

# Comparing biodiversity databases: Greater Caribbean reef fishes as a case study

Iliana Chollett<sup>1\*</sup>  | D. Ross Robertson<sup>2\*</sup>

<sup>1</sup>Sea Cottage, Louisburgh, Co. Mayo, Ireland

<sup>2</sup>Smithsonian Tropical Research Institute, Balboa, Panamá

Corresponding author

Iliana Chollett, Sea Cottage, Louisburgh, Co. Mayo, Ireland.

Email: [iliana.chollett@gmail.com](mailto:iliana.chollett@gmail.com)

**Funding information**

Smithsonian; National Geographic Society

## Abstract

There is a widespread need for reliable biodiversity databases for science and conservation. Among the many public databases available, we lack guidance as to how their data quality varies. Here, we compare species distribution data for a well known regional reef fish fauna extracted from five global online databases that supply “as is” data (GBIF, OBIS, IDigBio, FishNet2 and FishBase) and our own curated regional database (STRI, Smithsonian Tropical Research Institute) using quantitative criteria, and assess how they affect biogeographical analyses. We first describe the databases and quantify overlap between them. We then describe variation in the geographical distributions of species records and species richness and in the completeness of local species lists. Finally, we assess the consequences of using these different databases in biogeographical analyses by comparing patterns of species turnover (beta diversity) and bioregionalization. The databases vary considerably in size and show high overlap in species lists, but low overlap in georeferenced species records. Levels of completeness of local inventories are spatially heterogeneous and low in most databases. Spatial biases produced artefactual variation in patterns of species turnover and delineation of bioregions in all databases. Although not the largest, STRI database has the most complete geographic coverage of data, showed relatively low turnover and the clearest biogeographic regionality. Incorporating data from a wide range of other sources, curating data to reduce errors, and assessing effects of spatial biases in data is critical to obtaining an accurate picture of the geography of biodiversity and its change.

## KEYWORDS

beta diversity, biogeography, completeness, data quality, global aggregators, spatial biases

## 1 | Introduction

There is a general need for comprehensive biodiversity databases for science, conservation and promoting the sustainable use of natural resources. Data on species occurrence that reliably record the name of the species and precise locations where it is found (herein,

biodiversity databases) are central for the study of life on Earth. This information can be used as the basis for ecological studies, including research on species distributions (Elith & Leathwick, 2009) and their changes, for example in relation to climate change (Thomas et al., 2004). Data from biodiversity databases have also been used to assess progress towards targets on conserving biological diversity (Meyer, Kreft, Guralnick, & Jetz, 2015) and help identify management priorities that could allow the sustainable use of natural resources, for

\*Equal authorship.

example, by maximizing productivity through maintaining high biodiversity in grassland ecosystems (Tilman, Wedin, & Knops, 1996).

Currently, digital data are being produced at an accelerated rate, generating an information revolution that is affecting research around the globe. This big-data revolution is also manifested by a wealth of online biodiversity databases. Many biodiversity databases are now available through the internet, and their data are being gathered at a rapid rate with the aid of internet-based data sharing portals and biodiversity information networks. There are scores of international and national efforts that collate biodiversity data. There is, however, no guide that could help navigate the diversity of sources available or let the general user choose the most suitable database to fulfil a specific objective (Soberón & Peterson, 2004). To answer a scientific question, many researchers use only one database, without considering the alternatives or how they compare. It has been long acknowledged, however, that scientists should navigate multiple databases to obtain all information available on a particular species (Thomas, 2009)

Major aggregators that host marine biogeographic data, such as GBIF, began supplying georeferenced species records in 2004. GBIF obtained global reach with huge quantities of data (currently 1.4B records) much more recently. Consequently, few studies have examined the utility of aggregator data, focusing mainly on GBIF because it is the global aggregator of data from many subsidiary aggregators. Soon after GBIF started data release, when it hosted less than 10% of the records it now provides, Yesson et al. (2007) identified problems with geographic accuracy and regional biases in such data. Maldonado et al. (2015) compared geographic variation in species richness in a family of South American insects using data from both GBIF and from a purpose-built, taxonomically curated database, the only published study yet to take such an approach. They found spatial biases in the GBIF data. Robertson et al (Robertson, Visser, & Hui, 2016) and Zizka et al. (2019) developed automated filters that used inconsistencies in record metadata to “clean” georeferenced records from GBIF and other online sources, and the latter authors estimated questionable records at ~4%–7%. Zizka et al. (2020) evaluated a large suite of such filters (with “taxonomic curation” limited to flagging misspelled names) and databases. Their analysis indicated >30% potential error rates that varied among terrestrial and marine taxa and geographic areas in South America. For a diverse marine skate family, they flagged 38.5% of GBIF records, records that covered less than half the species known to occur in a large diversity hotspot, and thus provided a questionable description of its diversity. Here, we extend on those studies by employing automated data cleaning and comparing geographic distribution patterns of a large ecological suite of species within an entire, well-defined marine biogeographic region that are produced by records from various aggregators vs those produced by a purpose-built, taxonomically curated database.

Here, we compare six databases of reef fish species occurrence data in the Greater Caribbean. This fauna has a long history of intensive research, due in large part to its proximity to the USA, and is perhaps the best known among the reef fish faunas of the major tropical marine regions (Floeter et al., 2008). Our comparison

1. INTRODUCTION	001
2. THE SIX DATABASES USED IN THE ANALYSES	003
2.1 Global Biodiversity Information Facility (GBIF, <a href="https://www.gbif.org/">https://www.gbif.org/</a> )	003
2.2 Ocean Biogeographic Information System (OBIS, <a href="https://obis.org/">https://obis.org/</a> )	003
2.3 Integrated Digitized Biocollections (iDigBio, <a href="https://portal.idigbio.org/portal/search">https://portal.idigbio.org/portal/search</a> )	003
2.4 FishNet2 ( <a href="http://www.fishnet2.net/">http://www.fishnet2.net/</a> )	003
2.5 FishBase ( <a href="http://www.fishbase.org">http://www.fishbase.org</a> )	003
2.6 Smithsonian Tropical Research Institute (STRI; <a href="https://biogeodb.stri.si.edu/caribbean/en/pages">https://biogeodb.stri.si.edu/caribbean/en/pages</a> )	004
3. METHODOLOGY	005
3.1 Analyses	005
3.1.1. Attributes of databases	005
3.1.2. Overlap between databases	005
3.1.3. Completeness of databases	006
3.1.4. Assessing the consequences of using different databases in biogeographical analyses	006
4. RESULTS	007
4.1. Attributes of existing databases	007
4.2. Overlap between databases	007
4.3. Completeness of databases	008
4.4. Assess the consequences of using different databases in biogeographical analyses	010
5. DISCUSSION	012
DATA AVAILABILITY STATEMENT	016
REFERENCES	016

includes two major global databases: GBIF and FishBase and STRI (Smithsonian Tropical Research Institute), a regional database. Although GBIF is the largest aggregator of biodiversity data currently available, FishBase has a long history of use within fish biologists and ecologists. As a point of reference for readers not familiar with FishBase, a Google Scholar search in April 2020 for “biodiversity” and “GBIF” provided 19,500 records, while a search for “biodiversity” and “FishBase” provided 14,000 records. STRI is a regional, carefully curated database of fishes in the Greater Caribbean. In the comparisons, we are also including three additional online aggregator databases that supply some data to GBIF (OBIS, iDigBio and FishNet2), and are sometimes used instead of GBIF in research and conservation activities (Costello et al., 2017; Singer, Love, & Page, 2018; VanCompernelle, Knouff, & Ficklin, 2019). We first describe the databases in terms of database size, quantity of taxonomic errors, genus and species richness and abundances. We then quantify the amount of overlap between different databases. We describe the geographic distributions of species records and species richness, and calculate the level of completeness of local species lists provided by the databases. Finally, we assess the consequences of using

different databases in biogeographical analyses. Comparing available biodiversity databases, and understanding their weaknesses, strengths and potential is needed if we are to make informed use of biodiversity data and derive appropriate research conclusions and management advice.

## 2 | THE SIX DATABASES USED IN THE ANALYSES

Information we used in the present analyses came from various databases that provide georeferenced records of occurrence of fish species found in the Greater Caribbean (6–33°N, 57–100°W). Each record consists of a species name and associated latitude and longitude.

### 2.1 | Global Biodiversity Information Facility (GBIF, <https://www.gbif.org/>)

GBIF is an international network and research infrastructure aimed at providing open access to data about all types of life on earth. GBIF works through participant nodes using common standards and open-source tools that enable them to share information. Data from among the 49,000+ data sets hosted by GBIF that were used here range from those on museum specimens collected since the 18th century, to published scientific checklists, to curated local checklists produced by trained science sources such as the Atlantic and Gulf Rapid Assessment Program (<https://www.agrra.org/>), to geotagged smartphone photos (that act as vouchers allowing verification) shared by amateur and scientific naturalists through iNaturalist (<https://www.inaturalist.org/>), to unvouchered, unverified and unverifiable observation records from untrained divers, such as those contributing to DiveBoard (<http://www.diveboard.com>). GBIF data are standardized in Darwin Core format. GBIF data were obtained from a polygon of the study area and subject to taxonomic review after downloading (accessed through the GBIF portal, <https://www.gbif.org/>, on or about 2019-05-19).

### 2.2 | Ocean Biogeographic Information System (OBIS, <https://obis.org/>)

OBIS is a global open-access data and information clearing-house on marine biodiversity (OBIS, 2019) that was adopted as a project of the Intergovernmental Oceanographic Data and Information Exchange of the Intergovernmental Commission of UNESCO. Its range of sources is similar to that of GBIF. OBIS hosts data from organizations or programmes that join it as one of 13 “nodes,” and harvest the data from the IPT (Integrated Publishing Toolkit), where providers publish their data. The IPT is developed and maintained by the GBIF, and OBIS is a major contributor of marine data to GBIF. Data are standardized in Darwin Core format. OBIS data

were obtained for the region of study by downloading data on each family, then retaining only data inside the study area, which were then subject to taxonomic review and selection (accessed through the OBIS portal, <https://obis.org/>, on or about 2019-05-19).

### 2.3 | Integrated Digitized Biocollections (iDigBio, <https://portal.idigbio.org/portal/search>)

iDigBio is sponsored by the US National Science Foundation and run by the University of Florida. It provides digital data from public, non-federal, US collections. Data are standardized in a Darwin Core format, and provided “as is.” iDigBio joined the GBIF network in 2017. iDigBio records were downloaded from a polygon of the region of study and subject to taxonomic review and selection (accessed through the iDigBio portal, <https://portal.idigbio.org/portal/search>, on or about 2019-05-19).

### 2.4 | FishNet2 (<http://www.fishnet2.net/>)

FishNet2 is a collaborative effort that aggregates data on fish collections around the world to share and distribute data on specimen holdings from ~75 museums, universities and other institutions. FishNet2 distributes data in Darwin Core, and data are provided “as is.” FishNet2 is part of the network VerNet, which has contributed to GBIF since 2013 and became part of iDigBio in 2016. While FishNet2 has made substantial efforts to georeference location-record data it hosts, many hosted records still lack georeferencing. FishNet2 data were obtained from a polygon of the study area and subject to taxonomic review after downloading (accessed through the Fishnet2 Portal, [www.fishnet2.net](http://www.fishnet2.net/), 2019-05-19).

### 2.5 | FishBase (<http://www.fishbase.org>)

FishBase is a global biodiversity information system supervised by a consortium of nine non-USA international institutions, which hosts data on fin fishes and elasmobranchs (Froese & Pauly, 2009). Information presented in FishBase is extracted from the scientific literature, reports and museum or aggregator (GBIF) databases, and standardized by a team of specialists. FishBase was originally conceived as a fish encyclopaedia: a repository of information on the taxonomy, biology and ecology of fishes (Pauly & Froese, 1991). However, GBIF hosts 173,000 occurrence records of fishes provided by FishBase, and FishBase is routinely used as a source of information for biogeographic analyses of both global biogeographic analyses (Ready et al., 2010) and regional distributions (Sandin, Vermeij, & Hurlbert, 2008). Thus although it provided few relevant records for the present study, its inclusion provides a useful perspective for the scientific community. Data from FishBase were downloaded for the following ecosystems: Caribbean Sea, Gulf of Mexico, Southeast U.S. Continental Shelf, Atlantic Ocean, Sargasso Sea and Bermuda,

and subject to taxonomic review and selection after downloading (2019-05-19).

## 2.6 | Smithsonian Tropical Research Institute (STRI; <https://biogeodb.stri.si.edu/caribbean/en/pages>)

The STRI database was compiled by DRR and Ernesto Peña at STRI's Naos Marine Laboratory and represents almost 20 years accumulation of curated data (see below) from the following sources: data downloaded at roughly two year intervals from the five aggregators; data from online databases of various museums that supply aggregators (data directly downloaded from a museum sometimes differs from that available in an aggregator from the same museum), including the Swedish Museum of Natural History, the American Museum of Natural History, the Natural History Museum of Denmark, the Gulf Coast Research Laboratory, the Colombian Museum of Natural Marine History, the United States National Museum and the United States Geological Survey; data from national aggregators of Colombia (Sistema de Información Sobre Biodiversidad de Colombia (<https://sibcolombia.net/>), and Sistema de Información Ambiental Marina de Colombia, <https://siam.invemar.org.co/>), Mexico (La Comisión Nacional para el Conocimiento y Uso de la Biodiversidad, CONABIO; <http://www.conabio.gob.mx/informacion/gis/>), and Costa Rica (Museo de Zoología de la Universidad de Costa Rica, <http://museo.biologia.ucr.ac.cr/>); verified (by DRR) underwater photographs of fishes taken at known locations; peer reviewed publications containing location information (species descriptions; taxonomic revisions of species, genera and families; regional and local checklists); fisheries reports; digital tagging data for species such as elasmobranchs; diving surveys and collections of local faunas by DRR (Robertson, Domínguez-Domínguez, Aroyo, Mendoza, & Simões, 2019). In addition, selected data from two sources that collect species lists at sites scattered throughout the Greater Caribbean are incorporated: from the Atlantic and Gulf Rapid Reef Assessment programme (AGRRA, <https://www.agrra.org/>; Kramer & Lang, 2003) and from trained citizen scientists who contribute data on fishes to the Reef Environmental Education Foundation's database (REEF: Pattengill-Semmens & Semmens, 2003). The bibliographic module (<https://biogeodb.stri.si.edu/caribbean/en/library>) of Robertson and Van Tassell (2019) contains ~1,700 publications linked to species names, among them the publications from which location data were extracted.

Data from the aggregators are presented "as is" and the aggregators themselves do no data curation. Duplicates (and occasionally triplicates and quadruplicates) of the same museum record often are included from multiple sources (e.g., the original museum source, derivative checklists, an aggregator), sometimes with slightly different georeferenced coordinates. Data available in one year may subsequently disappear from an aggregator, and different data may be available for the same species under different names (e.g., both old and new names when a species is reassigned to another genus). Errors, sometimes large errors (Robertson, 2008), are common

in aggregator data, from museums as well as other sources, and longstanding errors can seem to take on a perpetual existence. For example, the damselfish Sergeant major (*Abudefduf saxatilis*, Pomacentridae) is a common and widespread inhabitant of tropical reefs on both sides of the Atlantic. Although it was once regarded as pantropical, it is now known to be restricted to the Atlantic Ocean. Despite the fact that its taxonomic status and range were resolved ~30 years ago (e.g. see Allen, 1991), museum data presented by all five aggregators (accessed December 10, 2019) that contributed to the multi-source database used in this study currently show large numbers of records of this species throughout the entire tropical Indo-Pacific, as well as across its native range in the Atlantic. Since many of the databases accumulating on aggregators are derivative (lists derived from records and from other derivative lists), it will become increasingly difficult to eliminate such errors as corrections to data in primary sources do not automatically propagate through the chain of usage by different databases. Due to increasing limitations on resources for taxonomic work, museums themselves have difficulty dealing with errors in specimen identity and location, and old specimens become unidentifiable, specimens never get returned when loaned out, or simply vanish, and entire collections can get destroyed by hurricanes or fires, or get dumped when museums close or experience a major change in mission. Georeferenced location data on fish distributions in the neotropics (and presumably most other areas) hosted by aggregators, particularly GBIF and OBIS, which take data from a broad range of source types, might best be described as messy, and the significant potential for errors in location records and an inability to verify records always needs to be taken into account when incorporating data from aggregators, primary museum sources and analog sources.

Data considered for inclusion in the STRI database were screened as follows to exclude questionable records. Data from two databases hosted by OBIS and GBIF were excluded entirely due to lack of reliability: BioGoMx (<https://www.gulfbase.org/project/biodiversity-gulf-mexico-biogomx-database>) and Diveboard (<http://www.diveboard.com>). The only REEF data used were from "expert" REEF recorders on readily identifiable species that are unlikely to be confused with similar species (e.g. data for some genera of sparids, gerreids, labrisomids and gobies that include various sympatric species with very similar appearances, were not used). After data from aggregators and museum sources were combined into a single database, duplicate records were filtered out by rounding all records to three decimal places and eliminating duplicates, a process that inevitably deleted some valid records as well as duplicates. The sizes of the databases and abundance of such duplicates precluded individual manual review and exclusion. Finally, all location data for each species were revised by DRR by examining the distribution of its georeferenced coordinates overlaid on a digital map of the current known distribution range of that species (for such range information see Carpenter & De Angelis, 2002; Ebert, Fowler, & Compagno, 2013; Last et al., 2016; Robertson & Van Tassell, 2019; and IUCN Redlist species accounts for most species considered here: <https://www.iucnredlist.org/search>). Such revision took into account recent

modifications to taxonomy and distributions due to new data and new publications, or as a result of discussions between DRR and experts in the taxonomy of particular species or genera. Source information of many individual questionable records provided by aggregators together with the hosted data was inspected to try and assess their validity. Records thought likely to be erroneous were deleted. Those included inexplicable records lacking adequate documentation located well outside the known distribution range, and records in unlikely habitats (e.g., on land for marine species; in deep water for shallow-water species). This revision process reduced the number of records by about 30%.

Construction of the STRI database started almost 20 years ago, and it has continued to grow through the addition of more species and location data, with the intention of providing the most accurate and current picture of the distributions of marine fishes in the Greater Caribbean. The entire location-record database is fully available and queryable online through a website (<https://biogeodb.stri.si.edu/caribbean/en/pages>). Data and analytical challenges preventing the transfer of STRI data into global databases such as GBIF include the lack of use of international data standards and protocols during data encoding and lack of standardized accompanying metadata for many records (see Feagraus, Andelman, Jones, & Schildhauer, 2005; Wieczorek et al., 2012). The STRI database started well before the global aggregator era and before the implementation of international standards for sharing information on biological diversity. Persistent identifiers to link a record to its source are lacking both for many records entered in the early stages of its construction, when such identifiers were not available, and for many individual records that are based on photographs or field surveys at specific locations that have been gathered over the years through personal communications. Addressing these issues would require major upgrades to the database, for which resources are not currently available.

Data from the five individual aggregator databases that are used in the comparisons described here were all downloaded from their online portals during May, 2019. However, data from those five aggregators that were incorporated in the STRI database were downloaded in March 2017, with data from other sources described above added to the STRI database intermittently between then and May 2019, when the entire dataset was curated as described above. Hence, the five individual aggregator databases analysed in this study undoubtedly contain additional data not included in the version of the STRI database used in the present analyses.

Only reef-associated fish species were included in the present analysis. These include demersal species known to occur on hard bottoms (coral, rock and oyster substrata), or on rubble, sand and vegetated bottoms within and around the immediate fringes of reefs, and pelagic species regularly found on reefs. All exotic, non-resident and species other than reef-associated fishes were excluded from all databases prior to comparisons. Non-residents were defined as widespread species only seen in the study area rarely. Reef fish assemblages dominated by shallow-water taxa are found in the waters of continental and insular shelves, that is, between 0 and 200 m (Baldwin, Tornabene, & Robertson, 2018). We used the shelf edge

as a breakpoint and excluded records in areas deeper than 200m, identifying those areas using the General Bathymetric Chart of the Oceans (GEBCO Compilation Group, 2019; Kapoor, 1981).

Before the analyses, for all databases, duplicate records were deleted. Subsequently, records in the Pacific or on land were deleted. We used the Global Self-consistent, Hierarchical, High-resolution Geography Database (Wessel & Smith, 1996) to identify these areas.

## 3 | METHODOLOGY

### 3.1 | Analyses

#### 3.1.1 | Attributes of databases

All databases were described using the metrics: (a) database size; (b) errors in species names; (c) number of genera; (d) number of species; (e) median number of records per species. The size of each database was estimated as the total number of records. The number of synonyms and errors in species names was quantified to flag both outdated and incorrect naming due to spelling errors by matching species names to those in Eschmeyer's Catalog of Fishes (Eschmeyer, Fricke, & Van der Laan, 2015), which is updated monthly. No data were excluded due to name issues.

After any corrections in species names were done to the database, the total number of genera and species present were quantified. We calculated the median and interquartile range of the number of records as a proxy for abundance. The interquartile range is a measure of statistical dispersion equal to the difference between the 75th and 25th percentiles of the data and is commonly used to describe datasets that are not normally distributed. Histograms of number of records for each database are shown in Figure S1.2.

For each of the six databases, the geographical distribution of the abundance of records per one-degree cell and species richness per one-degree cell were calculated and mapped.

#### 3.1.2 | Overlap between databases

Overlap between different databases was quantified both at the level of species lists and location records. A specific record consists of species name and associated latitude and longitude. Because different databases use a different number of significant figures when storing decimalized coordinates, we rounded every coordinate to two decimals (the minimum denominator among databases) before matching records. We are aware that matching records this way might provide some misleading results: it is possible that some matches are artificial because of the coarse coordinates; it is also possible that we deleted some true matches if data providers handle coordinates differently (e.g. if they truncate decimals instead of rounding them).

The overlap between databases was displayed as a two-dimensional matrix showing the percentage of overlap between database

$i$  and database  $j$ . The metric of overlap indicates the percentage of species or records from database  $i$  ( $x$  axis) observed also in database  $j$  ( $y$  axis). This matrix is not symmetrical, that is the percentage of species contained in the database  $i$  that are also in  $j$  is not the same than the percentage of species in  $j$  that are also in  $i$ .

### 3.1.3 | Completeness of databases

All databases contain point records from different sites, and survey effort in each database is uneven across the region of study. This introduces uncertainty into biodiversity assessments, because it is difficult to know if a species is actually absent or it was missed because sampling intensity was not high enough. We quantified how complete species lists were in each database using species accumulation curves (Soberón & Llorente, 1993). In a species accumulation curve, the number of recorded species is related to the number of sampled sites within a cell (a surrogate of survey effort). “Completeness” was calculated as the percentage difference between the observed number of species in a cell and the predicted, asymptotic value. Higher values of completeness imply more reliable inventories. Species accumulation curves were quantified using subsampling without replacement (Gotelli & Colwell, 2001), and the rational function, with the form  $y = (a + bx)/(1 + cx$ ; Ratkowsky, 1990), was used to calculate asymptotic extrapolated values. The asymptote of this function can be found by dividing the numerator’s leading coefficient and the denominator’s leading coefficient (i.e.,  $b/c$ ).

We first quantified one overall species accumulation curve and completeness value for each database. Accumulation curves were assumed to be stable when the increase between consecutive points was very low. Here, a threshold of 0.3 in the slope between consecutive points was used to identify this value. We then calculated variability in completeness of sampling throughout the study area for each database by fitting one species accumulation curve to each one-degree cell. One degree has been suggested as a reliable grid resolution that captures species distributions based on point-record data that have not been collected systematically (Hawkins, Rueda, & Rodríguez, 2008; Hurlbert & Jetz, 2007). For all completeness analyses, a cut-off value of one was used, that is, if the slope is higher than one, it is assumed that the spatial units were not sampled enough and completeness cannot be reliably calculated (Lobo et al., 2018).

### 3.1.4 | Assessing the consequences of using different databases in biogeographical analyses

To assess how use of the six databases affected the results of biogeographical analyses, we first looked at ordinations, then at patterns of species turnover (beta diversity) and then at bioregionalizations of the data (Kreft & Jetz, 2010). For these three analyses, we used species presence/absence data per one-degree cell.

Spatial variability in species composition per cell was visualized using classical multidimensional scaling (MDS). This analysis collapses the information from multiple dimensions into just two, so they can be easily visualized and interpreted. The Simpson’s index (beta sim) was used as a metric of dissimilarity and input to the ordination. This index is based on presence–absence and describes spatial turnover without the influence of richness gradients (Baselga, 2010). MDS ordinations were related to the most commonly used classifications of marine regions available: marine ecoregions of the world (Spalding et al., 2007). This was done by matching the colour of the ecoregion in the map to the colour of the dots in the ordination plots and comparing the distributions of dots from each ecoregion in the MDS plot for each database. For each database, we also assessed the statistical significance of the differences among species compositions in each of Spalding’s ecoregions using a pairwise PERMANOVA test and the Bonferroni correction for multiple tests.

Species turnover was calculated in each one-degree cell throughout the study area. To this end, mean community dissimilarity was calculated for each cell within a moving window of three neighbouring cells using Sorensen index as a measure of turnover (Laffan et al., 2016). Turnover cannot be calculated in isolated cells lacking neighbours.

We used each database to identify bioregional subdivisions of the study area and assess the consequences of using different databases in biogeographical analyses. Those bioregions were determined using “Infomap Bioregions,” a methodology developed by Vilhena and Antonelli (2015) and Edler, Guedes, Zizka, Rosvall, and Antonelli (2017). This method was developed precisely to minimize issues arising from unevenly spread species distributions records such as those in the databases considered here, and is claimed to represent a better method than clustering analyses or bioregionalizations using species distribution models (Vilhena & Antonelli, 2015). This method has outperformed approaches that use unipartite networks as inputs (Vilhena & Antonelli, 2015) or different clustering approaches (Aldecoa & Marín, 2013). However, unlike clustering, the output of Infomap Bioregions is not hierarchical, and it does not identify subdivisions of the bioregions it defines. Briefly, Infomap Bioregions bins species records consisting of species name, latitude and longitude, into discrete geographical grid cells. Data density per cell determines the spatial output resolution, with coarser grid cells in areas with sparser data. Infomap Bioregions then extract a bipartite network that includes both species and grid cells, and clusters the network using an information-theory clustering algorithm known as Infomap (Edler & Rosvall, 2015). To avoid creating regions with too few data points as well as very large regions, cell sizes were set to range between one and four degrees, and cell capacity between 10 and 100 species. This is the default, recommended parameterization for this analysis by Edler et al. (2017).

Bioregional maps produced by Infomap Bioregions were compared using Mapcurves (Hargrove, Hoffman, & Hessburg, 2006). This quantitative method compares the spatial overlap between categorical maps and summarizes the results with a global goodness of fit score. Goodness of fit was calculated along a grid with 0.1 degrees spacing over the study area.

All analyses were conducted in R. Packages “sp,” “rgdal,” “raster” and “rgeos” were used to handle spatial data (Bivand, Keitt, & Rowlingson, 2019; Bivand & Rundel, 2019; Hijmans, 2019; Pebesma & Bivand, 2005). Species accumulation curves were calculated using the package “vegan” (Oksanen et al., 2018) and dissimilarities in species composition of one-degree cells using the package “betapart” (Baselga, Orme, Villeger, Bortoli, & Leprieur, 2018). Beta diversity was calculated with the package “speciesRaster” (Title, 2017). Completeness was calculated using the package “KnowBR” (Lobo et al., 2018). The library “sabre” was used to compare categorical maps (Nowosad & Stepinski, 2018). The library RVAideMemoire (Hervé, 2019) was used to perform pairwise PERMANOVA tests.

## 4 | RESULTS

### 4.1 | Attributes of existing databases

FishBase and FishNet2 are the smaller data sets, with FishBase having only 2,793 records. GBIF is the largest dataset with 478,410 records (Table 1). There were very few inconsistencies in species names with FishNet2 and iDigBio having the most, about 1.5% in each database. Most inconsistencies are due to use of synonyms rather than misspellings. STRI had no name errors (Table 1).

STRI contains the largest number of species (842) and genera (318). iDigBio and GBIF also contain many genera (314 and 313 respectively) and species (801 and 795). FishBase contains the smallest numbers of both those taxa (225 and 452; Table 1). There is no relationship between the size of the data set and the number of genera or species it contains (Pearson correlation;  $R = 0.51$ ,  $p$ -value = .30 for genera,  $R = .52$ ,  $p$ -value .29 for species,  $n = 6$ ).

The median number of occurrences per species varied considerably among databases, ranging from only two records per species (FishBase) to 79 (STRI, Table 1, see Fig. S1.1 in Supporting information for a detailed description of these differences in terms of histograms of occurrences for each database).

In all databases, the greatest concentration of records is on the US shelf and the lowest concentration in the middle of the region, in a broad band running from the Nicaragua/Honduras

shelf, to Cuba, Hispaniola and the southern Bahamas (Figure 1). STRI has the most even coverage of records across the region and FishBase the least even. FishNet2, GBIF and iDigBio have similar patterns of coverage, which are more evenly distributed than that of OBIS (Figure 1).

STRI has the greatest number (22) of one-degree cells with more than 300 species. FishBase has no such cells, FishNet2 and OBIS have five, iDigBio has seven and GBIF has eight. Apart from FishBase, all databases show highest species richness in SE Florida and Puerto Rico. Other areas with high-richness cells in multiple databases include Meso-America, Panama, Colombia, the lesser Antilles and the northern Bahamas. STRI is the only database that has high-richness cells dispersed throughout the continental shelf and islands of most of the study area, except for the northern Gulf of Mexico, and the northeast and southeast fringes of the study area. OBIS is the only database that has high-richness cells scattered around the periphery of the northern Gulf of Mexico. None of the databases has high-richness cells in the central area that has the lowest concentration of records (Figures 1, 2).

### 4.2 | Overlap between databases

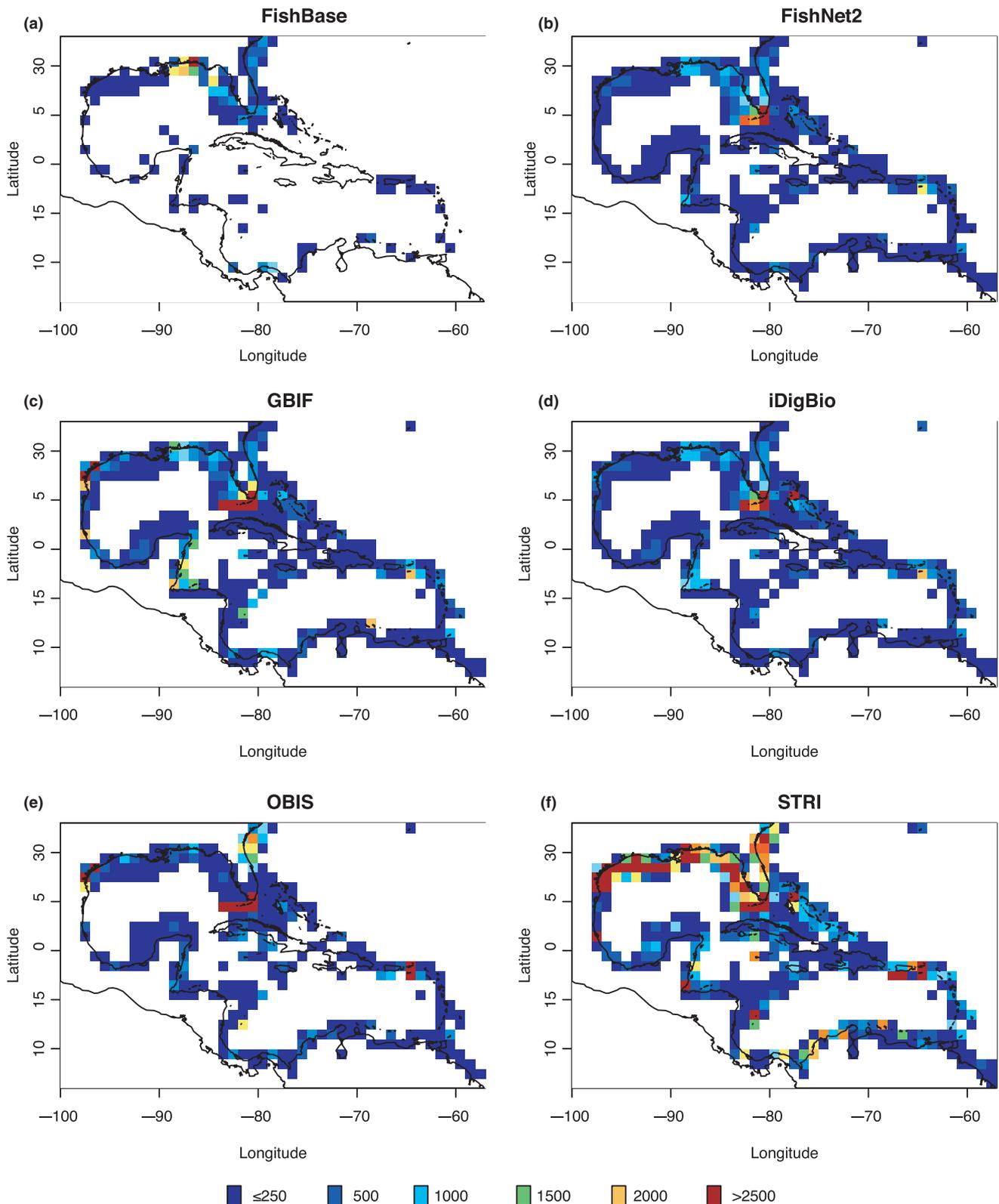
All databases except FishBase are similar in terms of species lists, and 84–100% of species overlap, while overlap between Fishbase and the others ranged from 53%–60% (Figure 3a). The list of species in FishNet2 is fully contained in iDigBio. STRI almost (>99%) fully contains species lists from all databases. OBIS, iDigBio and FishNet2 contribute a large proportion of species to GBIF (98–99%), but their species lists are not fully included in this larger aggregator.

Databases are more dissimilar in terms of records (Figure 3b). While a large percentage of records in iDigBio, FishNet2 and OBIS are contained in GBIF (85%, 81% and 73%), and overlap among the other three is relatively high, GBIF still lacks about 20% of the records found in each of those databases. In contrast, more than 90% of STRI records are found in no other database and it contains only 84% of FishBase records, and shares only 6%–32% of the records found in iDigBio, FishNet2, OBIS and GBIF.

**TABLE 1** Attributes of the databases.

Database	N	Synonyms (%)	Errors (%)	Genera	Species	Occurrence
FishBase	2,793	0.041	0.030	225	452	2 ± 5
FishNet2	51,149	1.486	0.036	305	761	30 ± 75
GBIF	478,410	0.000	0.040	313	795	68 ± 1389
iDigBio	65,345	1.708	0.042	314	801	38 ± 102
OBIS	152,993	0.002	0.008	302	709	31 ± 192
STRI	253,443	0.000	0.000	318	842	79 ± 264

Notes: Number of records (N), percentage of outdated species names (Synonyms), percentage of errors or misspellings in species names (Errors), total number of genus (Genera), total number of species (Species), median and interquartile range of number of occurrences per species (Occurrence). Note: STRI contains data on seven currently unnamed (undescribed) species that are not present in any other database.

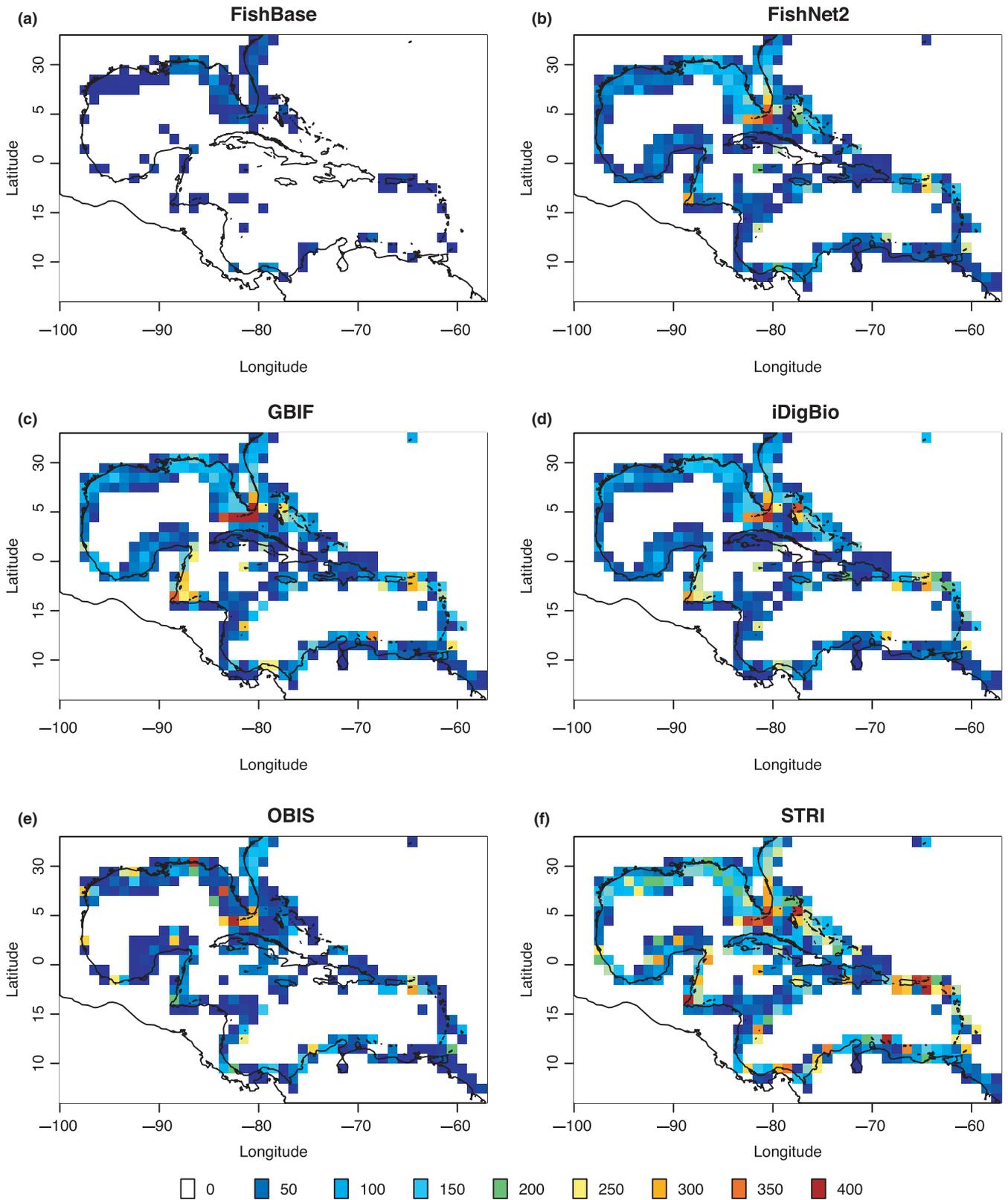


**FIGURE 1** Abundance of records per one-degree cell throughout the study area in six different databases. Colour figure available online.

### 4.3 | Completeness of databases

The number of species records from different one-degree cells varied between and within each of the six databases, and species

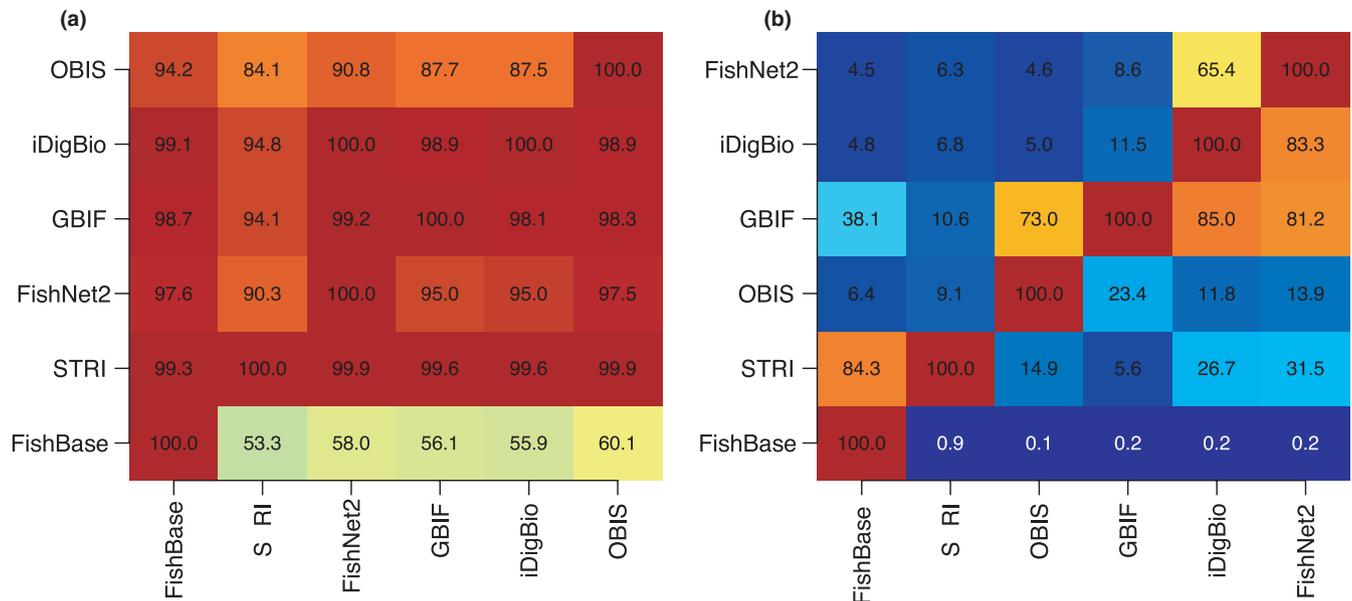
richness increases slowly with increasing numbers of cells (Figure S1.2). Species richness never stabilizes for FishBase but reaches steady values around 150 cells for all other databases. Levels of completeness are lowest for FishBase (maximum 71.5%), highest in



**FIGURE 2** Species richness in one-degree cells across the study area from six databases. Colour figure available online.

STRI (98.7%), and also achieve similar levels in GBIF, OBIS, FishNet2 and iDigBio (97.0, 97.3, 97.9 and 98.5%, respectively: Figure S1.2). For most databases, sampling seems adequate to describe the regional species list in the Greater Caribbean.

Levels of faunal completeness of individual one-degree cells were heterogeneous within databases (Figure 4). The average level of completeness was highest in STRI (55%), which also had the highest percentage of cells with data (97.4%, Table S1.1). Areas with



**FIGURE 3** Variability in (a) species lists and (b) records from each database. Databases are ordered according to similarity. Values and shade indicate the percentage of species from database *i* (x axis) contained in database *j* (y axis). So, for example, 100% of species in the FishNet2 data set are also in the iDigBio data set; but only 95% of the species in iDigBio are represented in FishNet2. Diagonals indicate 100% overlap. Colour figure available online.

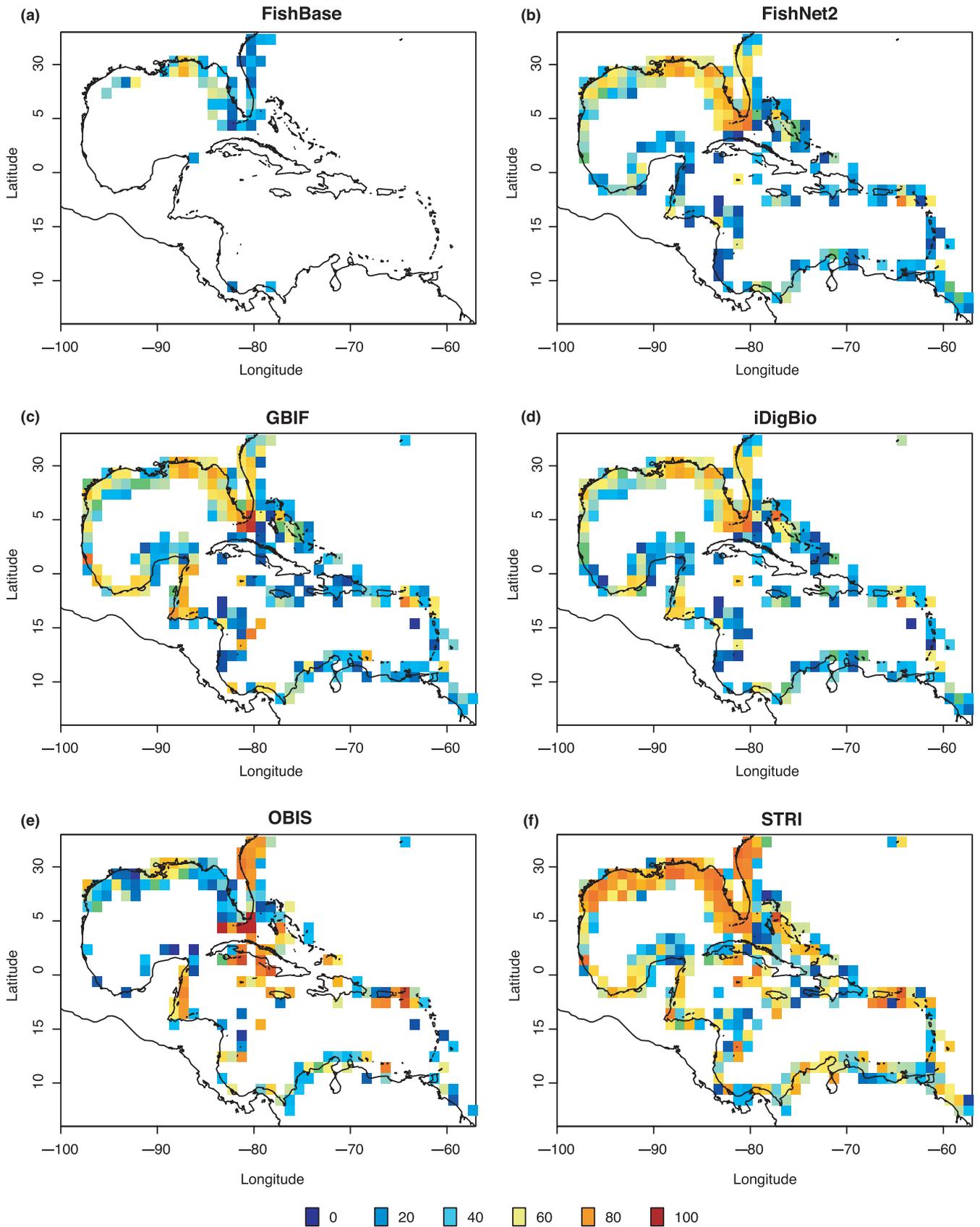
relatively complete inventories in all databases (except FishBase) include SE Florida and the Puerto Rico plateau. For GBIF, OBIS and STRI, the Meso-American Barrier Reef system also has a relatively complete inventory, while in GBIF, OBIS and STRI, the northeast Gulf of Mexico also has a relatively complete inventory. In the STRI database, levels of completeness are high in large areas of moderate richness across most of the Gulf of Mexico and NE Florida (Figure 2). STRI has the widest distribution and greatest abundance of cells with relatively complete inventories (Figure 4), followed by GBIF, then OBIS and iDigBio and then FishNet2. Areas that lack complete inventories in all databases include those with low sampling intensity: much of Campeche Bank, the Nicaragua/Honduras shelf and parts of Cuba, Hispaniola and the Bahamas (Figure 4). A table with values of completeness in each grid cell for each database can be found in the Supporting information (Table S1.1).

#### 4.4 | Assess the consequences of using different databases in biogeographical analyses

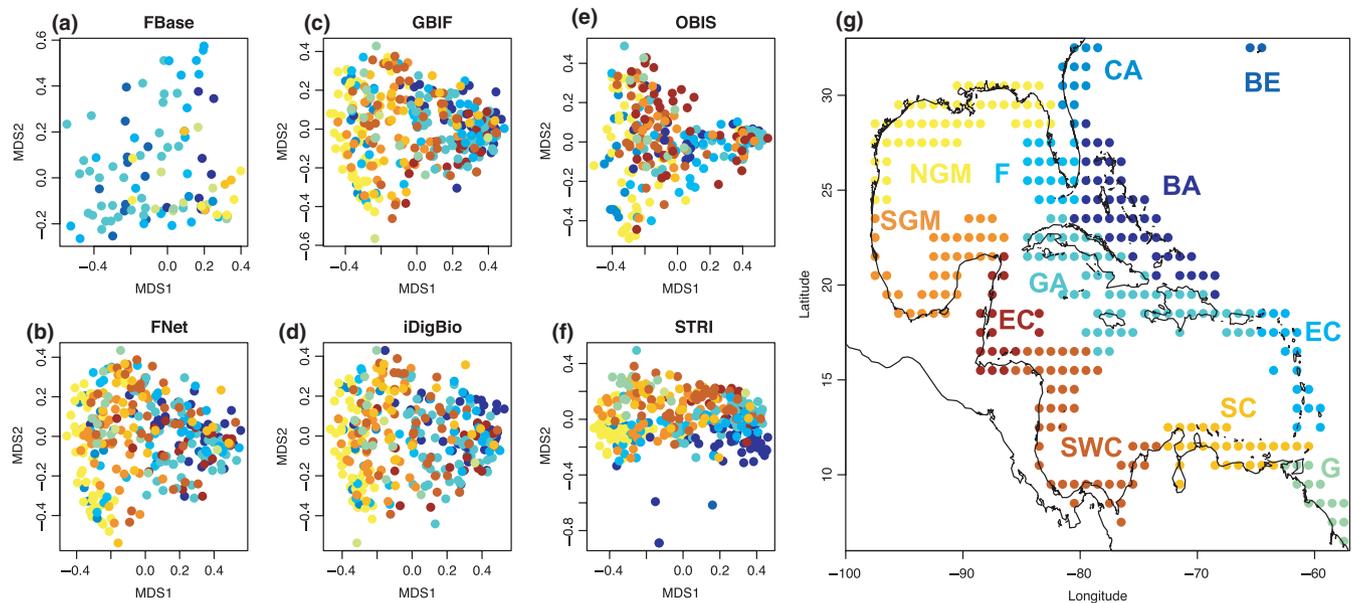
Metric multidimensional scaling (MDS) ordinations allowed visualization of geographic variation in the taxonomic composition of species assemblages in different one-degree cells. MDS ordinations of Greater Caribbean fish assemblages showed a continuous pattern, and sites do not form well-defined, separated groups along the bi-dimensional plane (Figure 5). Different ecoregions in the Caribbean do not form distinct groups as it would have been expected if each had a distinctive fauna. Rather, they are all well mixed in the ordination plots, and differences between ecoregion faunal lists are only significant for a few northern regions in the FishBase data set (PERMANOVA,  $p < .05$ , Table S1.2).

Maps of species turnover permit the identification of any homogeneous areas separated by boundary areas with high levels of compositional change. If that was the case with the Spalding et al.'s (2007) 12 ecoregions, one would expect to see clusters of cells with low species turnover within bioregions surrounded by areas of high-turnover cells separating different bioregions. However, that was not the case with any database. The entire Greater Caribbean is very heterogeneous, and most one-degree cells have relatively high species turnover and are different, often very different, from each other (Figure 6). The northern Gulf of Mexico is the only region that is relatively homogeneous, with similar assemblages spread across that area for all databases. In FishNet2, GBIF, iDigBio and STRI, other areas of that Gulf, Atlantic USA, the Meso-American Barrier Reef system and the Bahamas are also homogeneous in species composition. STRI has the highest proportion of cells with relatively low turnover, and OBIS the lowest proportion among the major aggregators. There is a statistically significant negative relationship between turnover and completeness of cells in all databases (Pearson correlation;  $p < .01$  in all six databases), indicating that high turnover is an artefact related to insufficient sampling effort.

Occurrence records from different databases were used to identify bioregions in the Greater Caribbean (Figure 7). The optimal number of bioregions in the Greater Caribbean varied between seven and 14 according to the source database (Figure 7). In the maps, different regions are shown in different colours. Regions are larger for FishBase, with fewer records and more limited spatial coverage. STRI provided the clearest definition of three major bioregions: (a) the northern Gulf of Mexico and NE USA, (b) the north coast of South America and (c) the remaining centre of the study area. The other major databases showed the large central area as discrete



**FIGURE 4** Completeness of species inventories in one-degree cells throughout the Greater Caribbean in six databases. White cells indicate regions with no data or where completeness could not be calculated. Colour figure available online.



**FIGURE 5** MDS plots of species assemblages within one-degree cells, coloured according to 12 marine ecoregions of Spalding et al. (2007). Each dot in the scatter plots represents a grid-cell assemblage. Colours in the MDS plots correspond to colours in the map. Ecoregions are “Bahamian” (BA), “Bermuda” (BE), “Carolinian” (CA), “Eastern Caribbean” (EC), “Floridian” (F), “Greater Antilles” (GA), “Guianan” (G), “Northern Gulf of Mexico” (NGM), “Southern Caribbean” (SC), “Southern Gulf of Mexico” (SGM), “Southwestern Caribbean” (SWC) and “Western Caribbean” (WC). Colour figure available online.

and tended to separate out a northern Gulf of Mexico bioregion as well, but displayed considerable small-scale heterogeneity between them, and varied widely in the way they handle variability along coastal areas. Infomap analyses identified substantial numbers of small bioregions in all databases (Figure 7), and 88% per cent of eight such bioregions represented by a single cell were characterized by low (<50%) completeness in their faunas, demonstrating effects of insufficient sampling.

Bioregional maps produced by the six databases were compared quantitatively (Table 2). As expected from Figure 7, the overlap between the different maps was poor in all cases. The largest overlaps occur among OBIS, GBIF and iDigBio (goodness of fit of 0.56 to 0.40) and the smallest between FishBase and FishNet2 (goodness of fit of 0.23). Overlap between STRI and the other databases was at the low end of the scale (goodness of fit 0.27–0.29) in all cases except OBIS (goodness of fit 0.37).

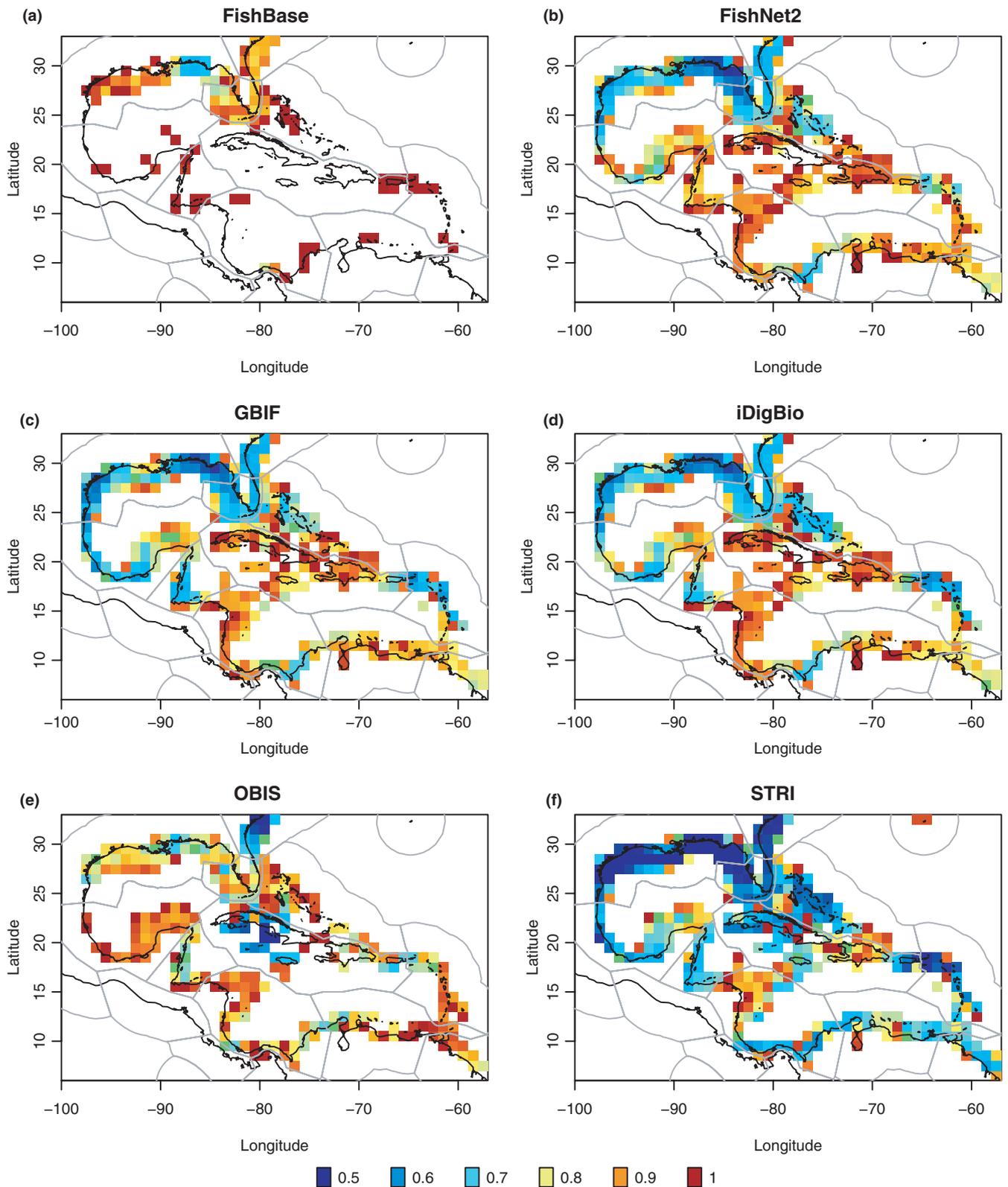
## 5 | DISCUSSION

Due to a long history of intensive exploration of the Greater Caribbean, the shore fishes of the region represent possibly the best known such fauna of any major marine bioregion in the world (Floeter et al., 2008; Linardich et al., 2019; Miloslavich et al., 2010). Consequently, the present study could be considered as a “best case scenario” for comparing biodiversity databases and determining the state of the knowledge. We found that the six databases assessed differed greatly in the abundance, and especially, distribution of georeferenced species records, and in their geographic patterns of

species richness and the completeness of local species inventories. Spatial sampling biases were pervasive in all databases, influencing the results of biodiversity analyses.

Gaps in data and sampling bias are a common issue in biodiversity databases, where sampling effort rarely is uniform in space (Meyer et al., 2015). To correct sampling bias, it has been suggested to decrease the spatial resolution of the data (Soberón, Jiménez, Golubov, & Koleff, 2007). This strategy, however, produced no better results in our data set for reef species (unpublished data): the marine environment is naturally patchy, with heterogeneous habitat distribution, and increasing the grain of the analysis tends to merely increase coverage over data-free, deep, off-shelf areas. Effects of sampling bias arose even when using methods devised to circumvent this issue such as Infomap Bioregions (Edler et al., 2017). These results highlight the need for caution when interpreting the results of analyses from occurrence data, even large sets of such data (García-Roselló et al., 2015; Meyer et al., 2015).

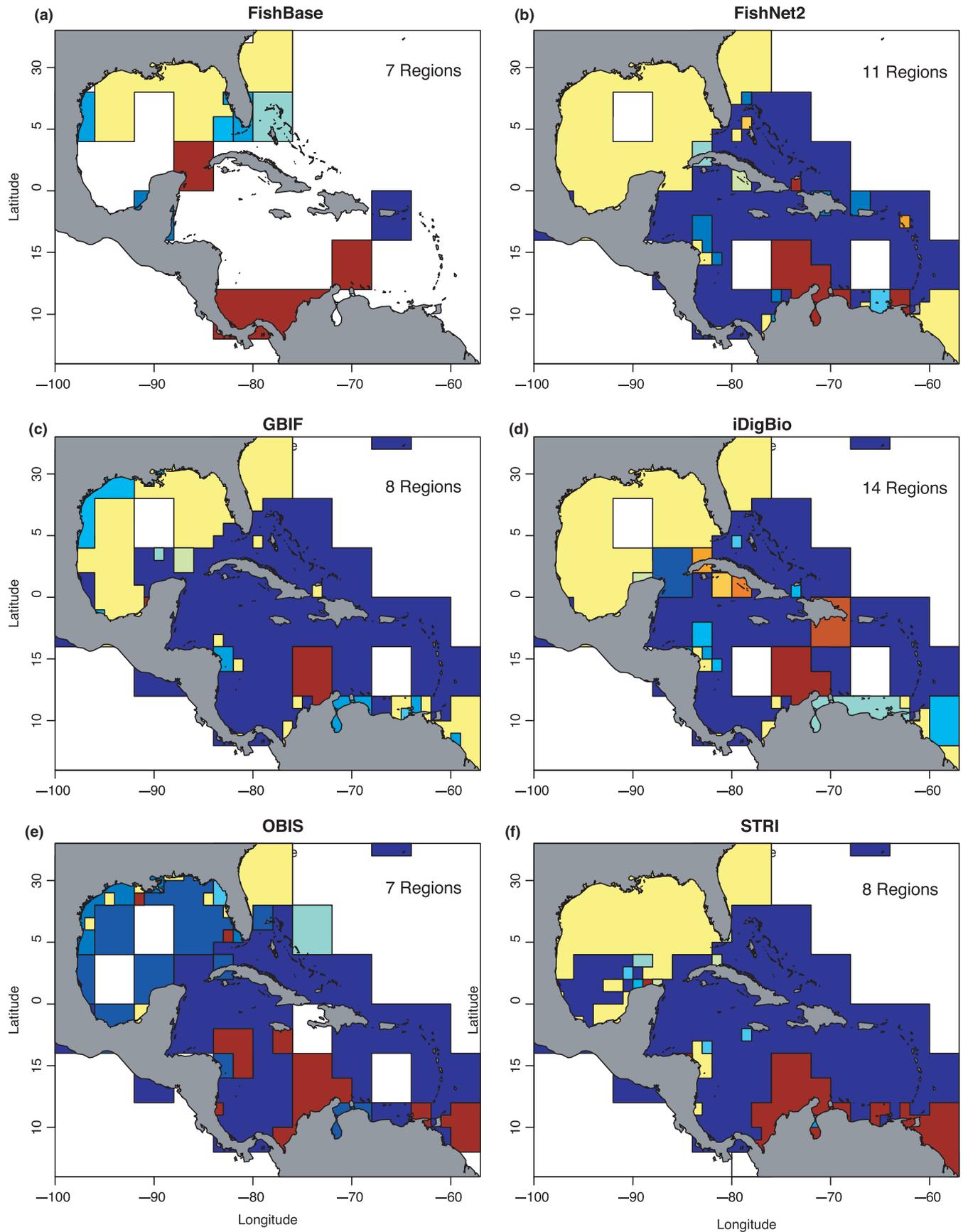
Substantial areas of the Greater Caribbean clearly lacked data in all databases. Data likely are available for some of those areas, but remain in analog format or have not been georeferenced (Feeley & Silman, 2011; Peterson, Soberón, & Krishtalka, 2015). FishNet2, for example, contains large numbers of museum collection records that lack georeferencing. Five years ago, it was estimated that only about 10% of information in museums worldwide was accessible via GBIF, the largest global aggregator of records (Peterson et al., 2015). Given the fading interest by funders in basic taxonomy, it is unlikely that the situation has radically improved since then. In other cases, data simply do not exist and some gaps in marine data likely reflect lack of sampling (Canónico et al., 2019). The mostly unexplored, but



**FIGURE 6** Maps of species turnover per one-degree cell based on data from different databases (a) FishBase; (b) FishNet2; (c) GBIF; (d) iDigBio; (e) OBIS; (f) STRI. Grey lines show outlines of the 12 Spalding et al. (2007) ecoregions. Colour figure available online.

extensive reefs and banks of the Honduras/Nicaragua shelf are a prominent case in point (Chollett et al., 2017; Chollett, Stoyke, & Box, 2014). There are now in place partnerships within scientific

communities, monitoring networks and observing systems to increase global coverage (Canonico et al., 2019). Results produced here could be used to bridge some of these data gaps and improve data



**FIGURE 7** Reef fish bioregions in the Greater Caribbean identified by Infomap using different databases (a) FishBase with 7 regions; (b) FishNet2 with 11 regions; (c) GBIF with 8 regions; (d) iDigBio with 14 regions; (e) OBIS with 7 regions; (f) STRI with 8 regions. Colour figure available online.

**TABLE 2** Goodness of fit between Infomap bioregions derived from each database.

	FishNet2	GBIF	iDigBio	OBIS	STRI
FishBase	0.26	0.31	0.25	0.35	0.29
FishNet2		0.31	0.35	0.29	0.28
GBIF			0.42	0.56	0.27
iDigBio				0.40	0.28
OBIS					0.37

coverage in the Greater Caribbean. Excluding highly under-sampled areas from some biogeographical analyses, such as species distribution models, can improve the explanatory value of these analyses and their reliability (Yang, Ma, & Kreft, 2013).

Given the extensive effort that has been put into developing the STRI database, which comprised curated data from the major aggregators and other online museum sources, and also incorporated much non-museum, “analog” data, how did that database, although not the largest, perform relative to the major aggregator databases? It did improve the Infomap bioregional resolution by showing the clearest subdivision of three major bioregions. These three as depicted by Infomap are broadly similar to the three produced by an earlier analysis by Robertson and Cramer (2014) that used a much less comprehensive database and different methodology: coarser scale sampling (presence/absence of species in 45 variably sized, irregularly shaped cells) and cluster analyses. This enhanced performance relative to the five aggregators can be attributed to more intense sampling throughout a greater proportion of the study area, leading to the highest levels of completeness among local faunas and lowest levels of turnover. However, STRI, along with the other databases, still suffered from low data availability in certain areas, notably parts of the southwest Gulf of Mexico, the Nicaragua/Honduras shelf and Hispaniola.

STRI also detected many more hotspots of species richness than any other database. Those hotspots were scattered throughout the continental shelf and various types of islands in the more tropical part of the region. Rather than being related to mesoscale environmental variation (Chollett, Mumby, Müller-Karger, & Hu, 2012), those sites simply correspond to sites of intensive sampling by ichthyologists at locations on both the continental shelf and offshore islands: SE Florida, northern Bahamas, the Puerto Rico plateau, some of the lesser Antilles, Colombia (mainland and southwest Atolls), Panama, Cayman Is, the Meso-American Barrier Reef System and sites on the Caribbean and Gulf shelves of Mexico. The richest hotspots in multiple databases—southeast Florida, the northern Bahamas and Puerto Rico—are the sites of the most intensive, long-duration research by USA ichthyologists. While there is high-intensity sampling of the northern Gulf of Mexico that region has only a moderate level of richness, a consequence of subtropical temperatures, large inflows of freshwater and a lack of shallow reefs.

STRI database provided the clearest picture of bioregionalism of the study area, with three large subdivisions evident. Various parts of those three areas were more weakly evident in the other

five databases. Relative contributions of historical vs contemporary processes to the production of these three faunal regions are far from clear. Historical biogeography of the Greater Caribbean is well developed at large (ocean basin) spatial scales through studies that have dealt with effects of the closure of the isthmus of Panama on the marine environments of that region (O’Dea & Collins, 2013); evolutionary relationships between the fish faunas of the Greater Caribbean and the tropical Eastern Pacific (Lessios, 2008); effects of the formation of the Amazon Barrier on relations between northern and southern West Atlantic reef fish faunas (Rocha, 2003); and historical (Floeter et al., 2008) and modern (Luiz et al., 2012) connections among shore fishes across the Atlantic and the Greater Caribbean and Brazil. Approaches relevant to the within-Greater Caribbean scale are much less well developed and have been based on qualitative assessments rather than quantitative analyses using location data. These, which were summarized in Robertson and Cramer (2014), include relations of the northern Gulf of Mexico and the eastern USA to the rest of the Greater Caribbean, between the Gulf and the eastern USA (e.g. see Briggs & Bowen, 2012 for both), between the mainland and offshore islands of the Caribbean, and between the northern coast of South America and Brazil. Previous work has focused mainly on how distribution patterns relate to modern variation in marine environments. The three major bioregions identified by Robertson and Cramer (2014) correspond to areas with major environmental differences: the Gulf of Mexico (particularly the northern part) and the southeast US coast have reduced temperatures, and relatively heightened productivity, the narrow northern coastal shelf of South America also is a high productivity zone with a set of substantial coastal upwelling systems and major river runoff areas, and the large central area of the Caribbean and offshore islands are found in an area of warm, oligotrophic waters that hosts large areas of coral reef (Chollett et al., 2012). The present analysis of an expanded set of STRI data identified the same three bioregions seen in Robertson and Cramer (2014), but with a substantial difference in the location of the border between the northern and central regions. Whereas the entire Gulf of Mexico and Florida were in the northern region in Robertson and Cramer (2014), the lower half of the Gulf and parts of southern Florida are included in the central region in the present analysis. This northward shift of that boundary can be attributed in part to recent improvements in data coverage for coral reef fishes in the Mexican part of the Gulf. The movement of part of southeast Florida from the northern to the central region may reflect some combination of better data coverage, the much finer scale of data cells in the present analysis, and the fact that some cells span both the southeast Florida coast and the western side of the Bahamas, which was not the case in Robertson and Cramer (2014).

Analyses of none of the databases used here supported the existence of Spalding et al.’s. (2007) 12-ecoregion subdivision of the Greater Caribbean. However, that system necessarily was constructed from expert opinion rather than through quantitative data analysis, likely incorporated information about species abundance as well as simple occurrence and was based on a broad variety of taxa

rather than just fishes. Changes to boundaries of the 12 ecoregions (c.f. Robertson & Cramer, 2014) may also reveal greater faunal differences than appeared here. How that 12-ecoregion arrangement applies to other components of the regional biota and its general validity remains to be determined.

Limited overlap between species lists and, particularly, georeferenced location records of major aggregator databases calls for their unification (Thomas, 2009). In particular, FishBase, a commonly used resource for fish biodiversity studies, provides a very different picture of biodiversity, due to limited data. Although we expected to see overlap between some databases, for example, between the major aggregators GBIF and OBIS/iDigBio/FishNet2, given that the last three contribute to GBIF, this was not the case, indicating that the workflow of the end-point aggregators needs to be improved and coordinated. Even GBIF, the largest existing online aggregator of species occurrences, does not include all species of reef-associated fishes known from the Greater Caribbean. Unification of databases would require reprocessing and reformatting large amounts of data (Thomas, 2009). The use of common standards by many databases, however, (e.g., Darwin core) would in principle ease this challenge.

Conserving biodiversity in the face of anthropogenic changes requires the best available knowledge on the distribution on species. To be able to produce accurate answers, we need more comprehensive, curated data sets, which involves working in tandem with many museum collections and incorporating data from numerous analog sources to assemble a truly global, comprehensive biodiversity database and fill gaps on species distributions through collection of new data.

## ACKNOWLEDGEMENTS

The six databases used in this paper were assembled and managed by Ernesto Peña of STRI. Data acquisition and construction of the STRI database were supported by funds from STRI, the Smithsonian Marine Science Network, the Smithsonian Publications Fund, the Smithsonian's Deep Reef Observation Project, the National Geographic Society, the IUCN Red List programme, the Harte Research Institute, and CONABIO. We thank REEF and AGRRA for supplying species-location records, fish biologists who provided taxonomic and location-record information used to construct that database (principal among them C Baldwin, S Brandl, K Conway B Frable, T Menut, T Munroe, R Robins, L Tornabene, J Van Tassell and B Victor), and hundreds of citizen-scientist photographers (see [https://biogeodb.stri.si.edu/caribbean/en/contributors/citizen\\_scientists](https://biogeodb.stri.si.edu/caribbean/en/contributors/citizen_scientists)), whose images of fish on reefs acted as vouchers for location records of many less common species.

## DATA AVAILABILITY STATEMENT

Many of the primary data used here are freely available from GBIF, OBIS, iDigBio, Fishnet2 and FishBase. Data incorporated in the STRI database and used in the present analyses represent a subset of data publicly available on the website: Robertson DR, Van Tassell JVT (2019) Shorefishes of the Greater Caribbean: online information system. Version 2.0. Smithsonian Tropical Research Institute, Balboa, Panama <https://biogeodb.stri.si.edu/caribbean/en/pages>.

All database records used in this work, consisting on species name, latitude and longitude of its occurrence in each of the six databases, are available in a database archived at the repository Zenodo (<https://doi.org/10.5281/zenodo.3606645>).

## ORCID

Iliana Chollett  <https://orcid.org/0000-0003-0536-499X>

## REFERENCES

- Aldecoa, R., & Marín, I. (2013). Exploring the limits of community detection strategies in complex networks. *Scientific Reports*, 3, 2216.
- Allen, G. R. (1991). Damsel-fishes of the world. *Mergus*.
- Baldwin, C. C., Tornabene, L., & Robertson, D. R. (2018). Below the Mesophotic. *Scientific Reports*, 8(1), 4920. <https://doi.org/10.1038/s41598-018-23067-1>
- Baselga, A. (2010). Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography*, 19(1), 134–143. <https://doi.org/10.1111/j.1466-8238.2009.00490.x>
- Baselga, A., Orme, D., Villeger, S., Bortoli, J. D., & Leprieux, F. (2018). betapart: Partitioning beta diversity into turnover and nestedness components. Retrieved from <https://CRAN.R-project.org/package=betapart>
- Bivand, R., Keitt, T., & Rowlingson, B. (2019). rgdal: bindings for the "geospatial" data abstraction library. R Package Version 0.8-16. Retrieved from <https://CRAN.R-project.org/package=rgdal>
- Bivand, R., & Rundel, C. (2019). rgeos: interface to geometry engine - open source ('GEOS'). R Package Version 0.3-2. Retrieved from <https://CRAN.R-project.org/package=rgeos>
- Briggs, J. C., & Bowen, B. W. (2012). A realignment of marine biogeographic provinces with particular reference to fish distributions. *Journal of Biogeography*, 39(1), 12–30. <https://doi.org/10.1111/j.1365-2699.2011.02613.x>
- Canonico, G., Buttigieg, P. L., Montes, E., Muller-Karger, F. E., Stepien, C., Wright, D., ... Murton, B. (2019). Global observational needs and resources for marine biodiversity. *Frontiers in Marine Science*, 6, 367. <https://doi.org/10.3389/fmars.2019.00367>
- Carpenter, K. E., & De Angelis, N. (2002). *The living marine resources of the Western Central Atlantic* (Vols. 1–3). Rome, Italy: Food and Agriculture Organization of the United Nations.
- Chollett, I., Garavelli, L., Holstein, D., Cherubin, L., Fulton, S., & Box, S. J. (2017). A case for redefining the boundaries of the Mesoamerican Reef Ecoregion. *Coral Reefs*, 36(4), 1039–1046. <https://doi.org/10.1007/s00338-017-1595-4>
- Chollett, I., Mumby, P. J., Müller-Karger, F. E., & Hu, C. (2012). Physical environments of the Caribbean Sea. *Limnology and Oceanography*, 57(4), 1233–1244. <https://doi.org/10.4319/lo.2012.57.4.1233>
- Chollett, I., Stoye, G., & Box, S. (2014). Honduran Miskito Cays: among the last unexplored reef systems in the Caribbean. *Coral Reefs*, 33(1), 155. <https://doi.org/10.1007/s00338-013-1084-3>
- Costello, M. J., Tsai, P., Wong, P. S., Cheung, A. K. L., Basher, Z., & Chaudhary, C. (2017). Marine biogeographic realms and species endemism. *Nature Communications*, 8(1), 1057. <https://doi.org/10.1038/s41467-017-01121-2>
- Ebert, D. A., Fowler, S. L., & Compagno, L. J. (2013). *Sharks of the world: A fully illustrated guide*. Plymouth, UK: Wild Nature Press.
- Edler, D., Guedes, T., Zizka, A., Rosvall, M., & Antonelli, A. (2017). Infomap Bioregions: Interactive mapping of biogeographical regions from species distributions. *Systematic Biology*, 66(2), 197–204. <https://doi.org/10.1093/sysbio/syw087>
- Edler, D., & Rosvall, M. (2015). The infomap software package. Retrieved from: <http://www.mapequation.org>
- Elith, J., & Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual Review*

- of *Ecology, Evolution, and Systematics*, 40(1), 677–697. <https://doi.org/10.1146/annurev.ecolsys.110308.120159>
- Eschmeyer, W., Fricke, R., & Van der Laan, R. (2015). *Catalog of fishes: Genera, species, references*. San Francisco, CA: California Academy of Sciences.
- Feeley, K. J., & Silman, M. R. (2011). The data void in modeling current and future distributions of tropical species. *Global Change Biology*, 17(1), 626–630. <https://doi.org/10.1111/j.1365-2486.2010.02239.x>
- Fegraus, E. H., Andelman, S., Jones, M. B., & Schildhauer, M. (2005). Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation. *The Bulletin of the Ecological Society of America*, 86(3), 158–168. [https://doi.org/10.1890/0012-9623\(2005\)86\[158:MTVOED\]2.0.CO;2](https://doi.org/10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2)
- Floeter, S. R., Rocha, L. A., Robertson, D. R., Joyeux, J. C., Smith-Vaniz, W. F., Wirtz, P., ... Bernardi, G. (2008). Atlantic reef fish biogeography and evolution. *Journal of Biogeography*, 35(1), 22–47. <https://doi.org/10.1111/j.1365-2699.2007.01790.x>
- Froese, R., & Pauly, D. (2009). FishBase. World wide web electronic publication. Retrieved from <http://www.fishbase.org>
- García-Roselló, E., Guisande, C., Manjarrés-Hernández, A., González-Dacosta, J., Heine, J., Pelayo-Villamil, P., ... Lobo, J. M. (2015). Can we derive macroecological patterns from primary Global Biodiversity Information Facility data? *Global Ecology and Biogeography*, 24(3), 335–347. <https://doi.org/10.1111/geb.12260>
- GEBCO Compilation Group. (2019). GEBCO 2019 Grid. <https://doi.org/10.5285/836f016a-33be-6ddc-e053-6c86abc0788e>
- Gotelli, N. J., & Colwell, R. K. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, 4(4), 379–391. <https://doi.org/10.1046/j.1461-0248.2001.00230.x>
- Hargrove, W. W., Hoffman, F. M., & Hessburg, P. F. (2006). Mapcurves: A quantitative method for comparing categorical maps. *Journal of Geographical Systems*, 8(2), 187. <https://doi.org/10.1007/s10109-006-0025-x>
- Hawkins, B. A., Rueda, M., & Rodríguez, M. Á. (2008). What do range maps and surveys tell us about diversity patterns? *Folia Geobotanica*, 43(3), 345. <https://doi.org/10.1007/s12224-008-9007-8>
- Hervé, M. (2019). RVAideMemoire: Testing and plotting procedures for biostatistics. R Package Version 0.9-69. <https://CRAN.R-project.org/package=RVAideMemoire>
- Hijmans, R. J. (2019). raster: Geographic data analysis and modeling. R Package Version 2.8. Retrieved from <https://CRAN.R-project.org/package=raster>
- Hurlbert, A. H., & Jetz, W. (2007). Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proceedings of the National Academy of Sciences*, 104(33), 13384. <https://doi.org/10.1073/pnas.0704469104>
- Kapoor, D. C. (1981). General bathymetric chart of the oceans (GEBCO). *Marine Geodesy*, 5(1), 73–80. <https://doi.org/10.1080/15210608109379408>
- Kramer, P. R., & Lang, J. C. (2003). Appendix one: The Atlantic and Gulf Rapid Reef Assessment (AGRR) Protocols: Former Version 2. 2. *Atoll Research Bulletin*, 496, 611–624.
- Kreft, H., & Jetz, W. (2010). A framework for delineating biogeographical regions based on species distributions. *Journal of Biogeography*, 37(11), 2029–2053. <https://doi.org/10.1111/j.1365-2699.2010.02375.x>
- Laffan, S. W., Rosauer, D. F., Di Virgilio, G., Miller, J. T., González-Orozco, C. E., Knerr, N., ... Mishler, B. D. (2016). Range-weighted metrics of species and phylogenetic turnover can better resolve biogeographic transition zones. *Methods in Ecology and Evolution*, 7(5), 580–588. <https://doi.org/10.1111/2041-210X.12513>
- Last, P., Naylor, G., Séret, B., White, W., de Carvalho, M., & Stehmann, M. (2016). *Rays of the World*. Clayton, VIC: CSIRO Publishing.
- Lessios, H. A. (2008). The great American schism: divergence of marine organisms after the rise of the Central American Isthmus. *Annual Review of Ecology, Evolution, and Systematics*, 39(1), 63–91. <https://doi.org/10.1146/annurev.ecolsys.38.091206.095815>
- Linardich, C., Ralph, G. M., Robertson, D. R., Harwell, H., Polidoro, B. A., Lindeman, K. C., & Carpenter, K. E. (2019). Extinction risk and conservation of marine bony shorefishes of the Greater Caribbean and Gulf of Mexico. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 29(1), 85–101. <https://doi.org/10.1002/aqc.2959>
- Lobo, J. M., Hortal, J., Yela, J. L., Millán, A., Sánchez-Fernández, D., García-Roselló, E., ... Guisande, C. (2018). KnowBR: an application to map the geographical variation of survey effort and identify well-surveyed areas from biodiversity databases. *Ecological Indicators*, 91, 241–248.
- Luiz, O. J., Madin, J. S., Robertson, D. R., Rocha, L. A., Wirtz, P., & Floeter, S. R. (2012). Ecological traits influencing range expansion across large oceanic dispersal barriers: insights from tropical Atlantic reef fishes. *Proceedings of the Royal Society B: Biological Sciences*, 279(1730), 1033–1040. <https://doi.org/10.1098/rspb.2011.1525>
- Maldonado, C., Molina, C. I., Zizka, A., Persson, C., Taylor, C. M., Albán, J., ... Antonelli, A. (2015). Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? *Global Ecology and Biogeography*, 24(8), 973–984. <https://doi.org/10.1111/geb.12326>
- Meyer, C., Kreft, H., Guralnick, R., & Jetz, W. (2015). Global priorities for an effective information basis of biodiversity distributions. *Nature Communications*, 6(1), 8221. <https://doi.org/10.1038/ncomms9221>
- Miloslavich, P., Díaz, J. M., Klein, E., Alvarado, J. J., Díaz, C., Gobin, J., ... Ortiz, M. (2010). Marine biodiversity in the Caribbean: regional estimates and distribution patterns. *PLoS One*, 5(8), e11916. <https://doi.org/10.1371/journal.pone.0011916>
- Nowosad, J., & Stepinski, T. (2018). Spatial association between regionalizations using the information-theoretical V-measure. *International Journal of Geographical Information Science*, 1–16. <https://doi.org/10.1080/13658816.2018.1511794>
- O'Dea, A., & Collins, L. S. (2013). Environmental, ecological, and evolutionary change in seas across the Isthmus of Panama. *Bulletin of Marine Science*, 89(4), 769–777. <https://doi.org/10.5343/bms.2013.1018>
- OBIS. (2019). Ocean Biogeographic Information System. Intergovernmental Oceanographic Commission of UNESCO. Retrieved from [www.iobis.org](http://www.iobis.org)
- Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., ... Wagner, H. (2018). vegan: Community Ecology Package. Retrieved from <https://CRAN.R-project.org/package=vegan>
- Pattengill-Semmens, C. V., & Semmens, B. X. (2003). Conservation and management applications of the reef volunteer fish monitoring program. In B. D. Melzian, V. Engle, M. McAlister, S. Sandhu, & L. K. Eads (Eds.), *Coastal Monitoring through Partnerships: Proceedings of the Fifth Symposium on the Environmental Monitoring and Assessment Program (EMAP) Pensacola Beach, FL, U.S.A., April 24–27, 2001* (pp. 43–50). The Netherlands: Springer. [https://doi.org/10.1007/978-94-017-0299-7\\_5](https://doi.org/10.1007/978-94-017-0299-7_5)
- Pauly, D., & Froese, R. (1991). FISHBASE: Assembling information on fish. *Naga, the ICLARM Quarterly*, 14(4), 10–11.
- Pebesma, E. J., & Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News*, 5(2), 9–13.
- Peterson, A. T., Soberón, J., & Krishtalka, L. (2015). A global perspective on decadal challenges and priorities in biodiversity informatics. *BMC Ecology*, 15(1), 15. <https://doi.org/10.1186/s12898-015-0046-8>
- Ratkowsky, D. A. (1990). *Handbook of nonlinear regression models*. New York, NY: Marcel Dekker Inc..
- Ready, J., Kaschner, K., South, A. B., Eastwood, P. D., Rees, T., Rius, J., ... Froese, R. (2010). Predicting the distributions of marine organisms at the global scale. *Ecological Modelling*, 221(3), 467–478. <https://doi.org/10.1016/j.ecolmodel.2009.10.025>
- Robertson, D. R. (2008). Global biogeographical data bases on marine fishes: caveat emptor. *Diversity and Distributions*, 14(6), 891–892. <https://doi.org/10.1111/j.1472-4642.2008.00519.x>

- Robertson, D. R., & Cramer, K. L. (2014). Defining and dividing the Greater Caribbean: insights from the biogeography of shorefishes. *PLoS One*, 9(7), e102918. <https://doi.org/10.1371/journal.pone.0102918>
- Robertson, D. R., Domínguez-Domínguez, O., Aroyo, Y. M. L., Mendoza, R. M., & Simões, N. (2019). Reef-associated fishes from the offshore reefs of western Campeche Bank, Mexico, with a discussion of mangroves and seagrass beds as nursery habitats. *ZooKeys*, 843, 71–115.
- Robertson, D. R., & Van Tassell, J. (2019). *Shorefishes of the Greater Caribbean: Online information system. Version 2.0*. Balboa, Panamá: Smithsonian Tropical Research Institute. Retrieved from <https://biogeodb.stri.si.edu/caribbean/en/pages>
- Robertson, M. P., Visser, V., & Hui, C. (2016). Biogeo: an R package for assessing and improving data quality of occurrence record datasets. *Ecography*, 39(4), 394–401.
- Rocha, L. A. (2003). Patterns of distribution and processes of speciation in Brazilian reef fishes. *Journal of Biogeography*, 30(8), 1161–1171. <https://doi.org/10.1046/j.1365-2699.2003.00900.x>
- Sandin, S. A., Vermeij, M. J. A., & Hurlbert, A. H. (2008). Island biogeography of Caribbean coral reef fish. *Global Ecology and Biogeography*, 17(6), 770–777. <https://doi.org/10.1111/j.1466-8238.2008.00418.x>
- Singer, R. A., Love, K. J., & Page, L. M. (2018). A survey of digitized data from U.S. fish collections in the iDigBio data aggregator. *PLoS One*, 13(12), e0207636. <https://doi.org/10.1371/journal.pone.0207636>
- Soberón, J., Jiménez, R., Golubov, J., & Koleff, P. (2007). Assessing completeness of biodiversity databases at different spatial scales. *Ecography*, 30(1), 152–160. <https://doi.org/10.1111/j.0906-7590.2007.04627.x>
- Soberón, J., & Llorente, J. (1993). The use of species accumulation functions for the prediction of species richness. *Conservation Biology*, 7(3), 480–488. <https://doi.org/10.1046/j.1523-1739.1993.07030480.x>
- Soberón, J., & Peterson, A. T. (2004). Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359(1444), 689–698. <https://doi.org/10.1098/rstb.2003.1439>
- Spalding, M. D., Fox, H. E., Allen, G. R., Davidson, N., Ferdaña, Z. A., Finlayson, M., ... Robertson, J. (2007). Marine ecoregions of the World: A bioregionalization of coastal and shelf areas. *BioScience*, 57(7), 573–583. <https://doi.org/10.1641/B570707>
- Thomas, C. (2009). Biodiversity databases spread, prompting unification call. *Science*, 324(5935), 1632–1633. <https://doi.org/10.1126/science.324.1632>
- Thomas, C., Cameron, A., Green, R. E., Bakkenes, M., Beaumont, L. J., Collingham, Y. C., ... Williams, S. E. (2004). Extinction risk from climate change. *Nature*, 427(6970), 145–148. <https://doi.org/10.1038/nature02121>
- Tilman, D., Wedin, D., & Knops, J. (1996). Productivity and sustainability influenced by biodiversity in grassland ecosystems. *Nature*, 379(6567), 718–720. <https://doi.org/10.1038/379718a0>
- Title, P. (2017). speciesRaster: Generation of SpeciesRaster object and calculation of morphological and phylogenetic metrics. R Package Version 1.0.
- VanCompernelle, M., Knouft, J. H., & Ficklin, D. L. (2019). Multispecies conservation of freshwater fish assemblages in response to climate change in the southeastern United States. *Diversity and Distributions*, 25(9), 1388–1398. <https://doi.org/10.1111/ddi.12948>
- Vilhena, D. A., & Antonelli, A. (2015). A network approach for identifying and delimiting biogeographical regions. *Nature Communications*, 6(1), 6848. <https://doi.org/10.1038/ncomms7848>
- Wessel, P., & Smith, W. H. F. (1996). A global, self-consistent, hierarchical, high-resolution shoreline database. *Journal of Geophysical Research: Solid Earth*, 101(B4), 8741–8743. <https://doi.org/10.1029/96JB0104>
- Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., ... Vieglais, D. (2012). Darwin Core: An evolving community-developed biodiversity data standard. *PLoS One*, 7(1), e29715. <https://doi.org/10.1371/journal.pone.0029715>
- Yang, W., Ma, K., & Kreft, H. (2013). Geographical sampling bias in a large distributional database and its effects on species richness–environment models. *Journal of Biogeography*, 40(8), 1415–1426. <https://doi.org/10.1111/jbi.12108>
- Yesson, C., Brewer, P. W., Sutton, T., Caithness, N., Pahwa, J. S., Burgess, M., ... Culham, A. (2007). How global is the Global Biodiversity Information Facility? *PLoS One*, 2(11), e1124. <https://doi.org/10.1371/journal.pone.0001124>
- Zizka, A., Carvalho, F. A., Calvente, A., Baez-Lizarazo, M. R., Cabral, A., Ramos Coelho, J. F., ... Antonelli, A. (2020). No one-size-fits-all solution to clean GBIF. *BioRxiv*. <https://doi.org/10.1101/2020.03.12.974543>
- Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., ... Antonelli, A. (2019). CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, 10(5), 744–751. <https://doi.org/10.1111/2041-210X.13152>

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Chollett I, Robertson DR. Comparing biodiversity databases: Greater Caribbean reef fishes as a case study. *Fish Fish*. 2020;00:1–18. <https://doi.org/10.1111/faf.12497>

**-Supplementary information**

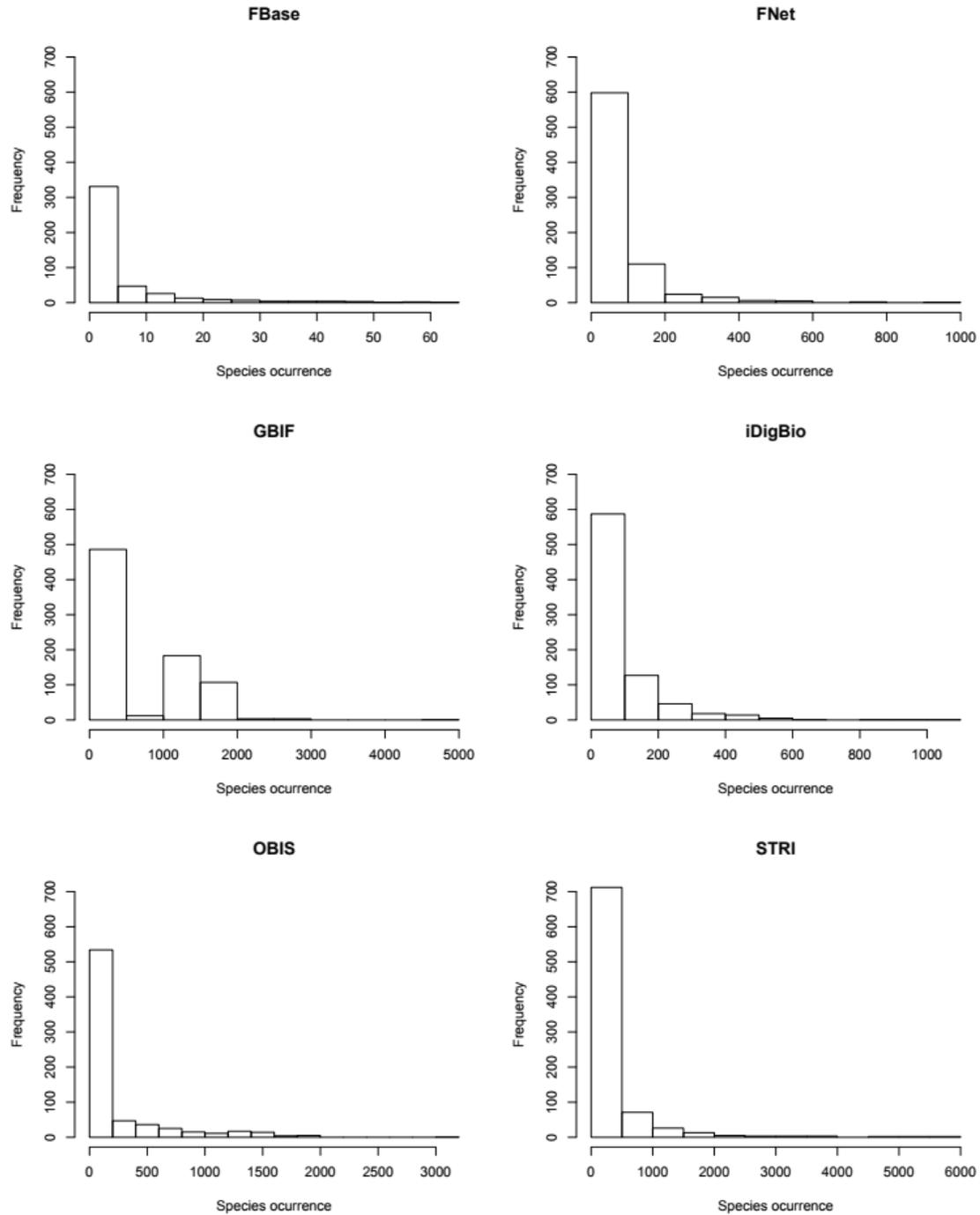


Figure S1.1. Histogram with the number of occurrences of each species in each dataset

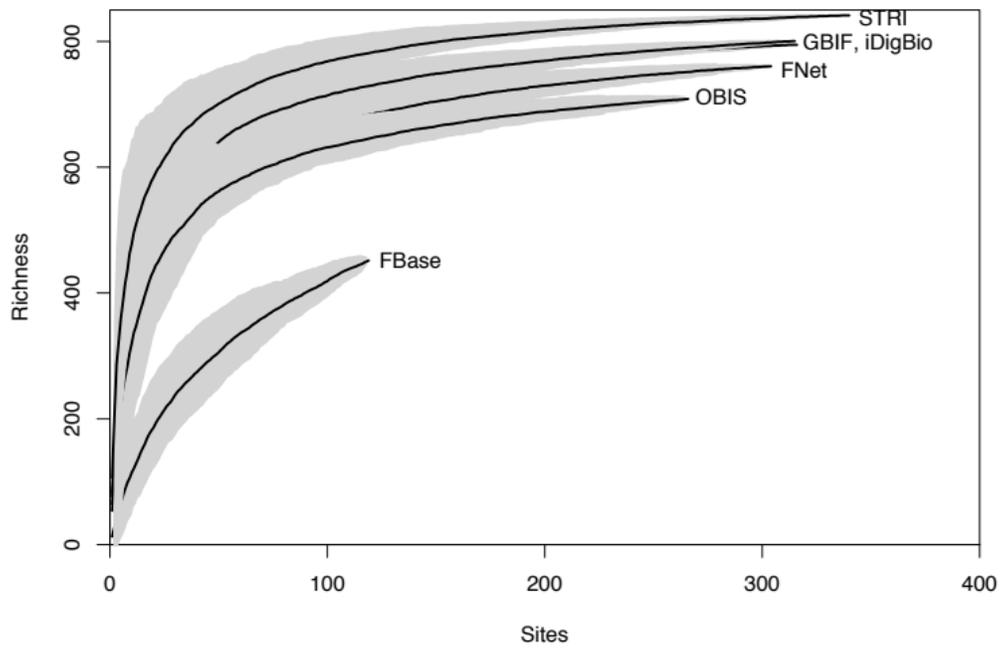


Figure S1.2. Species accumulation curves (average and 95% confidence interval) for each database. Each site is a one-degree cell.

Table S1.1. Values of completeness for each database in each one-degree cell. Latitude and Longitude indicate the center of a one-degree cell. NA indicates no data. At the end of the table there are summary statistics for each database.

Latitude	Longitude	FBase	FNet	GBIF	iDigBio	OBIS	STRI
32.5	-80.5	15.42	23.78	41.95	22.14	77.75	84.25
32.5	-79.5	24.30	28.76	47.31	33.79	80.12	71.59
32.5	-78.5	28.71	NA	54.23	NA	50.50	49.92
32.5	-65.5	NA	NA	NA	NA	NA	23.47
32.5	-64.5	NA	NA	35.53	51.72	28.13	65.54
31.5	-81.5	NA	64.65	66.32	69.07	81.80	88.08
31.5	-80.5	13.29	38.76	42.17	36.07	79.63	84.45
31.5	-79.5	NA	62.72	63.59	63.80	73.18	81.29
30.5	-90.5	NA	57.35	NA	36.82	57.35	15.67
30.5	-89.5	NA	63.56	41.43	59.57	73.03	61.79
30.5	-88.5	39.25	76.29	72.87	71.81	71.56	84.17
30.5	-87.5	70.13	75.63	78.40	74.71	64.57	79.95
30.5	-86.5	55.98	61.19	68.91	60.28	52.39	78.39
30.5	-85.5	23.99	58.16	58.30	57.37	17.36	74.36
30.5	-84.5	NA	56.71	60.26	60.09	45.02	61.26
30.5	-81.5	8.01	62.21	65.05	67.11	85.92	84.03
30.5	-80.5	25.08	67.13	69.76	66.98	81.02	81.55
30.5	-79.5	18.24	NA	9.54	NA	23.00	28.75
29.5	-95.5	NA	28.08	37.51	40.82	NA	37.19
29.5	-94.5	NA	63.10	64.22	63.80	12.14	78.15
29.5	-93.5	NA	58.16	63.96	49.43	10.57	84.28
29.5	-92.5	NA	42.14	NA	NA	2.56	78.10
29.5	-91.5	NA	50.23	25.76	51.12	54.73	67.74
29.5	-90.5	NA	69.68	24.27	63.47	68.08	73.23
29.5	-89.5	28.28	69.74	44.38	65.22	66.30	75.68
29.5	-88.5	58.54	72.93	68.94	72.87	62.92	88.01
29.5	-87.5	74.99	78.96	77.97	76.86	49.57	87.48
29.5	-86.5	59.18	73.25	74.05	73.34	20.43	78.31
29.5	-85.5	36.62	58.26	56.50	53.74	30.22	77.54
29.5	-84.5	NA	78.38	77.78	77.05	26.59	84.74
29.5	-83.5	32.75	75.20	70.44	75.07	28.27	85.80
29.5	-82.5	NA	38.89	17.06	17.06	NA	30.11
29.5	-81.5	NA	58.02	57.45	58.14	84.30	81.30
29.5	-80.5	18.47	67.43	68.31	66.77	78.87	83.00
28.5	-97.5	NA	39.28	49.20	55.96	72.79	50.72
28.5	-96.5	NA	54.43	58.52	59.06	39.56	78.18
28.5	-95.5	NA	67.57	64.99	62.71	25.71	85.34
28.5	-94.5	45.05	65.20	51.34	53.90	65.09	76.98
28.5	-93.5	13.36	49.62	45.79	42.20	19.54	64.48
28.5	-92.5	61.88	37.08	47.08	40.91	11.17	80.24
28.5	-91.5	NA	53.88	48.90	50.07	28.47	73.74

28.5	-90.5	NA	55.95	50.46	54.32	47.15	78.31
28.5	-89.5	NA	52.57	51.56	49.90	36.22	65.84
28.5	-88.5	NA	NA	NA	NA	NA	45.84
28.5	-85.5	46.46	62.92	57.46	58.09	21.96	72.61
28.5	-84.5	51.50	67.58	70.61	66.61	10.52	82.89
28.5	-83.5	NA	56.42	52.20	51.23	43.41	86.31
28.5	-82.5	18.24	67.46	39.47	NA	8.69	63.31
28.5	-80.5	10.04	63.00	67.03	65.84	79.91	72.05
28.5	-79.5	NA	21.95	21.05	21.05	29.04	26.11
27.5	-97.5	NA	64.04	70.46	71.19	51.75	80.03
27.5	-96.5	NA	66.08	62.02	50.86	29.83	85.65
27.5	-95.5	45.05	48.38	30.82	29.77	18.51	79.70
27.5	-94.5	NA	26.21	29.90	31.51	NA	62.60
27.5	-93.5	NA	44.33	23.53	22.81	52.55	82.18
27.5	-92.5	NA	12.73	NA	NA	13.36	58.88
27.5	-91.5	NA	NA	NA	NA	NA	16.78
27.5	-84.5	56.91	62.73	64.96	57.32	NA	77.05
27.5	-83.5	54.30	72.54	66.03	62.40	6.50	82.32
27.5	-82.5	40.31	75.85	67.33	65.61	36.41	80.63
27.5	-80.5	28.64	67.45	64.33	64.17	43.80	75.01
27.5	-79.5	NA	26.42	29.14	29.72	40.14	28.92
27.5	-78.5	NA	NA	NA	NA	NA	2.75
27.5	-77.5	NA	NA	NA	NA	NA	16.11
26.5	-97.5	NA	46.91	55.70	58.76	38.68	73.88
26.5	-96.5	NA	52.01	46.56	46.45	49.71	82.61
26.5	-84.5	28.05	40.89	45.12	34.90	NA	61.77
26.5	-83.5	27.91	68.82	57.66	52.87	31.84	81.59
26.5	-82.5	14.52	74.79	72.26	72.02	36.77	83.19
26.5	-81.5	NA	46.37	48.12	50.32	NA	65.61
26.5	-80.5	45.03	72.15	76.67	70.74	54.99	71.20
26.5	-79.5	NA	22.57	20.06	16.01	6.90	42.51
26.5	-78.5	NA	13.47	18.83	55.36	10.54	58.91
26.5	-77.5	NA	26.46	30.90	32.14	NA	51.02
26.5	-76.5	NA	24.69	25.67	26.26	58.71	54.96
25.5	-97.5	NA	34.58	51.78	51.83	NA	59.41
25.5	-96.5	NA	41.36	41.36	36.68	NA	39.25
25.5	-84.5	54.23	NA	25.67	24.31	NA	38.66
25.5	-83.5	NA	63.50	54.40	53.11	9.45	74.14
25.5	-82.5	11.79	66.92	69.49	70.26	18.15	77.47
25.5	-81.5	NA	80.91	76.94	76.64	48.44	81.73
25.5	-80.5	33.99	83.38	94.65	85.44	96.38	83.78
25.5	-79.5	23.02	47.91	62.13	59.39	22.35	70.33
25.5	-78.5	NA	10.54	4.26	12.48	34.37	30.00
25.5	-77.5	NA	68.30	71.28	86.14	40.64	91.55
25.5	-76.5	NA	20.20	20.57	35.67	NA	60.98
24.5	-97.5	NA	57.38	64.42	62.81	NA	64.49
24.5	-96.5	NA	NA	NA	NA	NA	28.28

24.5	-83.5	40.21	65.32	102.28	63.60	96.71	68.74
24.5	-82.5	4.65	78.12	99.77	80.14	77.53	81.20
24.5	-81.5	32.74	78.22	91.02	77.33	96.73	84.63
24.5	-80.5	13.36	85.49	95.15	87.34	96.68	84.22
24.5	-79.5	NA	6.46	4.88	4.88	NA	7.22
24.5	-78.5	NA	NA	NA	NA	NA	22.92
24.5	-77.5	NA	47.17	49.01	62.99	83.64	76.82
24.5	-76.5	NA	53.07	54.18	60.92	NA	70.55
24.5	-75.5	NA	49.36	49.36	52.33	NA	55.04
24.5	-74.5	NA	9.16	14.93	26.31	28.46	55.07
23.5	-97.5	NA	50.18	49.81	51.06	NA	56.80
23.5	-89.5	NA	29.23	28.13	29.23	NA	39.74
23.5	-88.5	NA	15.42	NA	NA	NA	15.35
23.5	-87.5	NA	40.18	52.52	48.55	NA	39.94
23.5	-82.5	NA	11.88	12.02	NA	NA	34.45
23.5	-81.5	NA	3.51	19.02	4.49	NA	NA
23.5	-80.5	NA	NA	NA	42.94	82.03	70.50
23.5	-79.5	NA	NA	7.31	11.20	NA	6.06
23.5	-78.5	NA	NA	NA	NA	NA	10.73
23.5	-77.5	NA	10.20	9.74	21.72	67.56	42.34
23.5	-76.5	NA	59.16	59.89	60.46	59.75	69.79
23.5	-75.5	NA	49.40	48.83	48.90	NA	59.50
23.5	-74.5	NA	18.18	20.38	19.01	34.36	NA
23.5	-73.5	NA	NA	NA	4.96	NA	24.32
22.5	-97.5	NA	46.10	38.37	46.88	33.04	45.42
22.5	-91.5	NA	11.96	8.68	5.13	NA	36.63
22.5	-90.5	NA	20.43	23.16	24.40	NA	21.33
22.5	-89.5	NA	25.99	21.01	22.05	10.54	69.88
22.5	-88.5	NA	NA	41.97	41.98	2.66	43.79
22.5	-87.5	NA	15.32	38.48	38.48	NA	15.92
22.5	-86.5	NA	17.96	18.17	18.11	1.24	28.65
22.5	-84.5	NA	NA	NA	NA	NA	38.88
22.5	-83.5	NA	NA	NA	NA	NA	23.10
22.5	-82.5	NA	NA	NA	NA	28.71	23.09
22.5	-81.5	NA	NA	NA	NA	85.85	68.24
22.5	-80.5	NA	NA	NA	NA	NA	6.90
22.5	-79.5	NA	NA	14.92	NA	77.70	64.15
22.5	-78.5	NA	29.20	34.59	25.71	72.47	67.20
22.5	-77.5	NA	NA	NA	NA	49.86	45.33
22.5	-75.5	NA	16.77	15.50	18.82	NA	69.44
22.5	-74.5	NA	19.24	21.95	37.76	NA	74.20
22.5	-73.5	NA	NA	NA	22.89	NA	74.81
22.5	-72.5	NA	NA	NA	9.52	NA	25.16
21.5	-97.5	NA	48.67	86.42	46.63	NA	86.80
21.5	-92.5	NA	42.87	26.48	29.68	NA	42.52
21.5	-91.5	NA	NA	55.65	55.73	32.85	54.84
21.5	-90.5	NA	NA	23.88	16.70	NA	56.46

21.5	-89.5	NA	NA	27.30	22.03	NA	34.59
21.5	-88.5	NA	NA	54.65	NA	NA	7.41
21.5	-87.5	NA	NA	9.09	10.52	NA	10.89
21.5	-86.5	20.69	32.58	54.94	41.48	27.63	55.49
21.5	-84.5	NA	NA	NA	NA	NA	49.33
21.5	-83.5	NA	8.00	10.89	14.75	49.04	55.83
21.5	-82.5	NA	NA	NA	NA	83.63	73.79
21.5	-81.5	NA	NA	NA	NA	92.42	80.03
21.5	-79.5	NA	35.78	6.87	35.36	89.83	39.97
21.5	-77.5	NA	NA	NA	NA	NA	45.48
21.5	-75.5	NA	NA	NA	NA	NA	40.78
21.5	-73.5	NA	NA	NA	7.29	NA	66.97
21.5	-72.5	NA	5.48	36.49	9.38	77.30	60.36
21.5	-71.5	NA	50.23	46.90	49.37	63.81	54.03
20.5	-97.5	NA	NA	22.68	22.77	NA	17.51
20.5	-96.5	NA	NA	65.94	10.54	NA	67.39
20.5	-92.5	NA	24.52	NA	19.00	NA	63.97
20.5	-91.5	NA	60.54	51.06	53.10	20.91	64.77
20.5	-90.5	NA	NA	19.48	17.85	NA	34.68
20.5	-87.5	NA	16.70	72.39	21.93	68.69	67.89
20.5	-86.5	NA	6.46	76.12	20.63	6.17	56.47
20.5	-79.5	NA	NA	NA	NA	82.86	84.87
20.5	-78.5	NA	NA	NA	NA	91.08	59.69
20.5	-74.5	NA	NA	NA	NA	NA	45.72
20.5	-73.5	NA	23.75	18.08	17.06	NA	7.74
20.5	-70.5	NA	NA	NA	NA	NA	6.06
20.5	-69.5	NA	28.13	28.13	28.13	NA	21.69
19.5	-96.5	NA	21.04	73.07	26.10	8.71	NA
19.5	-95.5	NA	18.24	61.06	3.39	NA	65.24
19.5	-92.5	NA	55.36	69.27	68.70	NA	71.33
19.5	-91.5	NA	39.30	67.08	66.10	NA	62.46
19.5	-90.5	NA	NA	20.61	5.26	NA	32.98
19.5	-87.5	NA	14.99	65.51	51.35	78.12	71.91
19.5	-81.5	NA	61.68	68.42	67.36	74.95	83.79
19.5	-80.5	NA	NA	10.34	NA	58.14	67.96
19.5	-79.5	NA	25.34	21.93	25.34	68.56	67.91
19.5	-75.5	NA	NA	10.51	NA	NA	17.59
19.5	-72.5	NA	NA	NA	NA	NA	52.30
19.5	-71.5	NA	NA	11.57	NA	73.37	64.39
19.5	-70.5	NA	NA	4.26	NA	NA	NA
19.5	-69.5	NA	20.53	21.96	20.53	34.99	33.58
18.5	-95.5	NA	31.47	42.55	26.00	NA	54.10
18.5	-94.5	NA	45.68	66.15	51.92	NA	74.48
18.5	-93.5	NA	45.79	64.66	61.19	7.47	65.49
18.5	-92.5	NA	51.28	65.93	65.62	NA	67.12
18.5	-91.5	NA	9.98	58.27	48.55	NA	55.30
18.5	-88.5	NA	NA	47.31	40.45	NA	41.47

18.5	-87.5	NA	12.22	75.82	62.12	82.98	71.47
18.5	-83.5	NA	NA	NA	NA	NA	28.71
18.5	-78.5	NA	NA	44.91	NA	79.36	69.94
18.5	-77.5	NA	NA	12.51	15.38	61.08	61.34
18.5	-76.5	NA	19.52	14.01	21.01	76.48	55.42
18.5	-74.5	NA	NA	8.32	NA	NA	7.14
18.5	-73.5	NA	NA	NA	NA	NA	10.15
18.5	-72.5	NA	23.50	24.33	48.80	NA	63.61
18.5	-71.5	NA	40.21	22.31	40.21	NA	NA
18.5	-70.5	NA	NA	NA	NA	NA	13.36
18.5	-69.5	NA	19.99	22.30	15.96	28.09	37.59
18.5	-68.5	NA	NA	43.30	NA	51.21	33.47
18.5	-67.5	NA	38.03	37.01	32.10	14.35	58.97
18.5	-66.5	NA	15.91	29.29	30.52	8.01	69.81
18.5	-65.5	NA	24.86	NA	59.09	79.58	76.09
18.5	-64.5	NA	54.58	63.93	58.70	93.52	89.46
18.5	-63.5	NA	18.78	31.12	34.05	50.62	58.16
18.5	-62.5	NA	32.85	18.12	32.85	NA	10.00
17.5	-88.5	NA	33.61	44.00	41.58	63.49	55.61
17.5	-87.5	NA	50.37	74.27	69.44	78.41	63.23
17.5	-83.5	NA	3.39	3.39	3.39	NA	NA
17.5	-78.5	NA	37.59	38.35	38.35	NA	43.39
17.5	-77.5	NA	53.35	53.13	53.74	14.58	66.02
17.5	-76.5	NA	43.34	45.60	48.60	NA	57.04
17.5	-75.5	NA	11.53	15.34	16.53	NA	8.39
17.5	-71.5	NA	11.10	10.52	11.10	56.85	44.25
17.5	-67.5	NA	35.50	54.53	49.88	80.73	81.49
17.5	-66.5	NA	17.49	32.91	33.60	82.41	86.30
17.5	-65.5	NA	NA	NA	NA	NA	4.95
17.5	-64.5	NA	83.84	83.57	82.88	90.63	85.02
17.5	-63.5	NA	62.36	67.81	61.66	70.04	69.90
17.5	-62.5	NA	5.75	30.39	35.35	37.01	66.26
17.5	-61.5	NA	63.51	64.00	65.30	NA	74.74
16.5	-88.5	NA	59.84	79.67	67.42	62.75	82.78
16.5	-87.5	NA	NA	67.49	67.36	78.72	40.75
16.5	-86.5	NA	17.68	73.73	61.57	29.26	60.27
16.5	-85.5	NA	21.00	17.82	23.62	NA	66.28
16.5	-83.5	NA	4.65	4.65	4.65	NA	19.24
16.5	-82.5	NA	46.17	43.62	46.17	NA	57.77
16.5	-81.5	NA	6.15	11.11	11.11	12.48	21.02
16.5	-80.5	NA	NA	NA	NA	NA	10.51
16.5	-77.5	NA	NA	NA	NA	NA	76.12
16.5	-61.5	NA	NA	61.45	NA	NA	30.86
15.5	-88.5	NA	NA	20.53	NA	34.42	64.39
15.5	-87.5	NA	NA	NA	NA	NA	49.71
15.5	-86.5	NA	8.01	37.18	39.29	NA	72.75
15.5	-85.5	NA	NA	NA	NA	NA	24.82

15.5	-84.5	NA	NA	28.28	38.68	NA	NA
15.5	-83.5	NA	64.53	64.18	62.77	20.37	64.06
15.5	-82.5	NA	23.25	18.08	21.99	NA	26.49
15.5	-81.5	NA	15.87	14.35	18.17	NA	20.26
15.5	-79.5	NA	NA	74.95	NA	75.14	40.02
15.5	-63.5	NA	NA	2.41	2.41	6.46	NA
15.5	-61.5	NA	19.99	23.80	25.06	NA	51.33
14.5	-81.5	NA	12.86	8.77	10.07	3.08	33.54
14.5	-80.5	NA	NA	84.57	8.99	NA	64.48
14.5	-61.5	NA	19.96	28.71	25.08	NA	31.07
14.5	-60.5	NA	8.88	41.78	36.44	7.05	59.39
13.5	-82.5	NA	NA	NA	NA	NA	9.52
13.5	-81.5	NA	58.19	74.57	59.85	18.85	85.72
13.5	-80.5	NA	NA	NA	NA	NA	35.24
13.5	-61.5	NA	6.06	6.85	3.51	NA	32.20
13.5	-59.5	NA	34.56	44.47	39.48	33.09	69.34
12.5	-83.5	NA	7.41	6.46	15.15	65.88	67.77
12.5	-82.5	NA	12.55	11.27	11.04	55.21	42.63
12.5	-81.5	NA	NA	19.25	4.00	79.48	74.78
12.5	-72.5	NA	44.01	45.77	42.71	31.62	64.26
12.5	-71.5	NA	49.14	26.82	48.47	38.85	64.83
12.5	-70.5	NA	35.31	33.63	34.68	30.77	43.65
12.5	-69.5	NA	11.20	33.68	15.51	40.99	43.85
12.5	-68.5	NA	38.66	72.29	47.43	66.80	71.14
12.5	-61.5	NA	19.32	30.48	61.17	29.20	72.40
11.5	-83.5	NA	6.90	6.46	6.90	NA	45.53
11.5	-75.5	NA	12.48	12.48	11.75	23.03	51.95
11.5	-74.5	NA	24.37	62.22	26.52	61.28	76.76
11.5	-73.5	NA	NA	7.79	7.40	50.10	60.40
11.5	-72.5	NA	28.32	28.32	28.32	60.54	65.11
11.5	-71.5	NA	NA	NA	NA	NA	58.18
11.5	-70.5	NA	36.95	38.03	38.03	57.01	55.67
11.5	-69.5	NA	8.69	8.69	8.69	NA	38.25
11.5	-68.5	NA	NA	NA	NA	NA	50.92
11.5	-66.5	NA	38.76	38.13	39.29	84.56	74.65
11.5	-65.5	NA	NA	NA	NA	NA	54.51
11.5	-64.5	NA	25.31	25.31	25.31	NA	51.48
11.5	-63.5	NA	31.22	31.11	32.19	NA	49.51
11.5	-62.5	NA	23.69	25.33	NA	NA	34.72
11.5	-61.5	NA	NA	NA	NA	NA	10.83
11.5	-60.5	NA	61.48	66.10	63.89	59.27	68.52
10.5	-83.5	NA	5.88	NA	NA	29.71	66.32
10.5	-76.5	NA	NA	NA	NA	NA	42.68
10.5	-75.5	NA	41.85	45.47	43.11	26.40	68.65
10.5	-74.5	NA	NA	25.76	NA	26.47	52.78
10.5	-71.5	NA	NA	40.21	45.05	NA	77.42
10.5	-68.5	NA	NA	12.34	NA	9.16	12.73

10.5	-67.5	NA	NA	NA	NA	NA	15.48
10.5	-66.5	NA	NA	39.56	NA	45.91	29.80
10.5	-65.5	NA	NA	18.32	17.47	NA	38.63
10.5	-64.5	NA	19.25	21.48	44.01	NA	61.96
10.5	-63.5	NA	13.51	12.82	31.54	NA	44.34
10.5	-62.5	NA	NA	NA	NA	NA	32.27
10.5	-61.5	NA	20.83	37.51	37.16	34.96	45.89
10.5	-60.5	NA	29.19	22.13	24.78	NA	43.86
9.5	-83.5	NA	NA	NA	NA	NA	6.33
9.5	-82.5	13.89	29.80	68.88	34.43	58.53	78.69
9.5	-81.5	NA	NA	NA	NA	NA	29.74
9.5	-80.5	NA	19.97	20.68	21.09	NA	23.38
9.5	-79.5	NA	57.28	64.53	63.61	43.43	73.19
9.5	-78.5	26.96	48.62	66.26	54.74	67.82	67.79
9.5	-77.5	NA	NA	NA	NA	NA	42.31
9.5	-76.5	NA	47.45	49.95	48.31	53.81	60.52
9.5	-75.5	NA	45.72	51.22	45.23	36.57	75.17
9.5	-61.5	NA	NA	NA	NA	NA	46.97
9.5	-60.5	NA	23.75	NA	NA	55.96	73.24
9.5	-59.5	NA	48.55	43.89	51.93	19.92	62.35
8.5	-81.5	NA	NA	NA	NA	NA	30.18
8.5	-80.5	NA	NA	NA	NA	NA	28.71
8.5	-77.5	NA	30.34	31.62	30.89	52.84	46.22
8.5	-76.5	NA	55.40	54.98	54.52	26.20	50.39
8.5	-59.5	NA	22.05	18.71	19.03	8.71	59.20
8.5	-58.5	NA	47.85	45.78	48.84	NA	69.62
8.5	-57.5	NA	28.13	28.13	28.13	NA	4.65
7.5	-76.5	NA	NA	NA	NA	34.56	29.34
7.5	-58.5	NA	27.61	25.05	25.05	NA	51.90
7.5	-57.5	NA	13.36	NA	11.79	22.31	34.46
<b>Mean</b>		<b>33.20</b>	<b>39.46</b>	<b>42.56</b>	<b>41.02</b>	<b>47.57</b>	<b>54.77</b>
<b>st.dev</b>		<b>18.02</b>	<b>21.86</b>	<b>23.26</b>	<b>21.22</b>	<b>26.53</b>	<b>22.94</b>
<b>N</b> cells with data		<b>45</b>	<b>214</b>	<b>244</b>	<b>233</b>	<b>176</b>	<b>298</b>
<b>%</b> cells with data		<b>14.71</b>	<b>69.93</b>	<b>79.74</b>	<b>76.14</b>	<b>57.52</b>	<b>97.39</b>



Table S1.2. Pairwise PERMANOVA comparisons of faunal composition across the 12 ecoregions of Spalding et al. (2007). Probability values are shown, with values below 0.05 highlighted in bold. Ecoregion code in brackets also according to Figure 5(g) in main text. Ecoregions are "Bahamian" (BA), "Bermuda" (BE), "Carolinian" (CA), "Eastern Caribbean" (EC), "Floridian" (F), "Greater Antilles" (GA), "Guianan" (G), "Northern Gulf of Mexico" (NGM), "Southern Caribbean" (SC), "Southern Gulf of Mexico" (SGM), "Southwestern Caribbean" (SWC), "Western Caribbean" (WC). Figure 5 in the body of the paper includes a map showing the distribution of the 12 ecoregions. Note: FishBase lacks sufficient data in BE and G for a full set of comparisons.

FishBase

	BA	CA	EC	F	GA	NGM	SC	SGM	SWC
CA	<b>0.045</b>	-	-	-	-	-	-	-	-
EC	1.000	0.090	-	-	-	-	-	-	-
F	0.720	0.270	0.135	-	-	-	-	-	-
GA	1.000	<b>0.045</b>	1.000	0.270	-	-	-	-	-
NGM	0.135	1.000	0.495	0.495	0.540	-	-	-	-
SC	1.000	0.180	1.000	0.360	1.000	0.855	-	-	-
SGM	1.000	<b>0.045</b>	1.000	0.270	1.000	0.180	1.000	-	-
SWC	1.000	<b>0.045</b>	1.000	<b>0.045</b>	1.000	0.090	1.000	1.00	-
WC	1.000	0.225	1.000	<b>0.045</b>	1.000	0.225	1.000	1.000	1.000

FishNet2

	BA	BE	CA	EC	F	GA	G	NGM	SC	SGM	SWC
BE	1.000	-	-	-	-	-	-	-	-	-	-
CA	0.078	1.000	-	-	-	-	-	-	-	-	-
EC	1.000	1.000	0.078	-	-	-	-	-	-	-	-
F	0.078	1.000	0.078	0.078	-	-	-	-	-	-	-
GA	0.234	1.000	0.078	0.702	0.078	-	-	-	-	-	-
G	0.078	1.000	0.078	0.078	0.078	0.078	-	-	-	-	-
NGM	0.078	1.000	0.780	0.078	0.078	0.078	0.078	-	-	-	-
SC	0.078	1.000	0.078	0.078	0.078	0.078	0.702	0.078	-	-	-
SGM	0.078	1.000	0.078	0.078	0.078	0.078	0.078	0.078	0.078	-	-
SWC	0.078	1.000	0.078	0.156	0.078	0.078	0.078	0.078	0.390	0.08	-
WC	1.000	1.000	0.078	1.000	0.078	1.000	0.078	0.078	0.234	0.08	0.08

GBIF

	BA	BE	CA	EC	F	GA	G	NGM	SC	SGM	SWC
--	----	----	----	----	---	----	---	-----	----	-----	-----

BE	1.000	-	-	-	-	-	-	-	-	-	-	-
CA	0.078	1.000	-	-	-	-	-	-	-	-	-	-
EC	0.390	1.000	0.078	-	-	-	-	-	-	-	-	-
F	0.078	1.000	0.078	0.078	-	-	-	-	-	-	-	-
GA	0.936	1.000	0.078	0.078	0.078	-	-	-	-	-	-	-
G	0.078	1.000	0.078	0.078	0.078	0.078	-	-	-	-	-	-
NGM	0.078	1.000	0.390	0.078	0.078	0.078	0.078	-	-	-	-	-
SC	0.078	1.000	0.078	0.078	0.078	0.078	0.468	0.078	-	-	-	-
SGM	0.078	1.000	0.078	0.078	0.078	0.078	0.078	0.078	0.078	-	-	-
SWC	0.078	1.000	0.078	0.078	0.078	0.078	0.156	0.078	1.000	0.078	-	-
WC	0.390	1.000	0.078	1.000	0.078	0.078	0.078	0.078	0.078	0.078	0.078	0.08

### iDigBIO

	BA	BE	CA	EC	F	GA	G	NGM	SC	SGM	SWC
BE	1.000	-	-	-	-	-	-	-	-	-	-
CA	0.078	1.000	-	-	-	-	-	-	-	-	-
EC	1.000	1.000	0.078	-	-	-	-	-	-	-	-
F	0.078	1.000	0.234	0.078	-	-	-	-	-	-	-
GA	0.078	1.000	0.078	0.078	0.078	-	-	-	-	-	-
G	0.078	1.000	0.078	0.078	0.078	0.078	-	-	-	-	-
NGM	0.078	1.000	0.858	0.078	0.078	0.078	0.078	-	-	-	-
SC	0.078	1.000	0.078	0.078	0.078	0.078	0.468	0.078	-	-	-
SGM	0.078	1.000	0.078	0.078	0.078	0.078	0.078	0.078	0.078	-	-
SWC	0.078	1.000	0.078	0.078	0.078	0.156	0.078	0.078	0.624	0.078	-
WC	1.000	1.000	0.078	1.000	0.078	0.156	0.078	0.078	0.078	0.078	0.08

### OBIS

	BA	BE	CA	EC	F	GA	G	NGM	SC	SGM	SWC
BE	1.000	-	-	-	-	-	-	-	-	-	-
CA	0.088	1.000	-	-	-	-	-	-	-	-	-
EC	1.000	1.000	0.088	-	-	-	-	-	-	-	-
F	0.088	1.000	0.088	0.088	-	-	-	-	-	-	-
GA	0.880	1.000	0.088	0.264	0.088	-	-	-	-	-	-
G	0.088	1.000	0.088	0.088	0.088	0.088	-	-	-	-	-
NGM	0.088	1.000	0.088	0.088	0.264	0.088	0.088	-	-	-	-
SC	0.880	1.00	0.088	1.000	0.088	0.176	0.088	0.088	-	-	-

SGM	0.088	1.000	0.088	0.088	0.088	0.088	0.088	0.088	0.264	-	-
SWC	0.088	1.000	0.088	1.000	0.088	0.088	0.088	0.088	1.000	0.09	-
WC	1.000	1.000	0.088	1.000	0.088	1.000	0.088	0.088	1.000	0.088	0.176

---

STRI

---

	BA	BE	CA	EC	F	GA	G	NGM	SC	SGM	SWC
BE	1.000	-	-	-	-	-	-	-	-	-	-
CA	0.078	1.000	-	-	-	-	-	-	-	-	-
EC	0.702	1.000	0.078	-	-	-	-	-	-	-	-
F	0.078	1.000	0.078	0.078	-	-	-	-	-	-	-
GA	0.078	1.000	0.078	0.078	0.078	-	-	-	-	-	-
G	0.078	1.000	0.078	0.078	0.078	0.078	-	-	-	-	-
HGM	0.078	0.312	0.156	0.078	0.078	0.078	0.078	-	-	-	-
SC	0.078	1.000	0.078	0.078	0.078	0.078	0.078	0.078	-	-	-
SGM	0.078	1.000	0.078	0.078	0.078	0.078	0.078	0.078	0.078	-	-
SWC	0.078	1.000	0.078	0.312	0.078	0.078	0.078	0.078	0.078	0.078	-
WC	0.078	1.000	0.078	0.078	0.078	1.000	0.078	0.078	0.078	0.078	0.08

---