Linking dimensions of data on global marine animal diversity

Thomas J. Webb[1], Bart Vanhoorne[2]

1. Department of Animal and Plant Sciences, University of Sheffield, S10 2TN, UK, t.j.webb@sheffield.ac.uk

2. Flanders Marine Institute (VLIZ), Ostend, Belgium

Abstract

Recent decades have seen an explosion in the amount of data available on all aspects of biodiversity, which has led to data-driven approaches to understand how and why diversity varies in time and space. Global repositories facilitate access to various classes of species-level data including biogeography, genetics, and conservation status, which are in turn required to study different dimensions of diversity. Ensuring that these different data sources are interoperable is a challenge as we aim to create synthetic data products to monitor the state of the world's biodiversity. One way to approach this is to link data of different classes, and to inventory the availability of data across multiple sources. Here, we use a comprehensive list of >200,000 marine animal species, and quantify the availability of data on geographic occurrences, genetic sequences, conservation assessments, and DNA barcodes across all phyla and broad functional groups. This reveals a very uneven picture: 44% of species are represented by no record other than their taxonomy, but some species are rich in data. Although these data-rich species are concentrated into a few taxonomic and functional groups, especially vertebrates, data is spread widely across marine animals, with members of all 32 phyla represented in at least one database. By highlighting gaps in current knowledge, our census of marine diversity data helps to prioritise future data collection activities, as well as emphasising the importance of ongoing sustained observations and archiving of existing data into global repositories.

Introduction

The explosion in the availability of data describing the natural world has, in recent decades, transformed the kinds of questions that we can now ask as ecologists. Efforts to reconstruct the evolutionary relationships between all living species (e.g. Open Tree of Life; [1,2]) can draw upon over 200M sequences (https://www.ncbi.nlm.nih.gov/genbank/statistics/) from over 170,000 metazoan species stored in GenBank [3,4]. In 2018, the Global Biodiversity Information Facility (GBIF, [5]) passed a billion species occurrence records (https://www.gbif.org/news/5BesWzmwqQ4U84suqWyOQy/big-data-for-biodiversity-gbiforg-surpasses-1-billion-species-occurrences), providing an unparalleled resource for students of biogeography. The conservation status of >116,000 species has now been formally assessed [6]. Significant efforts are underway to collate data biological, physiological, metabolic and thermal traits [7-11] across multiple species, as well as information on animal movement [12,13] and ecological interactions [14].

Against this background of increased data availability, the oceans are still often characterised as the data-poor relative of the data-rich land. Various autonomous platforms operating throughout the world's oceans do now enable vast quantities of physical and biogeochemical data to be transmitted [15] but marine biodiversity data remain more challenging to collect. In part, the vastness of the oceans precludes routine and casual observation by the citizen scientists who have contributed so much to the collection of terrestrial biodiversity data [16,17], except in some more accessible coastal areas [18-20]. However coordinated global initiatives have made enormous progress in collating existing data and promoting systematic new data collection. The Census of Marine Life [21] drove this effort from 2000-2010, and its legacies include the Ocean Biodiversity Information System (OBIS, [22]), which currently holds nearly 60M occurrence records from over 120,000 marine species. Initiatives like this have built on sustained observations of marine ecosystems [23], and continue to be developed to deliver the Essential Biodiversity

Variables that we need to monitor progress towards Sustainable Development Goals (e.g. [24]). Application of technologies from satellites and drones to biologgers and molecular methods such as eDNA continue to expand the range of data available to marine biodiversity scientists [25]. Crucially, the accumulation of data has proceeded in parallel with massive improvements in data infrastructure, and much better tools (taking advantage of the improved computing power available even to casual users) with which to access and analyse it [26,27]. This is important because the challenge now is to extract meaning from the sea of data, to deliver effective outcomes for marine conservation and monitoring of the state of the global ocean [19,24].

Although access to biodiversity data of different types is now much improved, to extract full value from existing data requires linking together different datasets that were often collected for different purposes, by different organisations and at different times. This kind of interoperability of diversity data is central to the vision of a ‘macroscope’ to sample and monitor the entire biosphere [25], and is a fundamental principle of the Bari Manifesto of best practice in biodiversity informatics [28]. Progress towards such interoperability requires comparable coverage across multiple classes of data and dimensions of diversity, as well as parallel measures of the abiotic environment and of human pressures. An exemplar of successful data integration for terrestrial plant communities is the Botanical Information and Ecology Network [29] which combines standardised information on plant distributions, traits, and evolutionary relationships with the computational tools needed to work with them. An important step towards this kind of model is to fully understand the gaps and biases in available data. In the marine environment, key gaps in the overall knowledge of marine biodiversity have been documented [30-32], including estimates of the extent of unknown biodiversity [33] and undocumented extinction risk [34]. Efforts to quantify these gaps across different key variables and data sources have been limited to the regional scale, but have shown for instance that the species and taxonomic groups that we know most in one dimension (e.g. global occurrences) tend to be those that we also know most about in another (e.g. biological traits, extinction risk; [34,35]). To date we lack a global overview of how data (and gaps) are co-distributed across axes of marine diversity, to compare for example with previous global analyses of terrestrial plants [36].

Such a task is feasible however, given the availability of a standardised global taxonomy of marine species, the World Register of Marine Species (WoRMS, [37]), which includes links out to other key biodiversity datasets (Table 1). In this paper, we focus on key data sources which, when linked to robust taxonomy, individually or in combination can be used to construct different dimensions of marine diversity. We consider geographic occurrences and nucleotide sequences to be the fundamental building blocks of the spatial and phylogenetic dimensions of diversity, which interact to structure the distribution of key ecological traits across species [38]. A first step to adding the functional dimension of diversity is to classify species into broad ecological guilds, similar to the way in which species can be classified in global theories and models of biodiversity [39,40]. Supplementing these with information on conservation status and molecular taxonomy provides insights into how marine diversity is changing, and how we might efficiently monitor this. Throughout we use open source computational tools to link data across these components of marine diversity to take stock of the current state of data availability, identifying gaps and priorities for future work. In this way we summarise data availability across multiple axes for >200,000 marine animal species from 32 phyla and across broad ecological guilds (e.g. benthos, zooplankton, seabirds), and we assess the extent to which this availability is correlated across different classes of diversity data. Above all, our aim is to highlight the wealth of marine biodiversity data that we have amassed as a community over centuries, and the opportunities that we now have to link different classes of data in order to better understand the dimensions of marine diversity.

Methods

To provide an overview of the state of knowledge of marine animal biodiversity, we mine the World Register of Marine Species (WoRMS, [37]), the most comprehensive source of taxonomic information on marine species, consisting of over half a million distinct names checked by expert taxonomic editors. We focus our investigation on marine animals, and so filtered the WoRMS database to Kingdom Animalia, retaining only those species considered to be marine by WoRMS (flag `isMarine` is TRUE), and excluding any species only known from fossils. We consider only taxa identified at the species rank, with a current accepted name and valid WoRMS identifier (Aphia ID).

In addition to taxonomy, WoRMS has aggregated data on species attributes including broad 'functional groups'. In reality these are closer to ecological guilds, defining habitat affinity (e.g. benthos, zooplankton) rather than ecological function, but for transparency we retain the terminology employed by WoRMS. We use these attributes to assign each species to a functional group, using a dedicated R function (https://github.com/tomjwebb/WoRMS-functional-groups) which accesses the WoRMS API using the `worrms` R package [41]. We supplement these functional groups with taxonomic groups to identify fish (using the WoRMS paraphyletic Superclass Pisces; [42]), marine mammals, seabirds, and reptiles. We consolidate functional groups into broad categories for maximum coverage - for example, our 'benthos' group includes all species categorised in WoRMS as endobenthos, epibenthos, hyperbenthos, macrobenthos, meiobenthos, and microbenthos, as well as those originally classified simply as benthos. When separate functional groups are recorded for different life stages, we always use the group for the adult stage. We group together categories with very few species (including meso, macro, neuston) and species with no functional group classification into the single category 'other/unknown'. For fish we include an additional grouping variable based on the broad habitat categories recorded in FishBase [10] accessed using the `rfishbase` package [43], classifying 17,568 of 18,261 species as bathydemersal, bathypelagic, benthopelagic, demersal, pelagic-oceanic, pelagic-nertitic, or reef-associated.

The WoRMS database includes links to other major biodiversity databases (table 1), and we exploit these to compare the state of biodiversity information availability across axes of biogeography, genetics, conservation, and molecular taxonomy. Specifically, we record for each species its total number of occurrences in the Ocean Biogeographic Information System (OBIS, [22]), and its total number of nucleotide sequences in GenBank. The taxonomy in OBIS is standardised to WoRMS, making these links straightforward, and GenBank's taxonomic information is generally reliable for marine animals [4] meaning that links between WoRMS and GenBank are likely to robustly associate relevant sequences with the correct taxonomic identifier. We also record for each species its IUCN conservation assessment category (if available), and whether or not it has DNA barcodes listed in the Barcode of Life Data System (BOLD).

Using our tidy database linking the diversity data sources shown in table 1, we then summarise the availability of biodiversity data across all marine animals as follows. First, we consider the two major quantitative databases, OBIS and GenBank. We calculate the proportion of species within each phylum with records in each of these databases, and the distribution of records between species within each phylum. To derive an indication of relative data availability across functional groups, highlighting groups that are particularly highly likely (or unlikely) to occur in the dataset, and those which tend to have more records when they are present, we model data availability across functional groups. We apply a two-step hurdle process, because of the high degree of zero-inflation in our data [44]. To assess whether certain functional groups were better represented in the databases than others, we model presence of species in OBIS or GenBank using a binomial GLM of the form species presence ~ functional group, and we model the distribution of counts (OBIS records or GenBank nucleotides) between functional groups, for those species present in the data source, using a zero-truncated negative binomial GLM. These hurdle models are implemented using the `hurdle` function in the `pscl` package [44,45]. For visualisation, we plot the exponentiated binomial coefficients from the zero component of the model, which shows the ratio of the probability of getting a non-zero to a zero observation within a functional group. We also plot the predicted counts for the subset of species in each functional group with non-zero counts.

To assess whether data availability is correlated across data sources, we use categorical scales of numbers of records per species in both OBIS and GenBank, using categories bounded by upper limits of 0, 1, 10, 100, 1,000, 10,000, 100,000, and a final category of >100,000 records. We use mosaic plots [46], created using the `ggmosaic` R package [47], to illustrate the distribution of GenBank count categories for each OBIS count category. We also consider how IUCN conservation assessments are distributed across species in different functional groups, and between species present and absent in OBIS, and we compare the number of OBIS occurrence records between species in different IUCN categories. To simplify this analysis, we aggregate to the following IUCN assessment categories: Not Assessed, Data Deficient (i.e., formally assessed but insufficient data to assign the species to a threat category), Threatened (formally assessed as Vulnerable, Endangered, Critically Endangered, Conservation Dependent, Extinct in the Wild, or Extinct), and Non-threatened (formally assessed as Near Threatened or Least Concern). We perform a similar analysis comparing species presence or absence in the Barcode of Life database with presence in OBIS and number of OBIS records.

All data and links were extracted from WoRMS on 2020-01-11 and the statistics we report are correct as of that date. Manipulation, visualisation, and analysis is performed in R 3.6.2 [48] using RStudio 1.2.5033 [49] and the `tidyverse` suite of packages [50] as well as `worrms` [41] to access the WoRMS API and `rfishbase` {Boettiger:2012bz} to access FishBase, and the plotting packages `ggmosaic [47]`, `ggbeeswarm` [51] and `patchwork [52]`. All data used in this article are publicly available via WoRMS. The processed summary data we use for our analysis is openly available under a Creative Commons

Attribution 4.0 International License in the Marine Data Archive (https://doi.org/10.14284/417). R code to replicate our analyses and figures is available via https://github.com/tomjwebb/linking_marine_diversity_data and is archived on Figshare via the University of Sheffield's Online Research Data repository here: https://doi.org/10.15131/shef.data.12833891.

Results

Our final dataset consisted of 206,849 valid marine animal species, from 32 phyla and 89 classes. Of these, 106,213 (51%) have at least one occurrence record listed in OBIS (table 2). Of these, 18,869 (18% of species in OBIS, 9% of all species) are represented by just a single occurrence record (table 2), while one species (Atlantic Cod, Gadus morhua) has over a million occurrence records (1,108,463). Overall, there are 45,974,726 OBIS occurrence records across all species. 36,094 (17%) of all species have at least one nucleotide recorded in GenBank, while 8 species (five fish, the Antarctic Minke Whale Balaenoptera bonaerensis, the tunicate Ciona intestinalis and the California Sea Hare Aplysia californica) have more than a million. Overall the species in our database total 56,846,294 GenBank nucleotides. Furthermore, 13,179 species have had their conservation status assessed by the IUCN, and 25,272 have at least one DNA barcode in the Barcode of Life database.

The distribution of OBIS and GenBank records across animal phyla and functional groups is shown in Fig 1. At least one species from every phylum has records in either OBIS or GenBank, with all phyla except Loricifera (which has just 29 species) represented in both databases (Fig 1A). Across all phyla, just over half (55%) of all species are represented in one or other database. Most species that are present in OBIS have only a few occurrence records, with median values of records ranging from 1 to 92 across phyla (Fig 1B). A similar pattern is observed for GenBank nucleotides (fig 1C), with median values between 1 and 94 except in phyla Orthonectida and Placozoa, both of which have only two species represented in GenBank, one of which has several thousand nucleotides (in Orthonectida, Intoshia linei has 3,522, in Placozoa, Trichoplax adhaerens has 29,176).

Data availability is variable across functional groups (fig 1B, C; fig 2). Modelling the presence or absence of species in OBIS in a binomial GLM shows that species of fish, mammal, bird, and reptile are much more likely to have occurrences in OBIS than are benthic or zooplankton species, with nekton falling in between, and species with unknown or other functional group classification the least likely to have occurrence records (fig 2A). A broadly similar pattern holds when modelling the number of occurrence records for those species with at least 1 (fig 2B), with the vertebrate taxa again tending to have most records, although distinctions between vertebrates and other groups are less stark. Benthic invertebrates typically have few OBIS records, but zooplankton that do occur in OBIS tend to have more records than nekton. In GenBank, birds, reptiles and mammals are most likely to be present in the database, followed by fish, nekton, and zooplankton, with benthos and other/unknown functional groups least likely to be represented (fig 2C). The rank order changes somewhat when considering number of nucleotides across species present in GenBank (fig 2D), with most records from mammals and reptiles, followed but birds and fish. Nekton tend to have fewest records, but there is considerable variability within all major groups. Data availability in both major databases is broadly similar across fish habitat groupings (figure S1, S2).

Considering the joint distribution of species across OBIS and GenBank categorical scales, 93,519 (45%) species have no records in either database (table 2, fig 3A). In general, species with more records in OBIS also tend to have more nucleotides in GenBank (table 2, fig 3), indicating that these different biodiversity data aggregators have similar biases in terms of the known marine biodiversity that they encompass. There are exceptions though: in particular several species have many (>100,000) GenBank nucleotides but very few (if any) OBIS records (table 3).

A similar pattern is evident when examining the distribution of OBIS records across different IUCN assessment categories. In general, and across functional groups, the proportion of species with records in OBIS is higher in assessed species (threatened and non-threatened) than it is in unassessed or data-deficient species: overall, 84% of threatened and 94% of non-threatened species have occurrence records in OBIS, compared to 75% of data-deficient and 49% of unassessed species (table 4A). Considering only those species with records in OBIS, there is considerable variation within and between IUCN categories in the number of occurrence records per species, but a general tendency is apparent in all functional groups for species in threatened and non-threatened categories to have more occurrence records than those in data-deficient and unassessed categories (fig 4A).

Species with DNA barcodes are disproportionately likely to also have occurrence records in OBIS: 45% of species with no record in the Barcode of Life database have at least one occurrence record in OBIS, compared to 89% of species with a barcode (table 4B). In addition, in all functional groups, species with barcodes tend to have more OBIS records than those which do not (fig 4B).

Discussion

Using the taxonomic backbone of the World Register of Marine Species [37] we have summarised data availability across axes of biogeography, genetics, molecular taxonomy, and conservation status for 206,849 marine animal species. This presents a mixed picture. One the one hand, 91,828 (44%) species have no records in any of these databases, and are represented only by their name. This is considerably higher than the 27% of plant species with no information other than their name [36], although of course the marine environment represents far larger habitable volume [53] and marine animals are a much more diverse taxonomic group. Only 6,688 marine animal species (3%) have records in all four of the datasets that we consider – again, rather lower than the 18% of broadly-covered plant species [36]. At the same time, it is important to remember that presence in a dataset does not imply extensive knowledge: among the 106,203 species with records in OBIS, for example, the median number of recorded occurrences is just 7, and 18% of these species (18,869 species) are known from only a single occurrence. Nonetheless, the distribution of biogeographic and genetic information across the animal tree of life is extensive, with all animal phyla represented in at least one database (fig 1). Data availability tends to be biased towards well-known taxa and functional groups (especially vertebrates; figs 1, 2, 4), in agreement with previous assessments (e.g. [32]), but the subset of 225 species with >1,000 occurrences in OBIS and >1,000 nucleotides in GenBank is drawn from 10 phyla and 27 classes, representing all major functional groups, and most of them have a barcode in BOLD (214 species), and have been assessed by the IUCN as something other than data deficient (102 non-threatened, 23 threatened species). For these diverse marine animal species then, it is reasonable to propose that the information available across multiple sources can be translated into knowledge about their distribution, evolutionary relationships, and conservation status.

The broad positive correlation between data availability across different sources (table 2, table 4, fig 3) reinforces previous findings that species with good information on one facet of their biology and ecology tend to be well represented in other databases too, both in plants [36] and in marine species [35]. These information-rich species are likely to be those most easily and frequently observed, or those of high economic or cultural value, and so will not be a random subset of all species. However, the consequences of biases towards data availability from these common species will vary depending on the specific question of interest. For instance, ecosystem function may be driven largely by just those common species that tend to be so well known [54]; but rare species will clearly be of great interest to conservationists, and may indeed sometimes contribute unique trait combinations to marine communities [55].

In terrestrial conservation, considerable concern has been expressed over the likely conservation status of species too poorly known to formally assess, as they tend to have characteristics (rarity, small ranges, occurring in poorly studied regions) which will predispose them to be at risk [56]. For some marine taxa this appears to be the case too, with high rates of extinction risk predicted for European sharks and rays formally assessed as Data Deficient [57], and low levels of conservation assessment in poorly-known marine groups may contribute to low overall documented levels of extinction risk [58]. On the other hand, the fact that the biggest data gaps in marine biodiversity tend to be in remote habitats largely inaccessible to humans (e.g. the deep pelagic ocean; [59]), and the highest rates of discoveries of new species and habitats are also in the deep sea [60,61], provides some contrast with the terrestrial situation, and may insulate these poorly-known species somewhat from human pressures. However, some patterns still hold in the deep sea, such as the tendency for widespread species to be encountered and described first [62], meaning that many of the species not yet present in major databases may be genuinely rare. Given the acceleration of human activities into previously unexploited regions of the oceans [63], with new threats including deep sea mining [64] and exploitation of the mesopelagic [65], it seems unwise to assume that the large fraction of marine biodiversity that remains poorly known is not at risk. Given the fact that Data Deficient conservation assessments are twice as frequent in marine versus non-marine taxa [34], data-driven predictive conservation assessments [57,66,67] which rely on some of the kinds of data we consider here (spatial distribution, evolutionary relationships, ecological guilds) combined with biological traits may prove to be especially valuable tools.

An aim of this study was to flag priorities for future work. One important point is that the major publicly available databases on which we draw do not constitute the sum total of data ever collected on marine species. This is particularly the case for occurrence data, as globally researchers have yet to adopt the routine deposition of species occurrences in OBIS as a cultural norm, in the way that genetic sequence data is deposited in GenBank. To this end, improving incentives for researchers to add their data to global repositories in an important goal [25], while data archaeology and rescue initiatives can help to ensure that historical data are captured [68]. Equally, it remains vital that ongoing survey schemes are properly valued [69], even as novel exploration is planned. At the same time, our quantification exercise can help to identify groups of species where a little additional research effort in one area would quickly result in a more valuable dataset. One candidate set of species might be those that are frequently observed but poorly represented in other databases. For instance, 1,216 species have >1,000 OBIS records but <10 GenBank nucleotides; and over half of the 3,533 species with >1,000 OBIS occurrences are either not assessed by the IUCN (1,876 species) or data-deficient (82 species). The fact that almost 90% (3,163) of these species have DNA barcodes in BOLD is encouraging, however, suggesting considerable potential for an increasing role for molecular studies to address a wide range of questions in marine ecology [70].

Mining the spatial information already present in other databases also has potential for supplementing existing occurrence datasets. In this study we relied on existing links between WoRMS and GenBank and BOLD, which simply summarise the number of nucleotides or barcodes present for each species. The spatial meta-data stored in the sequence databases provides an additional source of information, although in GenBank this data is relatively unstructured. Searching the GenBank nucleotide database, we found just 1,437 records for animals which contained a lat-lon field; matching this to our list of marine animals reduced this further to 183 records from 42 species. Nonetheless, even from this small set of species, 21 do not have occurrence records in OBIS, suggesting that mining GenBank for spatial data would likely add valuable information for a small number of species. Various methods have been developed to attempt this, based around mining spatial information from the full text of associated publications [71,72] with initiatives such as the Genomic Observatories MetaDatabase (GEOME, https://geome-db.org) also seeking to simplify access to meta-data from sequence datasets.

BOLD typically does store spatial data for individual specimens in a well-structured manner, only some of which has been harvested by OBIS. In our dataset, 3,117 species have BOLD barcodes but no OBIS records. Several of these are parasites, which we know are not well recorded in OBIS (e.g. Schistocephalus solidus, 718 barcodes; Anguillicoloides crassus, 508 barcodes) but there are free-living marine species too, such as the Gastropod mollusc Littoraria sinensis (257 barcodes) and the Copepod Calanoides natalis (183 barcodes). Accessing the specimen data from BOLD using the `bold` R package [73] for these two species reveals that none of the L. sinensis have information in the latitude and longitude fields, but full geographic information is available for 227 specimens for Calanoides natalis. Although none of these locations are currently recorded in OBIS, some are in GBIF, highlighting the often complex pipelines from data providers to global data aggregators. Improving pipelines from genetic databases to occurrence databases is currently a priority for OBIS (W. Appeltans, OBIS Project Manager, pers. comm.).

Finally, the dimensions of diversity that we summarise in this study are somewhat limited. We did not consider the traits of species, for instance, beyond functional groups that indicate habitat affiliation in very broad terms (e.g. benthic vs planktonic). These groupings are already useful as global patterns of diversity are known to differ between them [39], and they can also be used to refine methods of matching species occurrences to global sea temperature datasets [74], helping to predict species responses to climate change [75]. Beyond these coarse functional groups, however, traits data remain scarce even in reasonably common marine species in well-studied regions [35], and despite many efforts at collating traits – including within WoRMS; [76] - there is still no widely-adopted central standard [77]. Certain groups are well covered by existing initatives (e.g. FishBase [10], the Coral Trait Database [11]), and whether a single overarching portal to cover the immense diversity of marine lifeforms is possible – or even desirable – remains open for discussion. However, it is certainly the case that multiple smaller-scale projects collect valuable traits data for a subset of species which is typically made available (if at all) via supplementary material or bespoke web portals, at risk of being lost to the community. A wider adoption of principles embedded in initiatives like the Open Traits Network [7] would ensure interoperability of these small, project-specific traits datasets, maximising the availability of information on key traits for the largest possible fraction of marine diversity. Readily availabilie information on even just a few traits (e.g. body size, longevity, fecundity, planktonic larval duration) would help to test predictions from biodiversity models, embed life history theory into marine conservation, and predict the consequences of human activities on marine diversity [39,78-80].

The stocktake of marine biodiversity data availability that we have undertaken here adds to previous efforts focused on occurrence data [19,32,81]. While we reveal a similar story of gaps and biases across other data sources, there is considerable overlap in coverage too, and overall the potential to link dimensions of marine animal diversity is now high. The priority now should be to build on the substantial community-built foundations and to improve the pipeline from raw data to interoperable data products, both as a resource for fundamental macroecological research and to facilitate effective stewardship of our blue planet.

Acknowledgements

The ideas for this study were conceived while TJW was a Royal Society University Research Fellow, and developed over the course of the Natural Environment Research Council and Department for Environment, Food and Rural Affairs Marine Ecosystems Research Programme (grant number NE/L003279/1) and EMODnet Biology. Thanks to Helmut Hillebrand for the invitation to present this work at the 2019 HIFMB Symposium on Functional Marine Biodiversity and for the opportunity to contribute to this volume. Constructive and insightful comments from two anonymous reviewers have helped to greatly improve this work.

References

1. 1. Redelings, B. D. & Holder, M. T. 2017 A supertree pipeline for summarizing phylogenetic and taxonomic information for millions of species. PeerJ **5**, e3058. (doi:10.7717/peerj.3058)

2. Hinchliff, C. E. et al. 2015 Synthesis of phylogeny and taxonomy into a comprehensive tree of life. P Natl Acad Sci USA **112**, 12764–12769. (doi:10.1073/pnas.1423041112)

3. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. 2016 GenBank. Nucleic Acids Res. **44**, D67–D72. (doi:10.1093/nar/gkv1276)

4. Leray, M., Knowlton, N., Ho, S.-L., Nguyen, B. N. & Machida, R. J. 2019 GenBank is a reliable resource for 21st century biodiversity research. P Natl Acad Sci USA **116**, 22651–22656. (doi:10.1073/pnas.1911714116)

5. GBIF 2020 GBIF Home Page. https://www.gbif.org.

6. IUCN 2020 The IUCN Red List of Threatened Species. Version 2020-1. httpswww.iucnredlist.org.

7. Gallagher, R. et al. 2019 The Open Traits Network: Using Open Science principles to accelerate trait-based science across the Tree of Life. ecoevorxiv.org. (doi:https://doi.org/10.32942/osf.io/kac45)

8. Bennett, J. M. et al. 2018 GlobTherm, a global database on thermal tolerances for aquatic and terrestrial organisms. Sci. Data **5**, 180022. (doi:10.1038/sdata.2018.22)

9. Makarieva, A., Gorshkov, V. & LI, B. 2005 Biochemical universality of living matter and its metabolic implications. Funct Ecol

10. Froese, R. & Pauly, D. 2019 FishBase. World Wide Web electronic publication. version (12/2019). www.fishbase.org.

11. Madin, J. S. et al. 2016 The Coral Trait Database, a curated database of trait information for coral species from the global oceans. Sci. Data **3**, 178–22. (doi:10.1038/sdata.2016.17)

12. Kranstauber, B., Cameron, A., Weinzerl, R., Fountain, T., Tilak, S., Wikelski, M. & Kays, R. 2011 The Movebank data model for animal tracking. Environmental Modelling & Software **26**, 834–835. (doi:10.1016/j.envsoft.2010.12.005)

13. Wikelski, M., Davidson, S. C. & Kays, R. 2020 Movebank: archive, analysis and sharing of animal movement data. www.movebank.org.

14. Poelen, J. H., Simons, J. D. & Mungall, C. J. 2014 Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. Ecological Informatics **24**, 148–159. (doi:10.1016/j.ecoinf.2014.08.005)

15. Tanhua, T. et al. 2019 Ocean FAIR Data Services. Front. Mar. Sci. **6**, 92. (doi:10.3389/fmars.2019.00440)

16. Silvertown, J. 2009 A new dawn for citizen science. Trends Ecol Evol **24**, 467–471. (doi:10.1016/j.tree.2009.03.017)

17. Chandler, M. et al. 2017 Contribution of citizen science towards international biodiversity monitoring. Biological Conservation **213**, 280–294. (doi:10.1016/j.biocon.2016.09.004)

18. Hyder, K., Townhill, B., Anderson, L. G., Delany, J. & Pinnegar, J. K. 2015 Can citizen science contribute to the evidence-base that underpins marine policy? Marine Policy **59**, 112–120. (doi:10.1016/j.marpol.2015.04.022)

19. Edgar, G. J., Bates, A. E., Bird, T. J., Jones, A. H., Kininmonth, S., Stuart-Smith, R. D. & Webb, T. J. 2015 New Approaches to Marine Conservation Through Scaling Up of Ecological Data. Annual Review of Marine Science **8**, 150807173619006. (doi:10.1146/annurev-marine-122414-033921)

20. Edgar, G. J. & Stuart-Smith, R. D. 2014 Systematic global assessment of reef fish communities by the Reef Life Survey program. Sci. Data **1**, 1–8. (doi:10.1038/sdata.2014.7)

21. Snelgrove, P. V. R. 2010 Discoveries of the Census of Marine Life: Making Ocean Life Count. 1st edn. Cambridge University Press.

22. OBIS 2020 Ocean Biodiversity Information System. www.iobis.org.

23. Mieszkowska, N., Sugden, H., Firth, L. B. & Hawkins, S. J. 2014 The role of sustained observations in tracking impacts of environmental change on marine biodiversity and ecosystems. Philos T R Soc A **372**, 20130339–20130339. (doi:10.1098/rsta.2013.0339)

24. Miloslavich, P. et al. 2018 Essential ocean variables for global sustained observations of biodiversity and ecosystem changes. Global Change Biol **105**, 10456. (doi:10.1111/gcb.14108)

25. Dornelas, M. et al. 2019 Towards a macroscope: Leveraging technology to transform the breadth, scale and resolution of macroecological data. Global Ecol Biogeogr **28**, 1937–1948. (doi:10.1111/geb.13025)

26. Basset, A. & Los, W. 2012 Biodiversity e-Science: LifeWatch, the European infrastructure on biodiversity and ecosystem research. Plant Biosystems - An International Journal Dealing with all Aspects of Plant Biology **146**, 780–782. (doi:10.1080/11263504.2012.740091)

27. La Salle, J., Williams, K. J. & Moritz, C. 2016 Biodiversity analysis in the digital era. Philos T R Soc B **371**. (doi:10.1098/rstb.2015.0337)

28. Hardisty, A. R. et al. 2019 The Bari Manifesto: An interoperability framework for essential biodiversity variables. Ecological Informatics **49**, 22–31. (doi:10.1016/j.ecoinf.2018.11.003)

29. Maitner, B. S. et al. 2018 The bien r package: A tool to access the Botanical Information and Ecology Network (BIEN) database. Methods in Ecology and Evolution **9**, 373–379. (doi:10.1111/2041-210X.12861)

30. Costello, M., Coll, M., Danovaro, R., Halpin, P. & Ojaveer, H. 2010 A Census of Marine Biodiversity Knowledge, Resources, and Future Challenges. PLoS ONE

31. Snelgrove, P. et al. 2016 Global Patterns in Marine Biodiversity. In The First Global Integrated Marine Assessment World Ocean Assessment, un.org.

32. Miloslavich, P. et al. 2016 Extent of Assessment of Marine Biological Diversity. In The First Global Integrated Marine Assessment World Ocean Assessment (eds L. Inniss & A. Simcock), United Nations.

33. Appeltans, W. et al. 2012 The magnitude of global marine species diversity. Curr. Biol. **22**, 2189–2202. (doi:10.1016/j.cub.2012.09.036)

34. Webb, T. J. & Mindel, B. L. 2015 Global Patterns of Extinction Risk in Marine and Non-marine Systems. Current Biology **25**, 506–511. (doi:10.1016/j.cub.2014.12.023)

35. Tyler, E. H. M., Somerfield, P. J., Berghe, E. V., Bremner, J., Jackson, E., Langmead, O., Palomares, M. L. D. & Webb, T. J. 2012 Extensive gaps and biases in our knowledge of a well-known fauna: implications for integrating biological traits into macroecology. Global Ecol Biogeogr **21**, 922–934. (doi:10.1111/j.1466-8238.2011.00726.x)

36. Cornwell, W. K., Pearse, W. D., Dalrymple, R. L. & Zanne, A. E. 2019 What we (don't) know about global plant diversity. Ecography, ecog.04481. (doi:10.1111/ecog.04481)

37. WoRMS Editorial Board 2020 World Register of Marine Species. http://www.marinespecies.org.

38. Freckleton, R. P. & Jetz, W. 2009 Space versus phylogeny: disentangling phylogenetic and spatial signals in comparative data. P R Soc B **276**, 21–30. (doi:10.1098/rspb.2008.0905)

39. Worm, B. & Tittensor, D. P. 2018 A Theory of Global Biodiversity. Princeton and Oxford: Princeton University Press. (doi:10.2307/j.ctt1zkjz6q)

40. Harfoot, M. B. J., Newbold, T., Tittensor, D. P., Emmott, S., Hutton, J., Lyutsarev, V., Smith, M. J., Scharlemann, J. P. W. & Purves, D. W. 2014 Emergent Global Patterns of Ecosystem Structure and Function from a Mechanistic General Ecosystem Model. PLoS Biol **12**, e1001841. (doi:10.1371/journal.pbio.1001841)

41. Chamberlain, S. 2019 World Register of Marine Species (WoRMS) Client. R package worrms version 0.4.0. httpsCRAN.R-project.orgpackageworrms.

42. WoRMS 2020 Pisces. httpswww.marinespecies.orgaphia.phpptaxdetailsid.

43. Boettiger, C., Lang, D. T. & Wainwright, P. C. 2012 rfishbase: exploring, manipulating and visualizing FishBase data from R. Journal of Fish Biology **81**, 2030–2039. (doi:10.1111/j.1095-8649.2012.03464.x)

44. Zeileis, A., Kleiber, C. & Jackman, S. 2008 Regression models for count data in R. Journal of Statistical Software **27**, 1–25.

45. Jackman, S. 2020 pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory. United States Studies Centre, University of Sydney. Sydney, New South Wales, Australia. R package version 1.5.5. httpsgithub.comatahkpscl.

46. Hofmann, H. 2008 Mosaic Plots and Their Variants. In Handbook of Data Visualisation (eds C.-H. Chen W. Härdle & A. Unwin), Berlin Heidelberg.

47. Jeppson, H., Hofmann, H. & Cook, D. 2018 ggmosaic: Mosaic Plots in the ‘ggplot2’ Framework. R package version 0.2.0. https://CRAN.R-project.orgpackageggmosaic.

48. R Core Team 2019 R: A language and environment for statistical computing. https://www.R-project.org.

49. RStudio Team 2019 RStudio: Integrated Development for R. httpwww.rstudio.com.

50. Wickham, H. et al. 2019 Welcome to the Tidyverse. Journal of Open Source Software **4**, 1686. (doi:10.21105/joss.01686)

51. Clarke, E. & Sherrill-Mix, S. 2017 ggbeeswarm: Categorical Scatter (Violin Point) Plots. R package version 0.6.0. httpsCRAN.R-project.orgpackageggbeeswarm.

52. Pedersen, T. L. 2019 patchwork: The Composer of Plots. R package version 1.0.0. httpsCRAN.R-project.orgpackagepatchwork.

53. Dawson, M. N. 2012 Species richness, habitable volume, and species densities in freshwater, the sea, and on land. Frontiers of Biogeography **4**, fb_12675.

54. Gaston, K. & Fuller, R. 2008 Commonness, population depletion and conservation biology. Trends Ecol Evol **23**, 14–19.

55. Mouillot, D. et al. 2013 Rare Species Support Vulnerable Functions in High-Diversity Ecosystems. PLoS Biol **11**, e1001569. (doi:10.1371/journal.pbio.1001569)

56. Scheffers, B. R., Joppa, L. N., Pimm, S. L. & Laurance, W. F. 2012 What we know and don’t know about Earth's missing biodiversity. Trends Ecol Evol **27**, 501–510. (doi:10.1016/j.tree.2012.05.008)

57. Walls, R. H. L. & Dulvy, N. K. 2019 Predicting the conservation status of Europe’s Data Deficient sharks and rays. bioRxiv **276**, 614776. (doi:10.1101/614776)

58. Mindel, B. L., Webb, T. J., Neat, F. C. & Blanchard, J. L. 2016 A trait-based metric sheds new light on the nature of the body size–depth relationship in the deep sea. J Anim Ecol **85**, 427–436. (doi:10.1111/1365-2656.12471)

59. Webb, T. J., Vanden Berghe, E. & O'Dor, R. 2010 Biodiversity's big wet secret: the global distribution of marine biological records reveals chronic under-exploration of the deep pelagic ocean. PLoS ONE **5**, e10223. (doi:10.1371/journal.pone.0010223)

60. Ramirez-Llodra, E. et al. 2010 Deep, diverse and definitely different: unique attributes of the world's largest ecosystem. Biogeosciences **7**, 2851–2899. (doi:10.5194/bg-7-2851-2010)

61. Danovaro, R., Snelgrove, P. V. R. & Tyler, P. 2014 Challenging the paradigms of deep-sea ecology. Trends Ecol Evol **29**, 465–475. (doi:10.1016/j.tree.2014.06.002)

62. Higgs, N. D. & Attrill, M. 2015 Biases in biodiversity: wide-ranging species are discovered first in the deep sea. Front. Mar. Sci. **2**, 717. (doi:10.3389/fmars.2015.00061)

63. Jouffray, J. B., Blasiak, R., Norström, A. V., Österblom, H. & Nyström, M. 2020 The Blue Acceleration: The Trajectory of Human Expansion into the Ocean. One Earth **2**, 43–54. (doi:10.1016/j.oneear.2019.12.016)

64. Jones, D. O. B., Amon, D. J. & Chapman, A. S. A. 2018 Mining Deep-Ocean Mineral Deposits: What are the Ecological Risks? Elements **14**, 325–330. (doi:10.2138/gselements.14.5.325)

65. Hidalgo, M. & Browman, H. I. 2019 Developing the knowledge base needed to sustainably manage mesopelagic resources. ICES Journal of Marine Science **76**, 609–615. (doi:10.1093/icesjms/fsz067)

66. Jetz, W. & Freckleton, R. P. 2015 Towards a general framework for predicting threat status of data-deficient species from phylogenetic, spatial and environmental information. Philos T Roy Soc B **370**, 20140016–20140016. (doi:10.1098/rstb.2014.0016)

67. González-del-Pliego, P., Freckleton, R. P., Edwards, D. P., Koo, M. S., Scheffers, B. R., Pyron, R. A. & Jetz, W. 2019 Phylogenetic and Trait-Based Prediction of Extinction Risk for Data-Deficient Amphibians. Current Biology **29**, 1557–1563.e3. (doi:10.1016/j.cub.2019.04.005)

68. Faulwetter, S. et al. 2016 EMODnet Workshop on mechanisms and guidelines to mobilise historical data into biogeographic databases. RIO **2**, e9774–28. (doi:10.3897/rio.2.e9774)

69. Mieszkowska, N., Sugden, H., Firth, L. B. & Hawkins, S. J. 2014 The role of sustained observations in tracking impacts of environmental change on marine biodiversity and ecosystems. Philos T R Soc A **372**, 20130339. (doi:10.1098/rsta.2013.0339)

70. Goodwin, K. D., Thompson, L. R., Duarte, B., Kahlke, T., Thompson, A. R., Marques, J. C. & Caçador, I. 2017 DNA Sequencing as a Tool to Monitor Marine Ecological Status. Front. Mar. Sci. **4**, e1002358. (doi:10.3389/fmars.2017.00107)

71. Tahsin, T., Weissenbacher, D., Rivera, R., Beard, R., Firago, M., Wallstrom, G., Scotch, M. & Gonzalez, G. 2016 A high-precision rule-based extraction system for expanding geospatial metadata in GenBank records. J Am Med Inform Assoc **23**, 934–941. (doi:10.1093/jamia/ocv172)

72. Tahsin, T., Weissenbacher, D., O'Connor, K., Magge, A., Scotch, M. & Gonzalez-Hernandez, G. 2017 GeoBoost: accelerating research involving the

geospatial metadata of virus GenBank records. Bioinformatics **34**, 1606–1608. (doi:10.1093/bioinformatics/btx799)

73. Chamberlain, S. In press. bold: Interface to Bold Systems API. R package version 1.1.0. httpsCRAN.R-project.orgpackagebold.

74. Webb, T. J., Lines, A. & Howarth, L. M. 2020 Occupancy-derived thermal affinities reflect known physiological thermal limits of marine species. Ecology and Evolution **75**, 209. (doi:10.1002/ece3.6407)

75. Pinsky, M. L., Selden, R. L. & Kitchel, Z. J. 2020 Climate-Driven Shifts in Marine Species Ranges: Scaling from Organisms to Communities. Annual Review of Marine Science **12**, 153–179. (doi:10.1146/annurev-marine-010419-010916)

76. Costello, M. J., Claus, S., Dekeyzer, S., Vandepitte, L., Tuama, É. Ó., Lear, D. & Tyler-Walters, H. 2015 Biological and ecological traits of marine species. PeerJ **3**, e1201. (doi:10.7717/peerj.1201)

77. Beauchard, O., Veríssimo, H., Queirós, A. M. & Herman, P. M. J. 2017 The use of multiple biological traits in marine community ecology and its potential in ecological indicator development. Ecol Indic **76**, 81–96. (doi:10.1016/j.ecolind.2017.01.011)

78. Kindsvater, H. K., Mangel, M., Reynolds, J. D. & Dulvy, N. K. 2016 Ten principles from evolutionary ecology essential for effective marine conservation. Ecology and Evolution **6**, 2125–2138. (doi:10.1002/ece3.2012)

79. Hiddink, J. G. et al. 2019 Assessing bottom trawling impacts based on the longevity of benthic invertebrates. J Appl Ecol **56**, 1075–1084. (doi:10.1111/1365-2664.13278)

80. Álvarez-Noriega, M., Burgess, S. C., Byers, J. E., Pringle, J. M., Wares, J. P. & Marshall, D. J. 2020 Global biogeography of marine dispersal potential. Nature Ecology & Evolution 2017 1:9 **4**, 1–8. (doi:10.1038/s41559-020-1238-y)

81. Menegotto, A. & Rangel, T. F. 2018 Mapping knowledge gaps in marine diversity reveals a latitudinal gradient of missing species richness. Nature Communications **9**, 4713. (doi:10.1038/s41467-018-07217-7)

82. Ratnasingham, S. & Hebert, P. 2007 BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). Mol. Ecol. Notes (doi:doi: 10.1111/j.1471-8286.2006.01678.x)

Table 1. Data sources used to link different dimensions of diversity across all marine animals.

Dimension of diversity	Data source	Data type	Reference
Taxonomy	WoRMS	Authoritative classification and catalogue of marine taxonomic names	[37]
Functional Groups	WoRMS	Classification of marine species into broad ecological groups	[37]
Biogeography	OBIS	Global database of marine species occurrence records	[22]
Genetics	GenBank	The NIH genetic sequence database, an annotated collection of all publicly available DNA sequences	[3]
Molecular taxonomy	BOLD	Barcode of Life Data System for DNA barcodes	[82]
Conservation status	IUCN Red List	The IUCN Red List of threatened species	[6]

Table 2. Breakdown of 206,849 marine animal species by number of global occurrence records in OBIS, and numbers of nucleotide sequences in GenBank.

Number of OBIS records	Number of GenBank nucleotides										
	0	1	2-10	11-100	101-1,000	1,001-10,000	10,001-100,000	100,001-1,000,000	>1,000,000	*Totals*	*In OBIS*
0	93,519	1,312	4,484	1,164	116	13	15	13	0	*100,636*	
1	16,905	356	1,253	314	33	3	3	2	0	*18,869*	*106,213*
2-10	35,613	1,086	3,714	990	122	17	11	8	0	*41,561*	
11-100	19,998	1,392	5,931	2,733	351	32	30	26	2	*30,495*	
101-1,000	4,274	594	3,334	2,917	512	51	35	37	1	*11,755*	
1,001-10,000	402	86	630	1,113	315	33	53	33	4	*2,669*	
10,001-100,000	42	4	107	406	167	31	22	31	1	*811*	
100,001-1,000,000	2	0	0	14	20	5	3	8	0	*52*	
>1,000,000	0	0	0	0	0	0	1	0	0	*1*	
Totals	*170,755*	*4,830*	*19,453*	*9,651*	*1,636*	*185*	*173*	*158*	*8*	**206,849**	
In GenBank		*36,094*									

Table 3. Species with high numbers of GenBank nucleotide records but few OBIS occurrences, or species with large numbers of OBIS occurrences but few GenBank nucleotides.

Species	Phylum	Class	Functional Group	GenBank Nucleotides	OBIS Records
Olavius algarvensis	Annelida	Clitellata	benthos	173,609	0
Capitella teleta	Annelida	Polychaeta	benthos	208,794	1
Platynothrus peltifer	Arthropoda	Arachnida	other/unknown	106,099	0
Caligus rogercresseyi	Arthropoda	Hexanauplia	other/unknown	628,843	0
Proasellus racovitzai	Arthropoda	Malacostraca	benthos	127,716	0
Proasellus ibericus	Arthropoda	Malacostraca	benthos	150,798	0
Bragasellus molinai	Arthropoda	Malacostraca	benthos	209,419	0
Proasellus beticus	Arthropoda	Malacostraca	benthos	228,033	0
Seriola quinqueradiata	Chordata	Actinopterygii	fish	105,911	6
Theragra finnmarchica	Chordata	Actinopterygii	fish	130,916	0
Takifugu flavidus	Chordata	Actinopterygii	fish	138,301	0
Takifugu rubripes	Chordata	Actinopterygii	fish	466,790	5
Molgula tectiformis	Chordata	Ascidiacea	benthos	106,904	0
Halocynthia roretzi	Chordata	Ascidiacea	benthos	116,123	4
Pelecanus crispus	Chordata	Aves	birds	231,775	0
Balaenoptera acutorostrata scammoni	Chordata	Mammalia	mammals	238,976	0
Emydocephalus ijimae	Chordata	Reptilia	reptiles	157,876	0
Hemicentrotus pulcherrimus	Echinodermata	Echinoidea	benthos	153,541	3
Apostichopus parvimensis	Echinodermata	Holothuroidea	benthos	166,764	1
Apostichopus japonicus	Echinodermata	Holothuroidea	benthos	401,310	4
Cumia reticulata	Mollusca	Gastropoda	benthos	144,517	2
Amphimedon queenslandica	Porifera	Demospongiae	benthos	142,554	9
Thunnus alalunga	Chordata	Actinopterygii	fish	0	114,485
Chrysophrys auratus	Chordata	Actinopterygii	fish	0	104,066

Table 4A Breakdown of marine animal species by functional group and IUCN Assessment status. Listed for each IUCN assessment status are the total number of species per functional group, the number of these species with occurrences in OBIS, and the associated percentage.

	IUCN Assessment Status											
	Not assessed			Data Deficient			Threatened			Non-threatened		
Functional Group	N(species)	N(species in OBIS)	% species in OBIS	N(species)	N(species in OBIS)	% species in OBIS	N(species)	N(species in OBIS)	% species in OBIS	N(species)	N(species in OBIS)	% species in OBIS
Benthos	144,097	73,610	51%	749	530	71%	305	258	85%	1,400	1,206	86%
Zooplankton	5,742	3,027	53%	0	0	-	4	2	50%	2	2	100%
Nekton	3,076	1,878	61%	160	127	79%	7	2	29%	156	151	97%
Fish	8,599	6,161	72%	1,780	1,350	76%	523	457	87%	7,359	6,997	95%
Mammals	66	26	39%	20	17	85%	36	29	81%	70	68	97%
Birds	179	71	40%	1	0	0%	125	92	74%	382	340	89%
Reptiles	20	9	45%	21	14	67%	11	11	100%	44	37	84%
Other / Unknown	31,891	9,725	31%	3	3	100%	10	4	40%	11	9	82%
Totals	***193,670***	***94,507***	***49%***	***2,734***	***2,041***	***75%***	***1,021***	***855***	***84%***	***9,424***	***8,810***	***94%***

Table 4B Breakdown of marine animal species by functional group and presence in the BOLD DNA Barcode database. Listed for species absente from or present in BOLD are the total number of species per functional group, the number of these species with occurrences in OBIS, and the associated percentage.

	In Barcode of Life Database?					
	No			Yes		
Functional Group	N(species)	N(species in OBIS)	% species in OBIS	N(species)	N(species in OBIS)	% species in OBIS
Benthos	131,390	62,316	47%	15,161	13,288	88%
Zooplankton	4,768	2,117	44%	980	914	93%
Nekton	2,506	1,355	54%	893	803	90%
Fish	8,683	5,842	67%	9,578	9,123	95%
Mammals	85	37	44%	107	103	96%
Birds	238	108	45%	449	395	88%

Reptiles	80	59	70%	16	15	94%
Other / Unknown	30,292	8,692	29%	1,623	1,049	65%
Totals	***178,042***	***80,523***	***45%***	***28,807***	***25,690***	***89%***

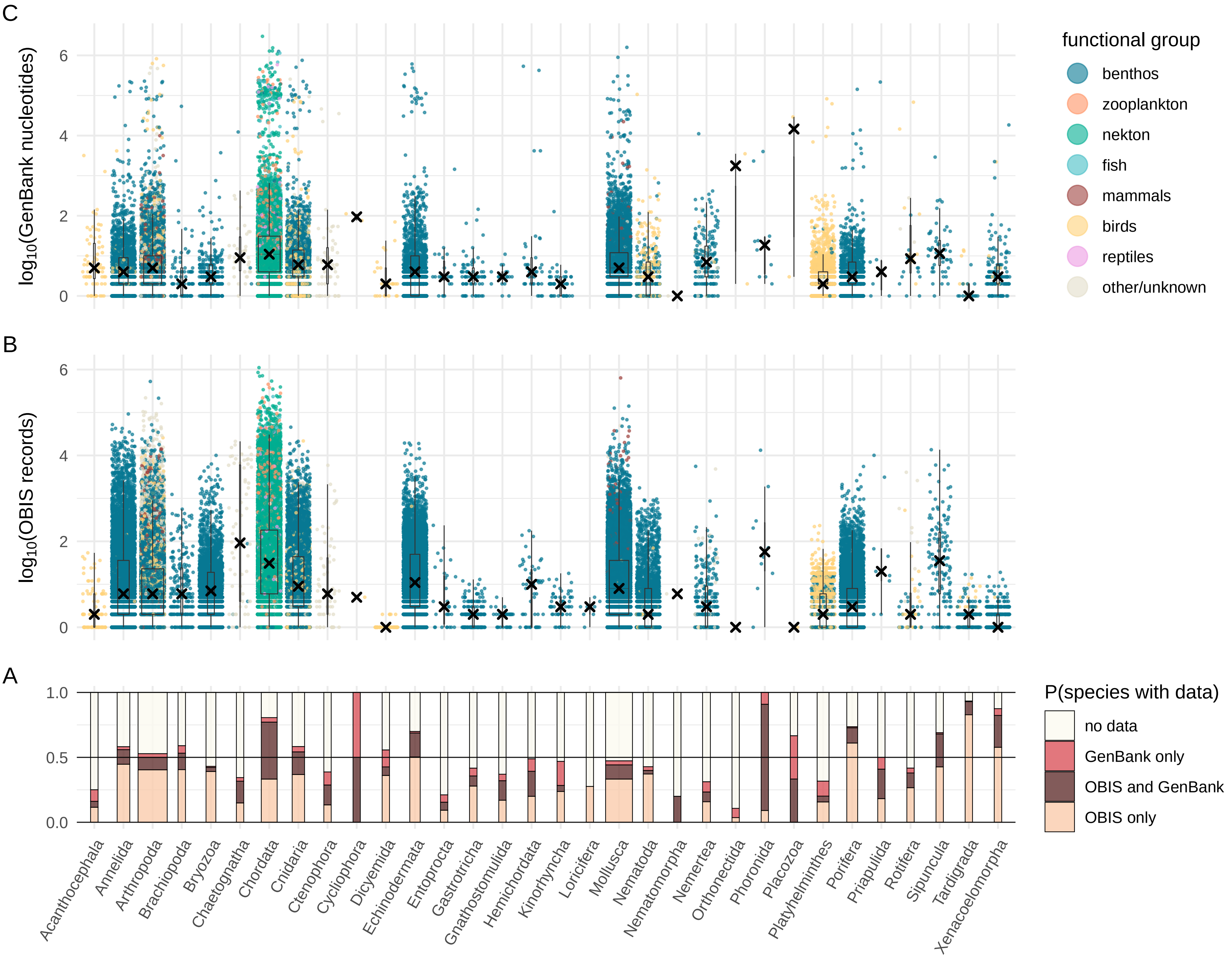
Figure Legends

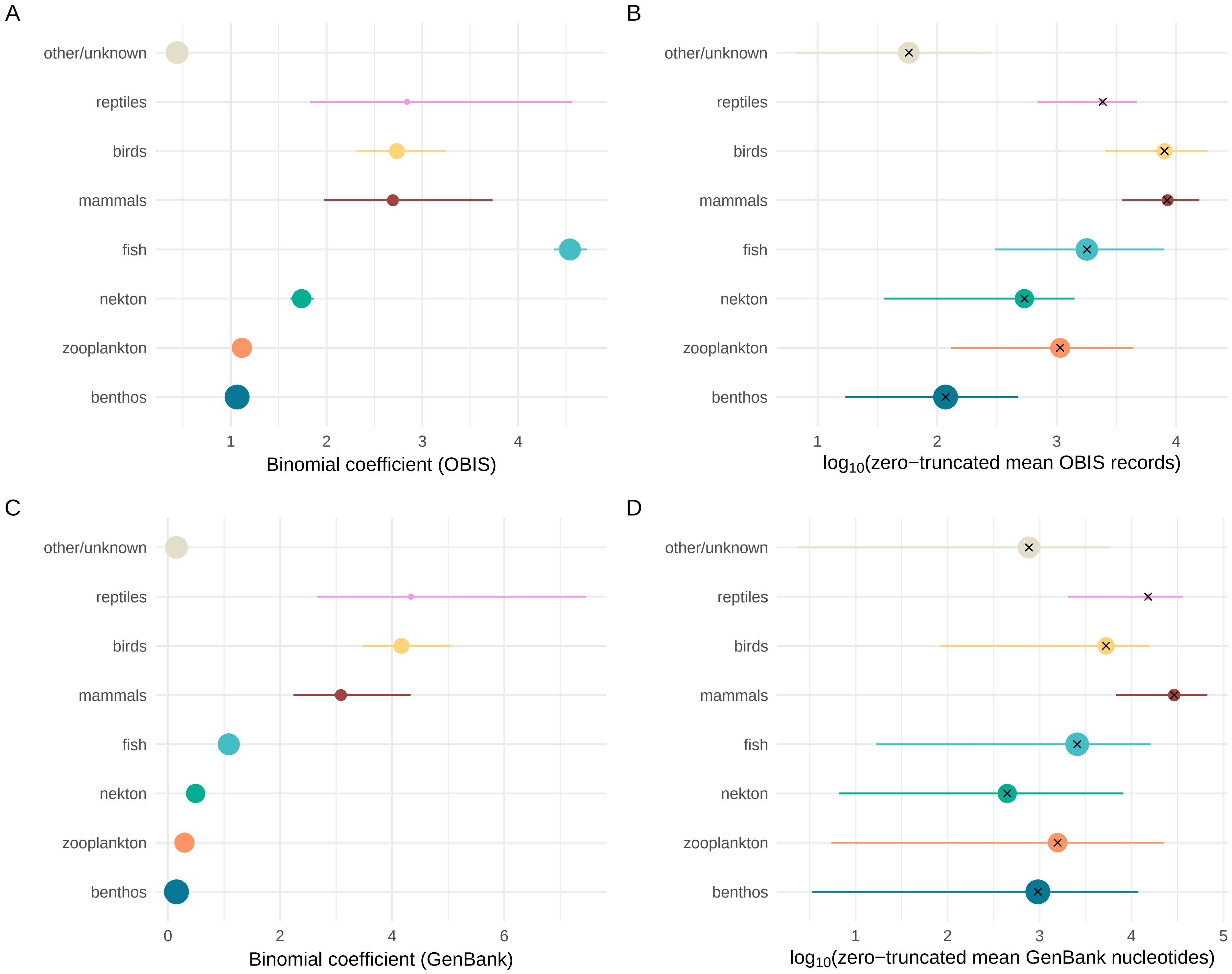
Figure 1. Availability of biogeographic (>45M OBIS occurrence records) and genetic (>56M GenBank nucleotides) data across 206,849 marine animal species, summarised by phylum and by broad functional group. **(A)** Proportion of species in each phylum with data in either database, both databases, or neither. Bar width is proportional to the number of species in each phylum. Number of **(B)** OBIS occurrence records and **(C)** Genbank nucleotide sequences are shown for species that occur in the respective database. Each point represents a species, coloured by functional group. Box plots are superimposed with X marking the median number of records within each phylum. Phylum size varies from 2 species (Cycliophora) to 57,336 species (Arthropoda).

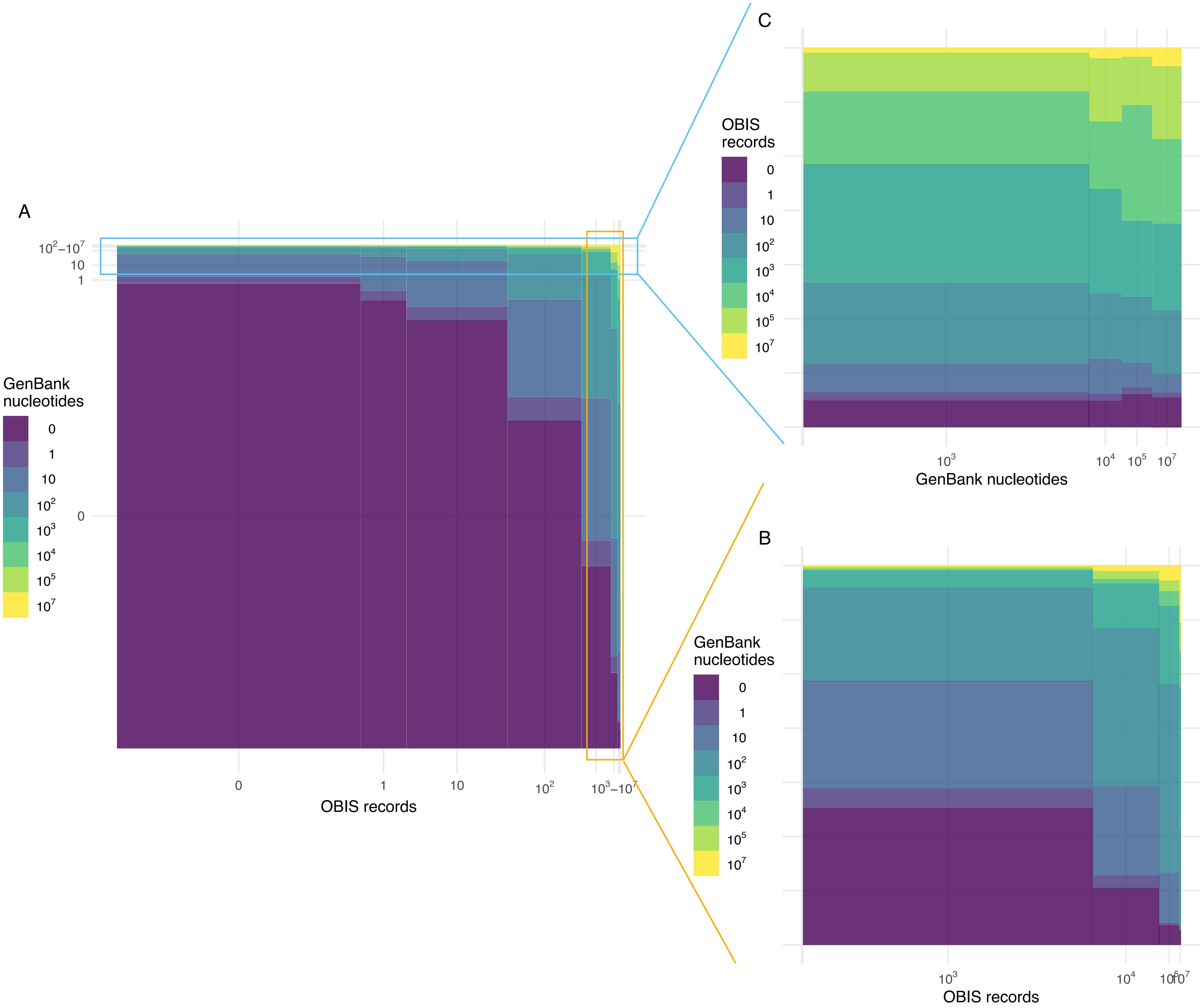
Figure 2. Coefficients from the hurdle models of data availability across functional groups, first modelling presence in a database with a binomial model, and then non-zero counts of records in a database as a negative binomial model. Species presence in OBIS **(A)** or GenBank nucleotide database **(C)** across functional groups is indicated with binomial coefficients (with 95% confidence intervals) on the response scale, representing the ratio of the probabilities of species within a group having records in the database versus not having records in the database. For the subset of species present in **(B)** OBIS or **(D)** GenBank, the empirical mean number of records per species is plotted together with bootstrapped 95% confidence intervals. For each group, the predicted non-zero count from the hurdle model is indicated with an X. Point size is scaled to the total number of species in each functional group (A, C, ranging from 96 reptiles to 146,551 benthos) and to the number of species in each group with records in OBIS (B, 71 reptiles to 75,604 benthos) or GenBank (D, 78 reptiles to 19,235 benthos).

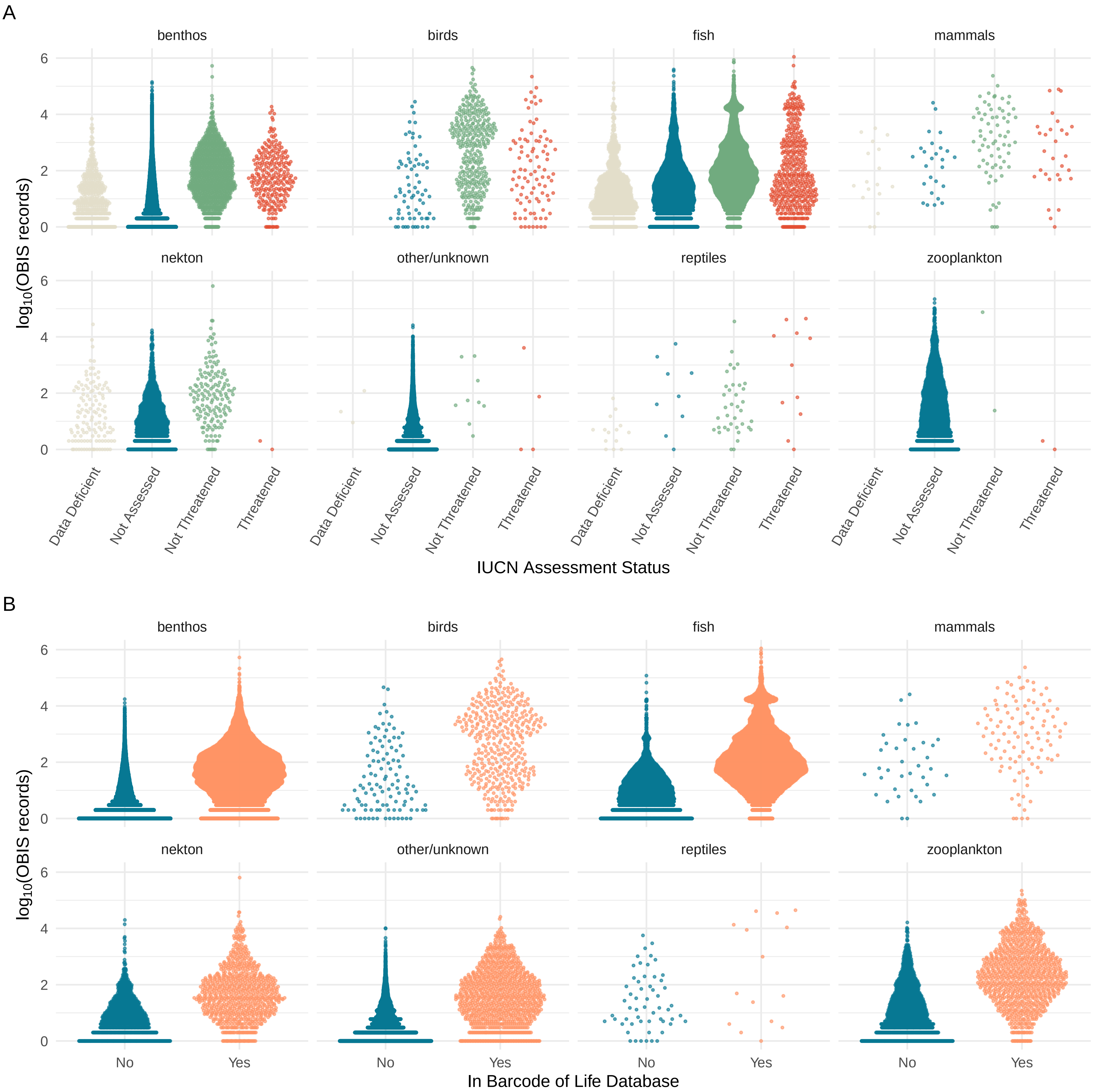
Figure 3. Mosaic plot showing the joint distribution of species between categories of OBIS records and GenBank nucleotides. (A) shows all species, and is dominated by species with no records in either database. (B) zooms in on species with high numbers (>100) of OBIS records, and (C) reverses the axes and zooms in on species with high numbers (>100) of GenBank nucleotides. Axis labels indicate the number of records at the right-hand bound of each category.

Figure 4. Distribution of occurrence records across 106,213 marine animal species present in OBIS by functional group and by **(A)** IUCN assessment status and **(B)** presence in the Barcode of Life Data System. Each point represents a species.


C
log$_{10}$(GenBank nucleotides)
0
2
4
6
functional group
benthos
zooplankton
nekton
fish
mammals
birds
reptiles
other/unknown
B
log$_{10}$(OBIS records)
0
2
4
6
A
0.0
0.5
1.0
P(species with data)
no data
GenBank only
OBIS and GenBank
OBIS only
Acanthocephala
Annelida
Arthropoda
Brachiopoda
Bryozoa
Chaetognatha
Chordata
Cnidaria
Ctenophora
Cycliophora
Dicyemida
Echinodermata
Entoprocta
Gastrotricha
Gnathostomulida
Hemichordata
Kinorhyncha
Loricifera
Mollusca
Nematoda
Nematomorpha
Nemertea
Orthonectida
Phoronida
Placozoa
Platyhelminthes
Porifera
Priapulida
Rotifera
Sipuncula
Tardigrada
Xenacoelomorpha


A
other/unknown
reptiles
birds
mammals
fish
nekton
zooplankton
benthos
1
2
3
4
Binomial coefficient (OBIS)
B
1
2
3
4
$\log_{10}$(zero−truncated mean OBIS records)
C
0
2
4
6
Binomial coefficient (GenBank)
D
1
2
3
4
5
$\log_{10}$(zero−truncated mean GenBank nucleotides)


A
10^2–10^7
10
1
0
GenBank nucleotides
0
1
10
10^2
10^3
10^4
10^5
10^7
0
1
10
10^2
10^3–10^7
OBIS records
C
OBIS records
0
1
10
10^2
10^3
10^4
10^5
10^7
10^3
10^4
10^5
10^7
GenBank nucleotides
B
GenBank nucleotides
0
1
10
10^2
10^3
10^4
10^5
10^7
10^3
10^4
10^5
10^7
OBIS records


A
benthos
birds
fish
mammals
nekton
other/unknown
reptiles
zooplankton
log$_{10}$(OBIS records)
0
2
4
6
Data Deficient
Not Assessed
Not Threatened
Threatened
IUCN Assessment Status
B
benthos
birds
fish
mammals
nekton
other/unknown
reptiles
zooplankton
log$_{10}$(OBIS records)
0
2
4
6
No
Yes
In Barcode of Life Database