COI barcoding and managing the data). Importantly, the network of Australian collections outlined here, were critical to advance the taxonomy of Australian fishes.

The sampling for marine fishes in this dataset was supported through CSIRO and our collaborators research efforts, primarily through collections on research voyages plus ships of opportunity, public donations, state-based field surveys and observer sampling onboard commercial fishing vessels. The dataset described here consists of 9767 specimens that have been taxonomically identified (with 91% identified to species level), mostly fixed and preserved, tissue sampled, and DNA barcoded with images taken of representatives of most taxa. There are at least 2220 species in the dataset representing 1010 genera and 288 families. More than 70% of species in the dataset are represented by multiple specimens, and 93% of these have associated geographic (i.e. latitude and longitude) metadata (Fig. 1). As most are curated in the ANFC, specimen metadata beyond that which are outlined in the BOLD dataset (e.g. depth, habitat) is also maintained in the ANFC's collection management Specify platform. Additionally, while most specimens were collected from open marine waters, we recognise the need for ongoing foci on rocky and coral reef environments and offshore waters to assist with addressing sampling gaps in the dataset, thereby helping to monitor changes in the marine estate.

Fig. 2 Overview of Australian COI reference sequence library data generation from the ANFC.

Our aim in publishing this publicly available dataset is to make this resource more widely known and used by the Australian and international ichthyology community. It will be of interest to researchers working in the Australian region, including those studying fish taxonomy and ecology, researchers undertaking barcoding and metabarcoding (of bulked samples such as fish eggs and larvae) and practitioners of marine community eDNA analyses. The dataset is Australian focused; however, it will also be of interest and have broad utility as an exemplar COI barcoding dataset of marine fishes at a regional and ocean-basin scale.

Methods

The curated reference sequence library for Australian marine fishes was developed over the last 17 years by ANFC taxonomists, ichthyologists and geneticists and their taxonomic collaborators from six Australian museums. A number of sampling, wet laboratory and molecular processing steps has enabled this publicly available resource (Fig. 2).

Sampling and specimen collection. All historical and contemporary specimen collecting was done in accordance with Australian Commonwealth and State regulations, with animal ethics approvals in place where required. The ANFC and CSIRO does not require permits to cover specimens and samples from routine commercial fishing operations nor for public donations of specimens if the material was legally collected using necessary licences when required.

Specimens were obtained from voyages, research projects, public donations and commercial and recreational fishing activities from the late 1980s through to 2022. All were collected within Australia's Exclusive Economic Zone, the world's third largest EEZ (see https://atlas.parksaustralia.gov.au/amps/underpinning-science). Most specimens were collected from 1994 onwards. The Australian EEZ is characterised by a lengthy latitudinal range including tropical, temperate, sub-Antarctic and Antarctic waters with high habitat diversity (see https://soe.dcceew.gov.au/). Many specimens were obtained during offshore expeditions on RV *Southern Surveyor* and more recently RV *Investigator* (Australia's national marine research vessel) voyages, including Great Australian Bight 2015 voyage IN2015_C02, Sampling the Abyss 2017 IN2017_V03, North West Shelf 2017 IN2017_V05 and Tasmanian seamounts 2018 IN2018_VO6.

Fishes were collected via a variety of methods and gear including beam trawls (in 30–60 minute tows, deployable to approximately 5500 m), demersal trawls (usually 15–30 minute tows, deployable to >2000 m), midwater trawls (sampling to at least 1000 m), benthic sleds, longlining, rotenone, seining, SCUBA, beach washups etc. Fishes were collected at a range of depths from 0–5500 m although depths >2000 m remain sparsely sampled. Descriptions of research surveying methodologies and specimen acquisition are presented in relevant literature[21,22]. Generally, where possible, fish specimens are identified to the lowest possible taxonomic level, photographed and tissue samples removed and frozen at −20 °C or −80 °C (depending on availability of low-temperature facilities). Depending on resources, specimens may be fixed at sea in 10% formalin and then stepped up and preserved in 70% ethanol when registered into the ANFC and/ or another Australian State based museum collection. Alternatively, whole frozen specimens from research voyages were later processed and preserved ashore in the ANFC.

Primers	Direction	Reference	Sequence
FishF1	Forward	[17]	TCAACCAACCACAAAGACATTGGCAC
FishF2	Forward	[17]	TCGACTAATCATAAAGATATCGGCAC
FishR1	Reverse	[17]	TAGACTTCTGGGTGGCCAAAGAATCA
FishR2	Reverse	[17]	ACTTCAGGGTGACCGAAGAATCAGAA
BCL	Forward	[25]	TCAACYAATCAYAAAGATATYGGCAC
BCH	Reverse	[25]	ACTTCYGGGTGRCCRAARAATCA

Table 1. Primers used in mtDNA COI barcoding for ANFC marine fish specimens.

Where fish have been part of specific research projects, whole specimens are identified in the field to the lowest taxonomic level, frozen at −20 °C and shipped to the ANFC in Hobart, where specimens are maintained at −20 °C until taxonomic assessment and tissue sampling are completed. Irrespective of how specimens are acquired, representative vouchered specimens are imaged, fixed in 10% formalin and stepped up into 70% ethanol for long term preservation. Contemporary tissue samples (i.e. muscle tissue, fin clips) are taken from vouchered specimens before the specimens are fixed in formalin. These tissues may be dissected from fresh, frozen (−20 °C or −80 °C), or ethanol preserved specimens and then sub-sampled for DNA extractions. Remaining tissues are stored at −80 °C in the ANFC.

Molecular laboratory processing. The barcode processing research was previously managed by R.D.W who established the Australian node of FISH-BOL[18], with tissue samples extracted with Chelex (Merck, USA) according to in-house protocols and largely sequenced at the University of Guelph (Canada).

Since 2016, DNA extractions, amplification of the COI gene and Sanger sequencing and barcoding have been managed by the senior author (S.A.A), with samples extracted at the CSIRO marine laboratories. Tissue amounts of 0.01–0.025 gm are extracted with silica binding plates (Wizard®SV 96 Genomic DNA Purification System, Promega, Australia) and plated into 96 well plates[14,23,24]. Resultant DNA ranges from 2 - > 100 ng/*u*l (average approximately 5–8 ng/*u*l), depending on the amount of starting tissue. Archival DNA has been stored at the CSIRO marine laboratories in plates at −80 °C for specimens processed since 2016. For pre-2016 samples, DNA can mostly be extracted from tissues that have been stored frozen (−80 °C & /−20 °C) at the marine laboratories.

Two sets of COI primers (Table 1) have generally been used to amplify the same COI region. Prior to 2016, most DNA samples were amplified and bi-directionally sequenced at the Canadian Centre for DNA Barcoding (Guelph, Canada) using FishF1 or FishF2 and FishR1 or FishR2 primers[17], with sequences uploaded directly by BOLD data managers. Since 2016, amplification and sequencing have been achieved with an equal molar combination of both forward primers FishF1&F2 and the FishR2 primer in the first instance (annealing temperature of 54 °C), with BCH and BCL primers[25] (annealing temperature of 50 °C) particularly for COI amplification in Elasmobranchii taxa and/or in teleost taxa that did not amplify with the FishF1, F2, R1 primers[17]. Sequencing then either occurred at the CSIRO marine laboratories (Hobart, Australia) or at the Ramaciotti Centre for Genomics (Sydney, Australia) on Applied Biosystems Sanger 3130XL and 3730 sequencers.

Since 2016, forward and reverse sequences were trimmed, *de novo* assembled and manually checked by eye for base pair calling accuracy before being converted into consensus sequences using various iterations of Geneious (currently Geneious Prime 2021.2.2; Biomatters Ltd., Auckland, New Zealand). Resulting consensus sequences for each sample were compared using the BOLD Identification Engine tool to check the similarity against existing database sequences. Identification was based on the percentage of sequence identity, with identity of ≥98%[26] as the primary criterion used here for genetic specimen identification. Concurrently, taxonomic identification of specimens based on the COI outcome was also checked by ANFC ichthyologists and collaborator taxonomists; BOLD also maintains records of taxonomic updates for each specimen. The integration of the two approaches (referred to here as integrated taxonomy) is used for species identification. Since 2016, sequence and metadata for each specimen has been backed up at CSIRO on cloud-based storage facilities.

Ongoing specimen identification and curation. All morphological identifications were made by the authors or collaborators at time of specimen registration into the ANFC. Using integrated taxonomy, the ANFC and BOLD data managers continually revise and curate the specimen and sequence information in this dataset and more broadly across our ANFC BOLD holdings. We have quality controlled and cleaned the dataset as fully as is currently practical, based on current taxonomic understanding and available resources. Our recent (February 2024) data cleaning and taxonomic updates to specimens in the dataset occurred across a number of families including (but not limited to) Apogonidae, Bathylagidae, Berycidae, Carangidae, Clupeidae, Congridae, Engraulidae, Ephippidae, Haemulidae, Moridae, Nemipteridae, Ophidiidae, Paralepididae, Pempheridae, Percophidae, Scorpaenidae, Serranidae, Synodontidae, Tetraodontidae, Trachichthyidae and Triglidae, and across Class Elasmobranchii. These families reflect the diversity of fish taxa that we are actively curating and managing, and we recognise that the dataset contains some level of misidentifications, incomplete identifications and, importantly, taxa requiring further taxonomic resolution. This is particularly so for some groups such as Ophidiidae, Apogonidae, *Uranoscopus* spp., *Champsodon* spp., *Saurida* spp. *Chimaera* spp., flatfishes (e.g. Cynoglossidae, Soleidae, *Poecilopsetta* spp.), *Torquigener* spp. and Elasmobranchii. The ongoing curation of our reference sequences in BOLD (by both ANFC researchers and BOLD data managers) supports our desire

Specimen Info Metadata	Field
Voucher information	Sample ID – BW-######
	Field ID – e.g. GT ****
	Museum ID – e.g. CSIRO H ****-**
	Collection Code – e.g. ANFC
	Institution Storing – CSIRO, Australian National Fish Collection
Taxonomy	Phylum
	Class
	Order
	Family
	Genus
	Species
	Identifier
Collection Data	Collection date
	Country/Ocean
	Exact site – within Australia's Exclusive Economic Zone
	Latitude (decimal degrees)
	Longitude (decimal degrees)

Table 2. Specimen metadata associated with each sequence in the BOLD DS-AUSDDP 'Australia's marine fishes ANFC reference COI library' dataset.

to provide accurate and up to date taxonomic outcomes for fishes, albeit one that offers a time-stamped view of Australian ichthyological taxonomy.

Data Records

This dataset presents an Australian focussed COI reference sequence library for marine fish species curated by the CSIRO ANFC and supported by co-authors. The dataset mostly consists of vouchered specimens archived and primarily stored in the ANFC (some specimens reside in co-author State museums), with their metadata, collection data and COI reference sequence. The specimen and sequence data for the 9767 marine fishes in this dataset are available via this permanent BOLD DOI[27], and through the CSIRO Data Access Portal[28]. Additionally, ANFC COI sequences (from specimens collected in the broader Australasian region and/or those resulting from ongoing collections post this dataset) are also publicly available in BOLD. All sequences are accessible for comparison purposes via the BOLD Identification Engine.

The metadata record set for each specimen includes at a minimum: date of collection and locality (given as latitudinal and longitudinal data points or, qualitatively, as a location description), the identifier, taxonomic assignment (to the lowest possible taxonomic level i.e. family, genus, species), ANFC and /or museum collection registration information (and /or field ID), tissue voucher information (see Table 2) and COI sequence. Each specimen record outlined here has been publicly released and is searchable in the Public Data Portal on BOLD or through BOLD's Data API (Application Programming Interface). The records in the dataset come from >20 ANFC publicly released projects (Fishes of Australia*, prefixed by FOA* and with ProcessIDs of FOA*###-##) and are aggregated within the BOLD DataSet DS-AUSDDP 'Australia's marine fishes ANFC reference COI library'. All sequences can also be identified by their Sample IDs beginning with BW-#. The same Sample ID can also be used to access the specimens' metadata and COI sequences in the DS-AUSDDP_specimen_sheets and DS-AUSDDP_COIsequences_Process_SampleID.fas FASTA files respectively on the CSIRO DAP.

We use the tools in BOLD as our primary workbench for submission, recording, curation, validation and editing of COI sequence records for the specimens in the dataset and indeed for our BOLD holdings more widely. We regularly update, curate, data clean and quality control the taxonomic identifications and sequence data (alongside and with support from collaborators and BOLD database managers) and add approximately 500–1000 new specimens into BOLD each year. As part of the aggregation of records into the dataset outlined here, we carried out additional cleaning and curation, focussed on COI data from ANFC specimens in taxonomic groups where ANFC staff and close collaborators have relevant experience in updating taxonomy. We also note here that BOLD follows the taxonomy of Eschmeyer's Catalog of Fishes and we are aware there will be some discrepancies/disagreements about some species' names that are not easily resolved for specimens in this dataset. Using the BOLD workbench, we removed sequencing errors (based on BOLD quality control expectations, the 9767 sequences in this dataset do not contain stop codons or contaminated sequences) and any specimens with obvious incongruencies following the construction of Neighbour-Joining (NJ) trees (using updated taxonomic information for these groups).

The dataset presented here contains 2220 species, across 288 Families and 68 Orders (see Table 3). The number of individuals barcoded per species ranges from 1 (e.g. *Argyropelecus aculeatus*, *Centropogon australis*, *Mobula birostris*) to over 40 (e.g. *Chimaera ogilbyi* and *Squalus chloroculus*) (Fig. 3) with an average of 4 sequences per species. The number of individuals per species reflects the ease of sampling from the field, taxonomic interest and current foci for ANFC taxonomists and collaborators, genetic studies within species and

Summary	N
Individuals	9767
Species	2220
BOLD BIN	2293*
Genus	1010
Family	288
Order	68
Class	5
Phylum	1

Table 3. Overview of publicly released COI reference sequences in BOLD DS-AUSDDP 'Australia's marine fishes ANFC reference COI library' dataset consisting of specimens, sequences and taxa at each taxonomic level. *30.6% of these are singleton BINs; BOLD BIN is the Barcode of Life Data System Barcode Index Number system that clusters barcode sequences algorithmically.

current resourcing for barcoding. Importantly, the taxonomic tree (as generated in BOLD - UPGMA tree based on K2P distances for the 9767 specimens, Supplementary Figure 1), visually highlights multiple ongoing challenges for Australian ichthyologists and taxonomists as discussed further below. These are most likely due to unrecognised cryptic species, ongoing assessment of lineages that are highly divergent in the COI gene fragment, and groups that may require many years work by taxonomists to resolve fully. Moreover, we are aware that the listed families in this dataset may be mismatched between different reference databases used in the taxonomic community; however, as we follow the taxonomy supported by BOLD, we direct the reader to BOLD if there are any queries about family-level classifications. Beneficially, the vouchered specimens in the dataset and their linked COI sequences help to focus and facilitate such ongoing taxonomic research and revisions.

Technical Validation

Table 4 outlines the genetic diversity statistics from the 9767 individuals and 2220 marine species identified from Australian marine territorial waters. The average within species, genus and family Kimura-2-Parameter (K2P)[29] distances were 1.27%, 11.99% and 20.40% respectively. However, as mentioned earlier, some of the taxa in the dataset present challenges, with some very high intraspecific distances, which therefore increase the mean intraspecific distance appreciably. These are perhaps due to unrecognised cryptic species, incomplete identifications (e.g. *Champsodon* sp., likely representing multiple species), taxonomy revisions required for specimens following barcoding outcomes, rare occurrences of discordance between mitochondrial gene trees and species trees and /or possible misidentification of some individuals. Figure 4 shows representative specimens from groups in which the COI sequence library is being used to help aid taxonomic clarification. These include the well-sampled teleost genus *Saurida* and the elasmobranch species *Chimaera ogilbyi* (see later). If these two taxa are omitted from the genetic diversity analyses, average within species K2P distance drops to 0.93%. In comparison, documented K2P distances of 0.39% for species, 9.93% for genus and 15.46% for family for 207 species (across 754 individuals) of mostly Australian marine fish were recorded previously[17].

We utilised the Taxon-ID Tree tool (based on Neighbour Joining - NJ)[30] trees using K2P distances and the Barcode Index Numbers (BIN; system that clusters barcode sequences algorithmically) discordance tool (based on Refined Single Linkage Algorithm, RESL[31]) in BOLD to check each of the sequence records in the dataset. All records were either assigned to an existing BIN or a new BIN was raised; records only enter the RESL analysis if they meet specific criteria[10].

Based on the clustering of records in the dataset with high COI sequence similarity (with separation of records with lower similarity)[31] across the 9767 records in the dataset, BOLD identified 2293 BINs; of these 704 (30.6%) are singleton BINs (Table 3). The similarity between the number of recognised species (n = 2220) and generated BINs indicates that in most cases there is good agreement between traditional taxonomy and COI sequences (some of this undoubtedly results from the integration of COI results into taxonomy considerations over the last 20 years). There are however also records that lack species-level resolution (i.e. cases where species do not align with sequence clusters or BINs), records that indicate cryptic speciation, and/or groups that require taxonomic revision, clarification, and description (e.g. *Bassozetus*, *Ipnops* and *Lepidotrigla*). For example, approximately 270 of the barcodes in the dataset do not have a species-level identification, while a further 400 sequences are identified to sp. or species like (i.e. 'cf.'). While lacking species identifications, these BINs are linked to voucher specimens that provide a traceable link between sequences and physical specimens to enable future revisions.

As further worked examples of some of these challenges, as Supplementary Figures 2, 3 outline, more in-depth updating of at least the genus *Saurida* (Lizardfishes) and species *Chimaera ogilbyi* is required. Based on currently published taxonomy, and noting taxonomy likely requires upgrading following barcoding outcomes for *Saurida* (and including non-Australian material) the COI sequencing results, outlined as a Maximum Likelihood (ML) K2P tree (produced in MEGA vers. X[32]) (Supplementary Figure 2) shows several apparent anomalies. For example, *S. undosquamis* is present in three distinct clades (VIII, IX and X) generating intraspecific distances of up to 14%. Clade VIII contains two subclades, differing by about 4% distance, one comprising fish from western and northern Australia, the other fish from eastern Australia. Some specimens of *Saurida* could not be allocated specific names and are currently classified as *S.* sp. (*Saurida* species undetermined) or *S.* n. sp. (*Saurida* presumptive new species) – each category generates intra-category distances (classified in

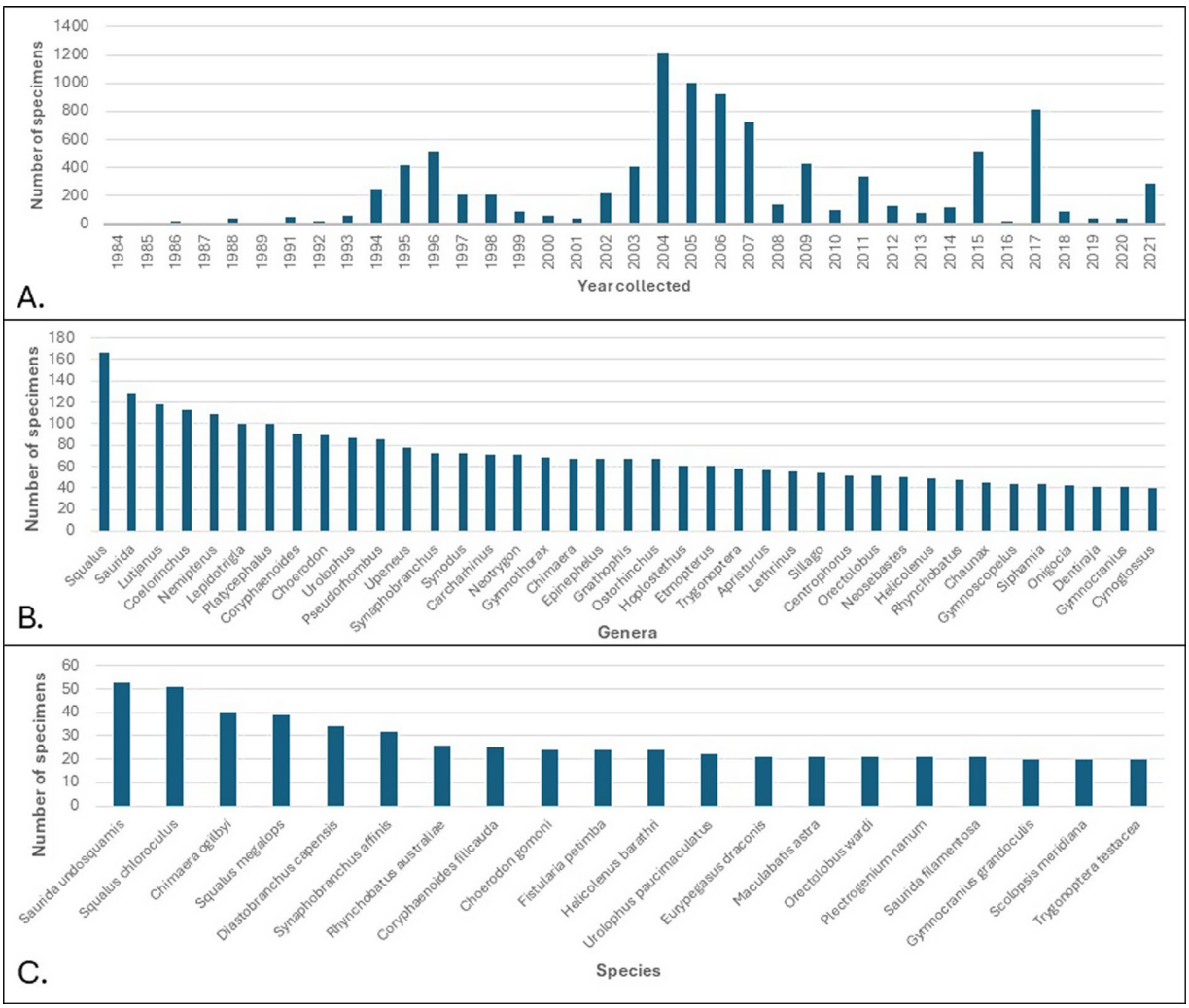

Fig. 3 Summary details for the BOLD DS-AUSDDP 'Australia's marine fishes ANFC reference COI library' dataset – (**A**) Year specimens were collected; (**B**) Genera represented by n ≥ 40 COI barcoded individuals; (**C**) Species represented by n ≥ 20 COI barcoded individuals.

Specimen level	Sample size	Min. K2P distance (%)	Mean K2P distance (%)	Max. K2P distance (%)
Within species	8900	0	1.27	30.30[a]
Within genus	7284	0	11.99	33.37
Within family	8499	0	20.40	37.52

Table 4. Genetic diversity (based on K2P distance (with pair wise deletion) sequence divergence) between individuals based on 9767 specimens in ANFC dataset. [a]*Champsodon* sp. (listed in BOLD as *Champsodon* sp. although up to 6 or 7 species may be represented under *Champsodon*).

BOLD and herein, as intraspecific distances) of up to 19%. For COI distances exceeding 2%, probabilities of conspecificity for fishes are very low[26]. *Saurida* is clearly a genus that requires further taxonomic examination as these COI data suggests possible errors in morphological identification, hybridisation, geographic isolation of taxa, ancestral haplotype sharing and /or the presence of new or cryptic species. Nomenclatural problems also abound in this genus. Old, potentially available species names exist, but type specimens may no longer be extant or are in such poor condition they are difficult to match morphologically. Historical DNA analyses of old museum material and comparison with COI data holds promise of resolving several taxonomic problems.

For the *Chimaera* 'complex' of cartilaginous fishes, the barcoding results are also outlined as a ML K2P tree (produced in MEGA vers. X[32]) (Supplementary Figure 3). Here it can be seen that *C. ogilbyi* comprises two distinct clades, V and VI, separated by about 7%. Clade V has been found exclusively from Western Australia, while Clade VI is largely an eastern Australia entity (although a few specimens have been located from Western Australia). The mitochondrial NADH2 gene showed similar separation between the two clades,

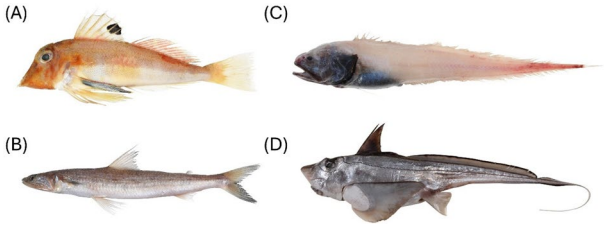

Fig. 4 Photographs of representative fish specimens from groups in which the COI sequence database is currently being used to help untangle taxonomic uncertainties. (**A**) genus *Lepidotrigla* (gurnard). *L. umbrosa*; COI sequence Sample ID: BW-A15889; Museum ID: CSIRO H 8570-07 (**B**) genus *Saurida* (lizardfish). *S. micropectoralis*; BW-A16821; CSIRO H 8763-90 (**C**) genus *Bassozetus* (cusk-eel). *B. squamosus*; BW-A13530; CSIRO H 7920-01) (**D**) species *Chimaera ogilbyi* (Ghostshark); BW-A8540; CSIRO H 7060-05.

but morphological examination and unspecified nuclear DNA analyses did not distinguish more than one species[33]. In the *Chimaera* 'complex', sequencing outcomes may reflect pairs of allopatric species with little gene flow between them – and further study of two largely allopatric groups is warranted.

Usage Notes

Our barcoding efforts over the last 17+ years represents an important resource for barcoding, metabarcoding and eDNA monitoring of Australian marine fish (for all life stages including egg, larva, juvenile and adult). The dataset outlined here is freely and publicly accessible on BOLD. The COI sequences from voucher specimens from the Australian marine estate are highly informative and currently being utilised for COI barcoding research, integrated taxonomy studies and phylogenetic analyses in classes Actinopterygii, Elasmobranchii, Holocephali and Myxini. This dataset is a curated sub-set of our larger ANFC voucher specimen holdings and BOLD COI sequences that extend to the broader Australasian region (i.e. south-east Asia, Papua New Guinea, South Pacific Islands and New Zealand) and includes some Indo/Pacific and Atlantic Oceans species and specimens added during or after 2022, the cut off period of this dataset.

In addition to single specimen barcoding, this dataset is a valuable resource for DNA metabarcoding. Metabarcoding is a special case/application applied to samples that contain more than one organism[34,35] – such as bulked DNA or eDNA samples (from environmental sources) that depend on reference sequence libraries to enable identification of taxa within mixed samples for which species identification is not otherwise practical[35,36]. With the advent of short read, high-throughput sequencing methods and with the mtDNA COI gene fragment recommended as the metabarcode for metazoans[37,38], interrogation of reference sequences (such as those in our dataset) for taxa assignment is imperative – particularly so if non-destructive and non-extractive metazoan biodiversity monitoring is to become commonplace. For Australian marine fish (adult and larval) and egg metabarcoding and Australasian regional eDNA studies, our dataset facilitates matching and identification of shorter COI fragments for identification purposes and unlike[35] GenBank, is characterised by data derived from well curated and vouchered fish specimens.

Our COI sequences are accessible through the BOLD website through the Taxonomy and Identification Engine for animal identification. As with all reference databases, we encourage users to consider the list of pairwise matches that are returned through the BOLD Identification Engine, particularly for taxa that may have high inter-clade variation (and share the same name) rather than basing a genetic identification on a single top match. Individual sequences or larger numbers of records from the dataset (with sequence data in FASTA format) can be accessed via the BOLD APIs (Public Data API, Taxonomy API, and ID Engine API). The records in our dataset are searchable in the Public Data Portal and for registered users, available via the BOLD Workbench.

Code availability

No custom code has been used in developing the dataset or the manuscript outlined here. Analysis pipelines that were used in the dataset are available on the BOLD website and in commercially available software such as MEGA-X.

Received: 28 August 2024; Accepted: 1 January 2025;
Published online: 07 January 2025

References

1. Hebert, P. D. N., Cywinska, A., Ball, S. L. & de Waard, J. R. Biological identifications through DNA barcodes. *Proc. R. Soc. B* **270**, 313–322 (2003).
2. Steinke, D., Zemlak, T. S., Boutillier, J. A. & Hebert, P. D. N. DNA barcoding of Pacific Canada's fishes. *Mar. Biol.* **156**, 2641–2647 (2009).
3. Vilgalys, R. Taxonomic misidentification in public DNA databases. *New Phytol.* **160**, 4–5 (2003).
4. Nilsson, R. H. *et al.* Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. *PLoS One* **1**, e59 (2006).
5. Collins, R. A. & Cruickshank, R. H. The seven deadly sins of DNA barcoding. *Mol. Ecol. Resour.* **13**, 969–975 (2013).
6. Bhattacharya, M. *et al.* DNA barcoding to fishes: current status and future directions. *Mitochondrial DNA A DNA Mapp. Seq. Anal.* **27**, 2744–2752 (2016).

7. Imtiaz, A., Nor, S. A. H. & Naim, D. M. D. Review: progress and potential of DNA barcoding for species identification of fish species. *Biodiversitas* **18**, 1394–1405 (2017).
8. Somervuo, P. *et al*. Quantifying uncertainty of taxonomic placement in DNA barcoding and metabarcoding. *Meth. Ecol. Evol.* **8**, 398–407 (2017).
9. Delrieu-Trottin, E. *et al*. A DNA barcode reference library of French Polynesian shore fishes. *Sci. Data* **6**, 114 (2019).
10. deWaard, J. R. *et al*. A reference library for Canadian invertebrates with 1.5 million barcodes, voucher specimens, and DNA samples. *Sci. Data* **6**, 308 (2019).
11. Collins, R. A. *et al*. Meta-fish-Lib: a generalised, dynamic DNA reference library pipeline for metabarcoding of fishes. *J. Fish Biol.* **99**, 1446–1454 (2021).
12. Grant, D. M. *et al*. The future of DNA barcoding: reflections from early career researchers. *Diversity* **13**, 313 (2021).
13. Hung, K.-W., Russell, B. C. & Chen, W.-J. Molecular systematics of threadfin breams and relatives (Teleostei, Nemipteridae). *Zool. Scri.* **46**, 536–551 (2017).
14. Appleyard, S. A., White, W. T., Vieira, S. & Sabub, B. Artisanal shark fishing in Milne Bay Province, Papua New Guinea: biomass estimation from genetically identified shark and ray fins. *Sci. Rep.* **8**, 6693 (2018).
15. Khalil, A. M., Gainsford, A. & van Herwerden, L. DNA barcoding of fresh seafood in Australian markets reveals misleading labelling and sale of endangered species. *J. Fish Biol.* **102**, 727–733 (2023).
16. Teramura, A. *et al*. Assessing the effectiveness of DNA barcoding for exploring hidden genetic diversity in deep-sea fishes. *Mar. Ecol. Prog. Ser.* **701**, 83–98 (2022).
17. Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R. & Hebert, P. D. N. DNA barcoding Australia's fish species. *Phil. Trans. R. Soc. B* **360**, 1847–1857 (2005).
18. Ward, R. D., Hanner, R. & Hebert, P. D. N. The campaign to DNA barcode all fishes, FISH-BOL. *J. Fish Biol.* **74**, 329–356 (2009).
19. Bray, D.J. Introduction to Australia's Fishes, in Bray, D. J. & Gomon, M.F. (eds) Fishes of Australia. Museums Victoria and OzFishNet, accessed 16.07.2024, http://fishesofaustralia.net.au/ (2018).
20. Steinke, D. *et al*. DNA barcoding the fishes of Lizard Island (Great Barrier Reef). *Biodivers. Data* **5**, e12409 (2017).
21. Williams, A. *et al*. Composition, diversity and biogeographic affinities of the deep-sea (200–3000 m) fish assemblage in the Great Australian Bight, Australia. *Deep-Sea Res.* **157-158**, 92–105 (2018).
22. Keesing, J. K. (Ed.). Benthic habitats and biodiversity of the Dampier and Montebello Australian Marine Parks. *Report for the Director of National Parks*. CSIRO, Australia (2019).
23. Appleyard, S. A. *et al*. Assessing DNA for fish identifications from reference collections: the good, bad and ugly shed light on formalin fixation and sequencing approaches. *J. Fish Biol.* **98**, 1421–1432 (2021).
24. Nielsen, J. G., Pogonoski, J. J. & Appleyard, S. A. Aphyonid-clade species of Australia (Teleostei, Bythitidae) with four species new to Australian waters and a new species of *Barathronus*. *Zootaxa* **4564**, 554 (2019).
25. Baldwin, C. C., Mounts, J. H., Smith, D. G. & Weigt, L. A. Genetic identification and color descriptions of early life-history stages of Belizena *Phaeoptyx* and *Astrapogon* (Teleostei: Apogonidae) with comments on identification of adult *Phaeoptyx*. *Zootaxa* **2008**, 1–22 (2009).
26. Ward, R. D. DNA barcode divergence among species and genera of birds and fishes. *Mol. Ecol. Resour.* **9**, 1077–1085 (2009).
27. Barcode of Life Datasystem https://doi.org/10.5883/DS-AUSDDP.
28. Appleyard, S. *et al*. Marine fishes (from Australia) COI barcode reference library. v1. *CSIRO. Data Collection*. https://doi.org/10.25919/8haz-aa91 (2024).
29. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol Evol.* **16**, 111–120 (1980).
30. Ratnasingham, S. & Hebert, P.D.N. BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Mol. Ecol. Notes* **7**, 355–364 (2007).
31. Ratnasingham, S. & Hebert, P. D. N. A DNA-based registry for all animal species: The Barcode Index number (BIN) system. *PLoS One* **8**, e66213 (2013).
32. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
33. Finucci, B., White, W. T., Kemper, J. M. & Naylor, G. J. P. Redescription of *Chimaera ogilbyi* (Chimaeriformes; Chimaeridae) from the Indo-Australian region. *Zootaxa* **4375**, 191–210 (2018).
34. Leray, M. & Knowlton, N. DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proc. Natl. Acad. Sci.* **112**, 2076–2081 (2015).
35. Dormontt, E. E. *et al*. Advancing DNA barcoding and metabarcoding applications for plants requires systematic analysis of herbarium collections – an Australian perspective. *Front. Ecol Evol.* **6**, 134 (2018).
36. Cristescu, M. E. From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *TREE* **29**, 566–571 (2014).
37. Elías-Gutiérrez, M., Hubert, N., Collins, R. A. & Andrade-Sossa, C. Aquatic organisms research with DNA barcodes. *Diversity* **13**, 306 (2021).
38. Andújar, C., Arribas, P., Yu, D. W., Vogler, A. P. & Emerson, B. C. Why the COI barcode should be the community DNA metabarcode for the metazoa. *Mol. Ecol.* **27**, 3968–3975 (2018).

Acknowledgements

The authors acknowledge the dedicated effort and support from many fisheries observers, field researchers, scientists (especially Alan Williams and John Keesing (CSIRO) and Tim O'Hara (Museum Victoria)) and coinvestigators, Marine National Facility support staff and crew, voyage managers, community and interested donors that provided specimens in this dataset. Additionally, ANFC researchers and photographers Gus Yearsley, Will White, Daniel Gledhill, Carlie Devine, Lou Conboy, Safia Maher, Helen O'Neill, William Zhang and Emily Gumina have greatly assisted the development of this dataset (and our larger ANFC voucher specimen holdings), Mark McGrouther and Sally Reader (Australian Museum), Ralph Foster (South Australian Museum) along with other researchers and volunteers from Australian and international fish collections. Eric Appleyard assisted the authors with specimen location mapping. Jodie van der Kamp and Lev Bodrossy gave us constructive feedback on our manuscript. Financial support for specimen sourcing, field work, field research, collection curation and molecular analyses came from several institutions – CSIRO, Museum of Victoria, Australian Museum, MAGNT, Queensland Museum, Western Australian Museum and Parks Australia. Fisheries Research Development Corporation (FRDC) supported our earlier Seafood Handbooks and associated tissue sampling. Thanks also go to the Centre for Biodiversity Genomics and BioPlatforms Australia for providing DNA sequencing capability and to BOLD for maintaining the dataset here.

Author contributions

S.A.A. drafted the manuscript, produced the DNA COI barcodes (from tissue extraction to DNA sequencing) in the dataset since 2016, currently curates the ANFC records in BOLD and analysed the technical components of the dataset. R.D.W. initiated the ANFC's BOLD reference library in 2004, curated the library between 2004 and 2014 and analysed the technical components of the dataset. J.J.P. is a current ANFC ichthyologist and identified or confirmed the identify of most of the teleost specimens in this dataset with the assistance of collaborators and helped with metadata curation. A.G. is the ANFC's collection manager and curated metadata and provided taxonomic assistance for ANFC specimens. B.H. produced the DNA COI barcodes (from tissue extraction to DNA sequencing) when B.D.W. established the reference library. P.R.L. and B.E.D. are past Directors of the ANFC; B.E.D. also analysed technical components of the dataset. M.F.G., D.J.B., J.W.J., A.C.H., G.I.M., M.P.H., B.R. and K.J.G. are Australian state museum collaborators, collected specimens and undertook taxonomic identifications of vouchered specimens. All authors provided extensive edits and comments and read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04375-4.

Correspondence and requests for materials should be addressed to S.A.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.