Title

Enhancing georeferenced biodiversity inventories: automated information extraction from literature records reveal the gaps

Authors

Bjørn Tore Kopperud[1], Scott Lidgard[2], Lee Hsiang Liow[1, 3,*]

Affiliations

[1] Natural History Museum, University of Oslo, PO Box 1172 Blindern, 0318 Oslo, Norway
[2] Negaunee Integrative Research Center, Field Museum, 1400 South Lake Shore Drive, Chicago IL, 60605, U.S.A.
[3] Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, PO Box 1066 Blindern, 0316 Oslo, Norway

Corresponding author

Lee Hsiang Liow. E-mail: l.h.liow@nhm.uio.no

Abstract

We use natural language processing (NLP) to retrieve location data for cheilostome bryozoan species (text-mined occurrences [TMO]) in an automated procedure. We compare these results with data from the Ocean Biogeographic Information System (OBIS). Using OBIS and TMO data separately and in combination, we present latitudinal species richness curves using standard estimators (Chao2 and the Jackknife) and range-through approaches. Our combined OBIS and TMO species richness curves quantitatively document a bimodal global latitudinal diversity gradient for cheilostomes for the first time, with peaks in the temperate zones. 79% of the georeferenced species we retrieved from TMO (N = 1780) and OBIS (N = 2453) are non-overlapping and underestimate known species richness, even in combination. Despite clear indications that global location data compiled for cheilostomes should be improved with concerted effort, our study supports the view that latitudinal species richness patterns deviate from the canonical LDG. Moreover, combining online biodiversity databases with automated information retrieval from the published literature is a promising avenue for expanding taxon-location datasets.

Keywords: cheilostomes, marine, latitudinal species richness, natural language processing (NLP), OBIS

Acknowledgements

We thank Mali H. Ramfjell for compiling part of our training dataset, the GeoDeepDive group, especially Ian Ross and Shanan Peters, for providing access to articles, and OBIS and their contributors for their georeferenced taxonomic data. We thank Phil Bock for maintaining bryozoa.net, Dennis Gordon for an updated version of the Working List of Genera and Subgenera for the Treatise on Invertebrate Paleontology, and both for their contributions to WoRMS. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 724324 to L.H. Liow).

Introduction

Global biogeographical and macroecological studies require data on aggregate entities, such as location-specific occurrences of taxa and regional species assemblages, in order to understand emergent patterns at global and/or temporal scales (McGill, 2019). Assembly of such detailed yet broad-scale data is highly labor-intensive; the sampling effort required for a specific research question can be daunting for any one researcher or single research team. This is one reason why collaborative and often public databases have gained traction (Klein et al., 2019). For instance, empirical global biogeographic analyses (Costello et al., 2017; Rabosky et al., 2018) are increasingly based on public databases of georeferenced taxonomic occurrences, such as the Ocean Biogeographic Information System (OBIS, www.iobis.org) and the Global Biodiversity Information Facility (GBIF, www.gbif.org). Analyzing such georeferenced databases with tools that alleviate incomplete or biased sampling (Saeedi et al., 2019) allows us to address questions on large-scale distributions of clades, especially those that are well-represented in such databases. Yet for less well-studied clades, prospects for obtaining large amounts of such data are lower. Answering pattern-based questions such as 'how many species of clade z are found in location y' and more process-oriented questions such as 'how did the current latitudinal diversity gradient form' both require location-specific taxonomic data in substantial volume. In addition, *generalized* biogeographic patterns and processes will be supported more robustly if they include a greater diversity of clades.

Cheilostome bryozoans, though less well-studied than several metazoan clades of similar size, are ubiquitous in benthic marine habitats. They are the most diverse order of Bryozoa with 83% of a conservatively estimated 5869 extant described species (Bock & Gordon, 2013). Bryozoans are ecologically important habitat builders (Wood, Rowden, Compton, Gordon, & Probert, 2013) and are vital components of the marine food chain (Lidgard, 2008). Despite important analyses of regional species distributions (Barnes & Griffiths, 2008; Clarke & Lidgard, 2000; López Gappa, 2000; Moyano, 1991), their global species richness distribution has never been quantified. We argue that even with concerns about the incompleteness of OBIS records for the purpose of inferring regional to global diversity patterns (e.g. Klein et al., 2019; Lindsay et al., 2017; Reimer et al., 2019), it is worth exploring cheilostome data in OBIS. We do so in order to identify spatial gaps in sampling but also to ask if automated information retrieval can enhance the species occurrence data available in OBIS.

Automated information retrieval (Hirschberg & Manning, 2015) is one recent approach to the time-consuming manual activity of data compilation from the scientific literature. Automated text-mining is well-established in the biomedical realm (Christopoulou, Tran, Sahu, Miwa, & Ananiadou, 2020; Percha, Garten, & Altman, 2012), but has only recently been adopted for biodiversity studies (Kopperud, Lidgard, & Liow, 2019; Peters, Husson, & Wilcots, 2017). As far as we are aware, automated text-mining has never been applied to the literature for extraction of taxon occurrences in given locations for the purpose to understanding biogeography (but see Page, 2019). We use natural language processing tools (Bojanowski, Grave, Joulin, & Mikolov, 2017; De Marneffe et al., 2014), to compile cheilostome text-mined occurrence data (TMO) to complement and potentially enhance data from OBIS.

Taxon occurrence data from OBIS and TMO are not expected to be the same. We ask if they could, separately or in combination, shed light on a long-standing biogeographic hypothesis in the bryozoological literature. Many different groups of organisms show the canonical latitudinal diversity gradient (LDG), a species richness peak in tropical regions and decreasing species richness towards

the temperate and polar zones (Hillebrand, 2004; Menegotto, Kurtz, & Lana, 2019). Despite being common across marine and terrestrial realms, and among diverse eukaryote clades, the LDG is not universal (Chaudhary, Saeedi, & Costello, 2016). Extratropical bimodal species richness peaks have been observed, for example in deep-sea brittle stars (Woolley et al., 2016), razor shells (Saeedi, Dennis, & Costello, 2017) and foraminiferans (Rutherford, D'Hondt, & Prell, 1999). Bimodality has also been suggested for cheilostome bryozoans (Barnes & Griffiths, 2008; Clarke & Lidgard, 2000; Schopf, 1970).

The TMO and OBIS data in combination support the view that the latitudinal diversity pattern of living cheilostomes is bimodal. These data reveal highest levels of estimated species richness in temperate latitudes, but TMO species richness has a peak in the northern hemisphere while OBIS has a peak in the temperate south. Moreover, the data sets differ significantly in the geographic richness patterns in Atlantic versus Pacific ocean basins (Barnes & Griffiths, 2008; Schopf, 1970). We discuss the pros and cons of TMO and public databases such as OBIS and how their differences can help us understand the uncertainties of the retrieved spatial diversity patterns, beyond what is estimated within the confines of each dataset.

Methods

OBIS Data Retrieval

We use the R-package robis (Provoost & Bosch, 2020) to access OBIS (28.11.2019) and retrieve latitude/longitude occurrence records of cheilostomes. We remove records without species epithets. For taxonomic ambiguities such as cf., aff., we disregard the uncertainty; for instance, *Microporella* cf. *ciliata* becomes *Microporella ciliata*. Records with genus names that are not accepted according to either the Working List of Genera and Subgenera for the Treatise on Invertebrate Paleontology (pers comm. Dennis P. Gordon, 2019), World Register of Marine Species (WoRMS Editorial Board, 2020) or www.bryozoa.net (Bock, 2020) are also removed. The result is 561 unique genus names and 2453 unique genus-species combinations (henceforth simply species) in 144917 retained OBIS records.

TMO (Text-Mined Occurrence) Data Retrieval

We follow a previously detailed text-mining procedure (Kopperud et al., 2019) with modifications. We extract text from two collections of published works, our own corpus (3233 pdf documents) and the GeoDeepDive archive (GDD, https://geodeepdive.org/), which contains full-text contents of journal articles. Only English language publications and those likely to feature extant bryozoans were used for information extraction (see Appendix S1 in Supporting Information).

We use CoreNLP (Manning et al., 2014) for an initial natural language analysis prior to information extraction, including tokenization, named-entity recognition, and dependency grammar annotation (Hirschberg & Manning, 2015). We use a pre-trained machine-learning model to recognize location names in the text (Finkel, Grenager, & Manning, 2005). To facilitate extraction of species, we compile names from the Working List of Genera and Subgenera for the Treatise on Invertebrate Paleontology (pers comm. Dennis P. Gordon, 2019), World Register of Marine Species (WoRMS Editorial Board, 2020) and www.bryozoa.net (Bock, 2020) that we then use in rule-based recognition (Chang & Manning, 2014). For example, consider a sentence from Tilbrook et al. (2001, p. 50):

121 "The avicularia resemble those seen in *B. intermedia* (Hincks, 1881b), from Tasmania and
122 New Zealand, but this species is only just over half the size of *B. cookae*."

123 This sentence contains two species names ("*B. intermedia*" and "*B. cookae*") and two location names
124 ("Tasmania" and "New Zealand"). Each species-location pair is a candidate relation. The sentence
125 implies that *B. intermedia* is found in New Zealand (a positive relation), but does not say anything
126 about where *B. cookae* is found (a negative relation). We automate this distinction using a machine-
127 learning classifier that we trained using a dataset of 4938 unique candidates labelled as positive or
128 negative by two persons. Part of our procedure resolves the genus name referred to as '*B.*' above
129 (see Appendix S1).

130 We use a test data set comprising 10% of the labelled candidates to evaluate several aspects of our
131 machine-classifier: (i) Accuracy, the ratio of correct predictions to all predictions; (ii) precision, the
132 ratio of true positive predictions to all positive predictions; (iii) recall, the ratio of true positive
133 predictions to all positive labels; (iv) false positive rate (FPR), the ratio of false positive predictions to
134 all negative labels; and (v) F1, the harmonic mean of precision and recall. Each of these metrics yields
135 different information on the reliability of the extracted data. We treat taxonomic ambiguities within
136 TMO data in the same manner as OBIS records (see previous section).

137 From TMO location names to spatial data

138 Location names (e.g., New Zealand, Tasmania) are submitted to the Google geocoding service
139 (https://developers.google.com/maps/documentation/geocoding/) to acquire a bounding box with
140 four latitude-longitude coordinates and a centroid (Fig. S1). We remove species occurrences in
141 locations represented by bounding boxes that are larger than about 2% of the Earth's surface using
142 area calculations assuming a spherical globe. See Fig. S2 for how the bounding box sizes are
143 distributed, and Fig. S3 for how alternative thresholds impact the result.

144 Estimating latitudinal species richness

145 We initially evaluate species richness in thirty-six 5° latitudinal bands using two standard richness
146 estimators that perform relatively well under a suite of conditions (Walther & Moore, 2005): Chao2
147 and Jackknife using the function specpool in the R package vegan (Oksanen et al., 2019). We treat
148 these latitudinal bands as independent. We then repeat the procedure using thirty-six equal area
149 bands, since areas represented within equal angle bands decrease poleward. To apply these
150 estimators, we divide each (equal angle or area) latitudinal band into 5° longitudinal sampling units.
151 We use the bias-corrected form of Chao2 = $S_{obs} + Q_1^2(N-1)/(2NQ_2)$, and incidence-based
152 Jackknife = $S_{obs} + Q_1(N-1)/N$. Here, S_{obs} is the number of observed species in each band, N is
153 the number of (longitudinal) sampling units, Q_1 is the number of species observed in only one
154 sampling unit, and Q_2 is the number observed in two sampling units. Because terrestrial regions are
155 not suitable habitats for marine cheilostomes, we mapped all landlocked longitudinal sampling bins
156 (Fig. S4) based on a 1:10 m map of global coastlines (Patterson, 2019). We removed the landlocked
157 bins prior to richness estimation. For OBIS data where spatial coordinates are points, it is trivial to
158 assign data to sampling units. For TMO, we assume that a species occurs in all of the sampling units
159 that intersect the bounding box associated with the location. TMO bounding boxes vary in size, but
160 most are smaller in area than our sampling units (Fig. S2).

In addition to Chao2 and Jackknife estimators, we also determined range-through species richness. Here, we assume that a species spans its southernmost and northernmost occurrence record, regardless of whether it is observed in any intermediate latitudinal band.

The code and data required to reproduce the analyses and the figures are stored datadryad.org and will be available upon submission.

Results

Capturing species diversity: comparing OBIS and TMO

Applying the text-mining procedure to our corpora, we retrieved 1780 species in 382 genera, and 1915 unique location names among 9653 TMO records. Only 27% of the species in the OBIS data that we retained were also in TMO. Similarly, only 41% of these combinations in the TMO occurred in OBIS. 20% of the species richness is common to both (Figs. 1, S5). In combination with OBIS data, we have species-location information from 3323 species or 68% of 4921 described cheilostome species (Bock & Gordon, 2013).

Our machine-classifier achieved an accuracy of 73.1%, F1 of 76.8%, recall of 78.9%, FPR of 34.3% and precision of 74.8% as estimated with our test set (Fig. S6b). These results are substantially better than a random classifier baseline, but not as good as the human annotator repeatability. Specifically, the FPR among annotators is about 15% (n = 200). A random classifier that is as unbalanced as our training data (60% positive labels) would yield 60% false positives, but a random classifier equaling our classifier's recall of 78.9% would have the same false positive rate of 78.9% (see Appendix S1).

Latitudinal species richness patterns

Combined TMO and OBIS data in plots of range-through species richness show a bimodal pattern with species richness peaks in both hemispheres surrounding 40° and -40° (Fig. 1). Inferred species richness in both of these peaks is about double that in the tropics. The two data sources contribute different latitudinal constituents, as suggested by the limited overlap in their species composition (Fig. 1 inset).

Chao2 and Jackknife estimated species richness from OBIS shows two peaks between -20° and -45° that are more than double the next highest peak between 25° and 50° (Fig. 2a). In contrast, TMO estimated richness shows a highest peak between 30° and 45° (Fig. 2b). With minor exceptions in the Antarctic where spatial distortion is largest, equiangular and equi-areal bands yield nearly identical inferences (compare Fig. 2a,c). The latitudinal pattern appears smoother when using larger latitudinal band sizes (Fig. S7), while retaining a qualitatively similar picture. Longitudinal sampling bins of varied sizes appear not to be important for the Jackknife and Chao2 estimators (Fig. S8).

The northern hemisphere peak in richness (Fig. 1) reflects TMO records from the Mediterranean and Japan, but also from the Atlantic Ocean (Fig. 3a,e), including the British Isles. Note that we did not include the Mediterranean as part of the Atlantic basin for Fig. 3. A portion of the TMO data are spatially imprecise, for example the location names "France", "Spain" or "Morocco" may be associated with Mediterranean endemics, yet these records could contribute to the Atlantic richness counts in Fig. 3. The spatially precise OBIS data show a much lower peak in the Eastern Atlantic (Fig. 3e, orange line shifted slightly northward), reflecting data from the British Isles and northern Europe. Conversely, OBIS data mainly from Australia and New Zealand contribute disproportionately to the

huge southern hemisphere peak. The richness captured by OBIS in Australia and New Zealand is not reflected by TMO species richness (Fig. 3b,d). The western Atlantic and eastern Pacific do not display such pronounced temperate zone peaks (Fig. 3c,f). Looking at individual ocean basins, TMO and OBIS are sometimes congruent and other times incongruent. For example, there is an absence of OBIS records in Japanese waters, and there are similarly few TMO and OBIS records in the Indian Ocean (Fig. 4).

Such varied regional species richness patterns are in part influenced by the geographic occurrence of samples. Figure 4 summarizes the relative distribution of species-location records for TMO and OBIS data as global heatmaps. For OBIS data, there are about one order of magnitude fewer records in tropical regions than for subtropical and temperate ones (Fig. S9a). While there are also fewer TMO records in tropical regions, the effect is not as pronounced (Fig. S9b). Northern and southern hemisphere species richness peaks in the two data sets (Fig. 1) correspond with high regional densities of TMO and OBIS records, respectively (Fig. 3e,d).

Discussion

Causal hypotheses for a LDG and contrarian patterns are plentiful and can sometimes be tested in groups with rich and relatively unbiased spatial data from both extant and extinct taxa (Jablonski et al., 2013; Jablonski, Roy, & Valentine, 2006; Krug, Jablonski, & Valentine, 2007) or those with independent molecular phylogenetic evidence (e.g. Rabosky *et al.* 2018). We believe ours is the first study to quantify global cheilostome species biogeographic patterns. Using a combined TMO and OBIS perspective, and a bimodal latitudinal diversity gradient in cheilostome species richness is quite apparent. Yet, at present, we can merely speculate about what processes that may have led to their latitudinal pattern. Given the biases and heterogeneity of the data we explored which are striking when comparing our two data sources, we also need to consider (i) how this pattern coincides with previous observations, and (ii) methodological, sampling, and taxonomic concerns.

Two patterns in our analyses are similar to Schopf's (1970) findings from then-scarce available data: higher species richness on the eastern margin of the Atlantic and the western margin of the Pacific compared to their opposite margins, and increasing richness with latitude away from the equator. Our combined data conforms with the first finding, but still doesn't capture the richness of the severely-understudied Philippine-Indonesian region and its many archipelagoes (Gordon, 1999; Okada & Mawatari, 1958; Tilbrook & De Grave, 2005). Changes to the second finding are more nuanced, and may partly reflect relatively lower equatorial sampling density (Chaudhary et al., 2016; Chaudhary, Saeedi, & Costello, 2017; Fernandez & Marques, 2017; Menegotto & Rangel, 2018) apparent in both of our datasets (Fig. S9). However, our observed peaks of species richness are at significantly higher latitudes than those reported for bryozoans in Chaudhary et al. (2016).

Fossil and modern patterns of bryozoan abundance in cool-water carbonate sediments suggest that the lower tropical species richness is not merely a sampling artifact. Modern bryozoan-dominated carbonate platforms are far more common on cool-water temperate shelves than on tropical ones (James & Clarke, 1997; Schlanger & Konishi, 1975). Cenozoic tropical bryozoan faunas are both less abundantly preserved and less diverse than those from temperate latitudes, possibly reflecting biotic interactions, preservational biases, and cryptic existence in shallower-water habitats dominated by corals, calcareous algae, and other photobiont organisms (Taylor & Di Martino, 2014; Winston, 1986). A far-reaching study by Taylor & Allison (1998) showed that 94% of bryozoan-rich post-

Paleozoic sedimentary deposits formed outside of the paleotropics, which may be especially significant if regional species richness and skeletal abundance are linked. About a third of all described bryozoan species occur south of -30°, and 87% of these are cheilostomes (Barnes and Griffiths, 2008).

We chose to discretize the data in latitudinal bands and longitudinal bins that are larger than those previously used e.g., in Rabosky et al. (2018). The choice of band- and bin sizes for species richness estimation is somewhat arbitrary. Differing choices suggest quantitatively dissimilar inferences, although the bimodality is still apparent in the cases we have explored (Figs. S7 and S8). A range-through latitudinal diversity approach (Fig. 3) assumes that any species that is not observed in a gap between two adjacent latitudinal bands should contribute to species richness in that gap, but this assumption is quite easily broken (Menegotto & Rangel, 2018). The bounding boxes used for TMO locations may also tend to bleed range margins to as opposed to OBIS point location data. Richness estimates may be inflated via range-through estimates, particularly in the tropics, compared to estimating richness independently in each latitudinal band which yields lower estimates (Fig. 2). Regardless, both methods for estimating species richness give a picture of bimodality.

Global biogeographic studies such as ours are more prone to the issues of sampling and taxonomic concerns than local or regional ones, simply due to their scope. Large sampling gaps are apparent in both TMO and OBIS datasets. The development and application of richness estimation models that distinguish true absences from non-observations (Iknayan, Tingley, Furnas, & Beissinger, 2014) may help improve inferences, but are likely insufficient to overcome blatant sampling gaps. Overall, there are relatively few records in the Indian Ocean, most of the South Atlantic, and eastern margin of the Pacific. TMO records for the Arctic are sparse, as are OBIS records for the northwest Pacific. Aside from a few extreme outliers from OBIS British Isles locations, species richness and number of records per 5° latitudinal band have a strong positive relationship (Fig. S10). Independent taxonomic surveys of underrepresented regions in one or both datasets corroborate the existence of significant gaps (Florence, Hayward, & Gibbons, 2007; Grischenko, Mawatari, & Taylor, 2000; Hirose, 2017; X. Liu & Liu, 2008; López Gappa, 2000; Moyano, 1991; Vieira, Migotto, & Winston, 2008). The OBIS records may partly reflect recent histories of active bryozoan research programs in the Antarctic (Barnes & Griffiths, 2008; Figuerola, Barnes, Brickle, & Brewin, 2017) and Australia and New Zealand (Wood et al., 2013) as well as contributions to OBIS that differ substantially among research institutions. On the other hand, TMO extracted extensive species-location information from the Mediterranean (27° to 50°) that are severely wanting in OBIS, demonstrating that combining disparate data sources can help bridge gaps in global biogeographic studies.

Taxonomic errors inevitably exist in large databases. Taxonomy is continuously subject to revisions (Bock & Gordon, 2013), not all of which are accounted for in our datasets. Many species await description; (Gordon, Bock, Souto-Derungs, & Reverter Gil, 2019) suggest that there are over 6,400 'known' cheilostome species without commenting on nomenclatural status, suggesting that there are up to 600 ‘known’ species that need naming. Our machine-classifier is currently unable to extract location information for 18% of the species that were detected in our corpus of published works (Fig. S5). Our conversion of taxonomic ambiguities into certainties likely deflated species richness estimates, while mistaken inclusion of fossil species names may have inflated richness estimates. We have assumed these do not necessarily introduce spatial bias. Additionally, many bryozoan species determined by traditional morphological methods may actually consist of unrecognized species

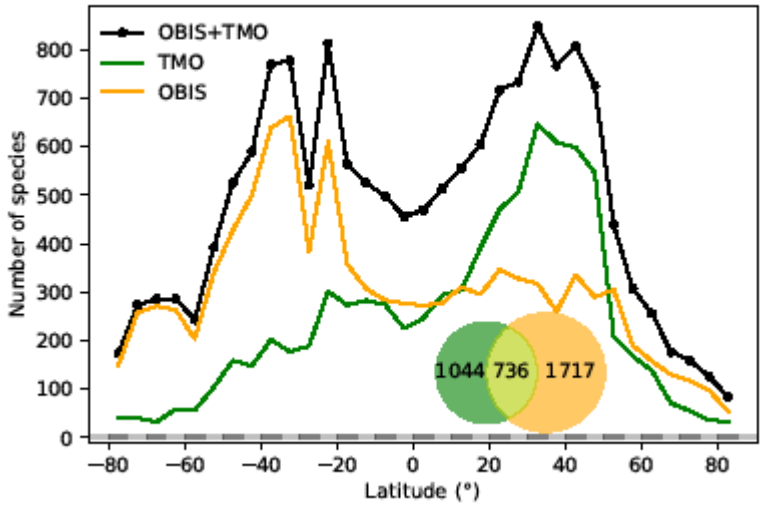
complexes (Fehlauer-Ale et al., 2014; Jackson & Cheetham, 1990; Lidgard & Buckley, 1994). While the portion of TMO data that is derived from the taxonomic literature may be less plagued by taxonomic misidentifications, the same cannot be easily argued for faunal lists or ecological surveys, which may also be part of OBIS data. However, in our experience, broad inferences based on synoptic, large-scaled databases tend to change significantly with different models, more so than data updates (Liow, Reitan, & Harnik, 2015; Sepkoski, 1993).

In terms of our text-mining task, we found that generating and classifying species-location candidates here is more challenging than classifying species-age candidates (Kopperud et al., 2019). An F1 result of about 77.5% is not uncommon for relation extraction studies (Henry, Buchan, Filannino, Stubbs, & Uzuner, 2020; Kim, Kim, & Lee, 2019), especially for datasets with low label assignment repeatability. Nonetheless, while the accuracy of the machine-classifier is less sensitive than human evaluation, its FPR is substantially lower than a null model. Note that the classifier merely provides a probabilistic measure of whether the sentence provides evidence that a species is present at a geographic location. In the event of a false positive, it is still possible that the species is actually present in that particular location. On the other hand, there is a wealth of species mentions for which we were not able retrieve any species-location candidates (Fig. S5). It is possible to extend our approach by considering cross-sentence candidates (Gupta, Rajaram, Schütze, & Runkler, 2019), although these methods are usually less accurate. Alternatively, we could go beyond standard NLP tools, which are relatively flexible and easy to adopt, and use non-linguistic features (such as tables and spatial layout) for information extraction, as has been suggested in the knowledge base creation literature (Schlichtkrull et al., 2018). However, methods for information extraction that combine linguistic and non-linguistic features are still at an early stage of development.

The main advantage of automatic information retrieval over collaborative data-entry is that of time and resource investment. The information retrieval procedure is largely independent of the size of the literature, or the taxonomic scope, say for cheilostomes versus all metazoans. Public biodiversity inventories such as GBIF and OBIS require large consortia and networks of research factions to contribute their data. Conversely, there is a wealth of biodiversity knowledge available in the published literature, and it is feasible for one person or a small team to extract substantial amounts of data quickly using automated information retrieval. We have used some supervised classification methods, which require us to generate training data. However as NLP is adopted in the biodiversity literature, it will become easier to use distantly supervised relation extraction (Hirschberg & Manning, 2015).

Biodiversity inventories such as OBIS are vital for supplying data for inferences of global biogeographic patterns. While we strongly support the continued development of these databases, we demonstrated that our automated information retrieval approach can enhance such inventories when answering global-scale questions, especially for under-studied taxa. To understand how the spatial diversity of cheilostomes has come to be will require continued and concerted efforts in taxonomic investigations (Bock & Gordon, 2013), compilation of more spatial data especially in areas currently devoid of deposited information (Klein et al., 2019), tool-development in automated data retrieval (Kopperud et al., 2019), and continued research in molecular phylogenetics (Orr et al., 2019).

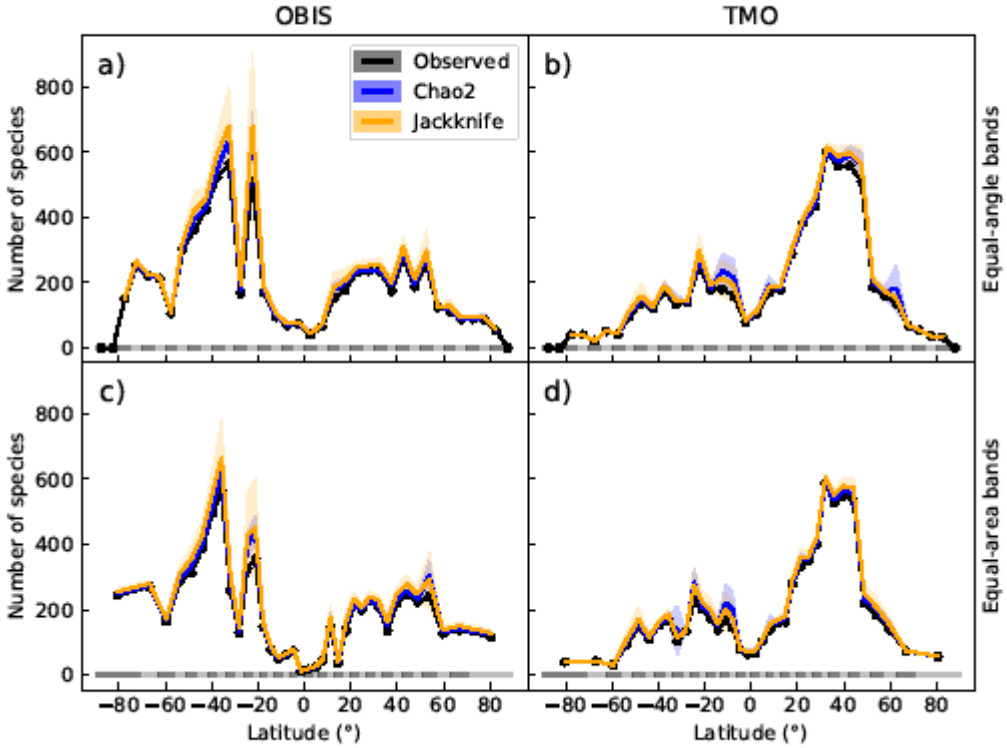
328 **Figure Captions**



329

330 **Fig. 1 Global range-through latitudinal species richness for cheilostome bryozoans.** The black line
331 shows combined Ocean Biogeography Information System (OBIS) and text-mined occurrence (TMO)
332 richness, and orange and green curves show range-through richness for OBIS and TMO separately.
333 The inset is a Venn-diagram showing the global overlap in species between OBIS and TMO.

334



335

336 **Fig. 2 Global latitudinal species richness for cheilostome bryozoans, estimated using Chao2 and**
337 **Jackknife.** The top panels show richness for Ocean Biogeography Information System (OBIS) and text-
338 mined occurrences (TMO) data in 5° equal-angle latitudinal bands. The lower panels show the
339 equivalent in 5° equal-area latitudinal bands. Black lines show the observed richness, while blue and
340 orange lines show the Chao2 and Jackknife estimates, respectively. The shaded areas are 95%
341 confidence intervals. See Figs. S7 and S8 for alternative band and bin sizes.

342

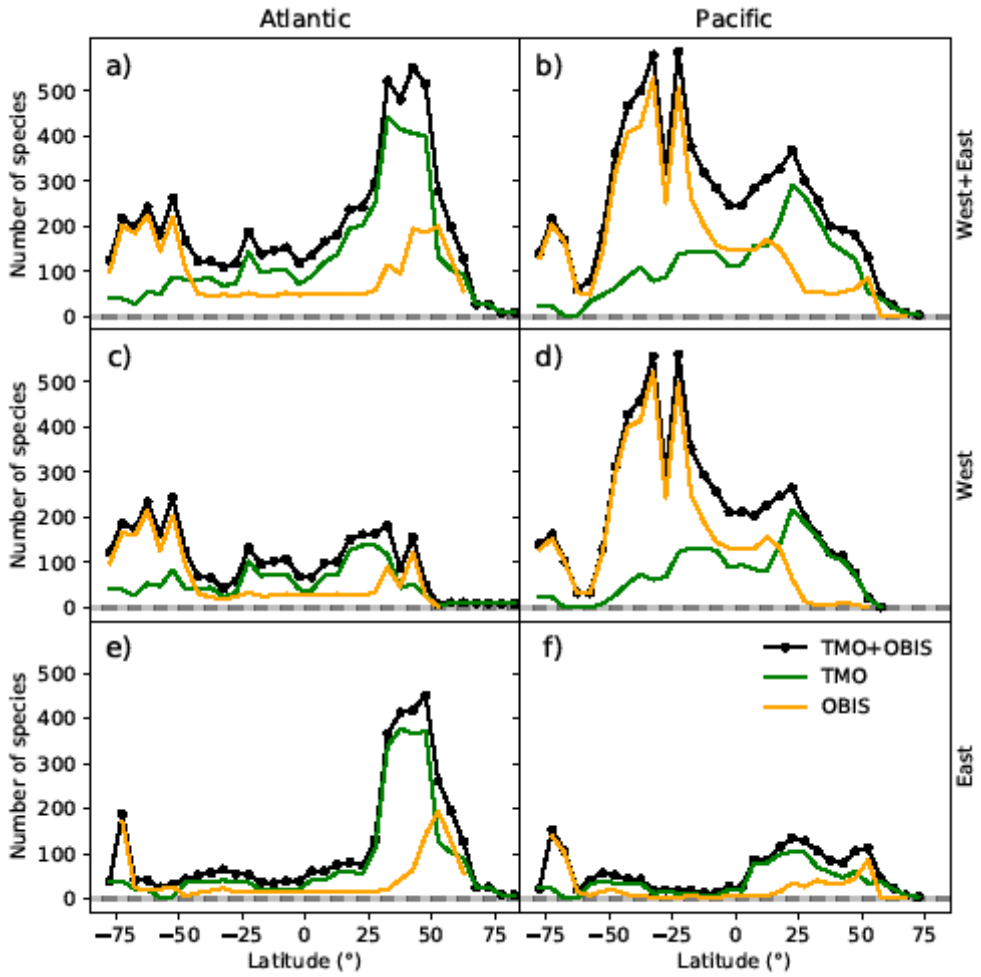


Fig. 3. **Range-through latitudinal species richness for cheilostome bryozoans in the Atlantic and Pacific Oceans.** The left column shows species richness in the Atlantic, and the right column shows that in the Pacific. The panel rows represent the eastern, western or the entire ocean basins. Orange and green lines represent Ocean Biogeography Information System (OBIS) and text-mined occurrences (TMO), respectively, and black lines are the joint data. Note that in this figure, the Atlantic borders Greenland and Iceland in the north, and the Antarctic in the south, but does not include the Gulf of Mexico, the Caribbean, the Baltic Sea or the Mediterranean. The Pacific borders the Bering Strait in the north, and includes the South China Sea, the Java Sea, north and east Australia, Tasmania as well as the Antarctic border.

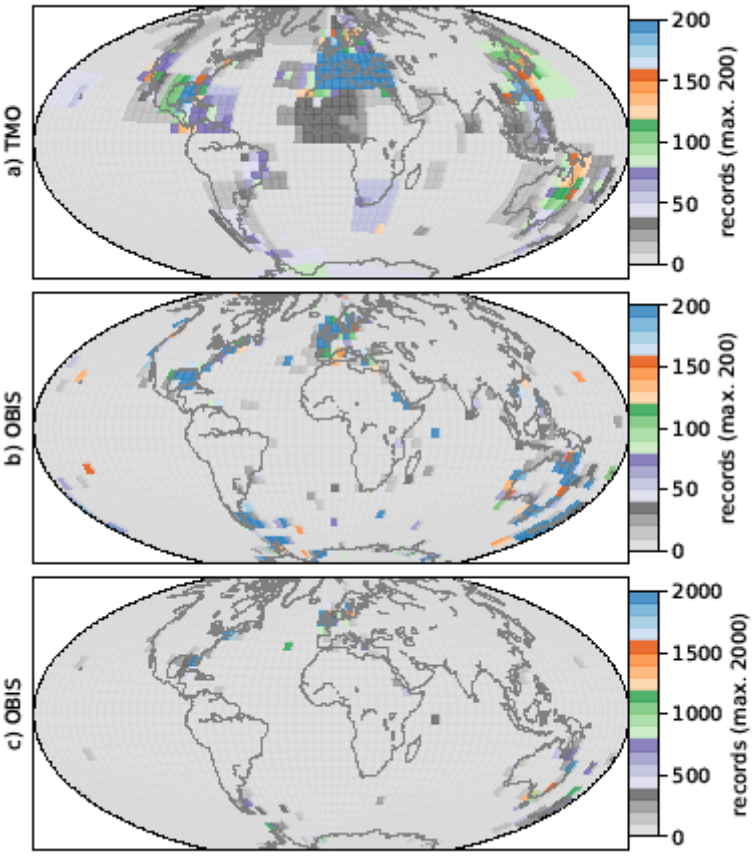


Fig. 4. Heatmaps for cheilostome bryozoan occurrence records per 5° latitude by 5° longitude bins. The color axes are truncated for visualization purposes, to a maximum of 200, 200 and 2000 in a), b), c), respectively. There are about 750 maximum records per bin in the Mediterranean for the text-mined occurrences (TMO), and about 35000 maximum records in the British Isles for the Ocean Biogeography Information System (OBIS) data. The globe is plotted using the Hammer equal-area projection. See Fig. S11 for the same figure plotted using the plate carrée projection.

Supporting Information:

Appendix S1: Extended methods.

Appendix S2: Supplementary figures.

Appendix S3: Bibliographic references for TMO data.

Code and data supplement: Will be available on datadryad.org upon submission.

References

Barnes, D. K. A., & Griffiths, H. J. (2008). Biodiversity and biogeography of southern temperate and polar bryozoans. *Global Ecology and Biogeography*, *17*, 84–99.

Bock, P. (2020). *Recent and Fossil Bryozoa*. Retrieved from http://www.bryozoa.net/

Bock, P., & Gordon, D. P. (2013). Phylum Bryozoa Ehrenberg, 1831. In: Zhang, Z.-Q.(Ed.) Animal Biodiversity: An Outline of Higher-level Classification and Survey of Taxonomic Richness (Addenda 2013). *Zootaxa*, *3703*(1), 67–74. Retrieved from http://www.biotaxa.org/Zootaxa/article/view/zootaxa.3703.1.14/0

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Chang, A. X., & Manning, C. D. (2014). TokensRegex: Defining cascaded regular expressions over tokens. *Stanford University Computer Science Technical Reports*.

Chaudhary, C., Saeedi, H., & Costello, M. J. (2016). Bimodality of Latitudinal Gradients in Marine Species Richness. *Trends in Ecology & Evolution*, *31*(9), 670–676. https://doi.org/10.1016/j.tree.2016.06.001

Chaudhary, C., Saeedi, H., & Costello, M. J. (2017). Marine Species Richness Is Bimodal with Latitude: A Reply to Fernandez and Marques. *Trends in Ecology & Evolution*, *32*(4), 234–237. https://doi.org/10.1016/j.tree.2017.02.007

Christopoulou, F., Tran, T. T., Sahu, S. K., Miwa, M., & Ananiadou, S. (2020). Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *Journal of the American Medical Informatics Association*, *27*(1).

Clarke, A., & Lidgard, S. (2000). Spatial patterns of diversity in the sea: Bryozoan species richness in the North Atlantic. *Journal of Animal Ecology*, *69*(5), 799–814. https://doi.org/10.1046/j.1365-2656.2000.00440.x

391 Costello, M. J., Tsai, P., Wong, P. S., Cheung, A. K. L., Basher, Z., & Chaudhary, C. (2017). Marine
392 biogeographic realms and species endemicity. *Nature Communications*, *8*(1).
393 https://doi.org/10.1038/s41467-017-01121-2
394 De Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. (2014).
395 Universal Stanford dependencies: A cross-linguistic typology. *Language Resources and*
396 *Evaluation Conference*, *14*, 4585–4592.
397 Fehlauer-Ale, K. H., Mackie, J. A., Lim-Fong, G. E., Ale, E., Pie, M. R., & Waeschenbach, A. (2014).
398 Cryptic species in the cosmopolitan *Bugula neritina* complex (Bryozoa, Cheilostomata).
399 *Zoologica Scripta*, *43*(2), 193–205.
400 Fernandez, M. O., & Marques, A. C. (2017). Diversity of Diversities: A Response to Chaudhary, Saeedi,
401 and Costello. *Trends in Ecology & Evolution*, *32*(4), 232–234.
402 https://doi.org/10.1016/j.tree.2016.10.013
403 Figuerola, B., Barnes, D. K. A., Brickle, P., & Brewin, P. E. (2017). Bryozoan diversity around the
404 Falkland and South Georgia Islands: Overcoming Antarctic barriers. *Marine Environmental*
405 *Research*, *126*, 81–94. https://doi.org/10.1016/j.marenvres.2017.02.005
406 Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information
407 extraction systems by gibbs sampling. *Proceedings of the 43rd Annual Meeting on Association*
408 *for Computational Linguistics*, 363–370. Association for Computational Linguistics.
409 Florence, W. K., Hayward, P. J., & Gibbons, M. J. (2007). Taxonomy of shallow-water Bryozoa from
410 the west coast of South Africa. *African Natural History*, *3*, 1–58.
411 Gordon, D. P. (1999). Bryozoan diversity in New Zealand and Australia. In W. Ponder & D. Lunney
412 (Eds.), *The other 99. The conservation and biodiversity of invertebrates* (pp. 199–204).
413 Mosman: Transactions of the Royal Zoological Society of New South Wales.
414 Gordon, D. P., Bock, P., Souto-Derungs, J., & Reverter Gil, O. (2019). A bryozoan tale of two
415 continents: Faunistic data for the Recent Bryozoa of Greater Australia (Sahul) and Zealandia,
416 with European comparisons. *Australasian Palaeontological Memoirs*, *52*, 13–22.

417 Grischenko, A. V., Mawatari, S. F., & Taylor, P. D. (2000). Systematics and phylogeny of the
418 cheilostome bryozoan *Doryporella*. *Zoologica Scripta*, *29*, 247–264.
419 Gupta, P., Rajaram, S., Schütze, H., & Runkler, T. (2019). Neural relation extraction within and across
420 sentence boundaries. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 6513–
421 6520.
422 Henry, S., Buchan, K., Filannino, M., Stubbs, A., & Uzuner, O. (2020). 2018 n2c2 shared task on
423 adverse drug events and medication extraction in electronic health records. *Journal of the*
424 *American Medical Informatics Association*, *27*(1), 3–12.
425 https://doi.org/10.1093/jamia/ocz166
426 Hillebrand, H. (2004). On the Generality of the Latitudinal Diversity Gradient. *The American*
427 *Naturalist*, *163*(2), 192–211. https://doi.org/10.1086/381004
428 Hirose, M. (2017). Diversity of freshwater and marine bryozoans in Japan. In M. Motokawa & H.
429 Kajihara (Eds.), *Species Diversity of Animals in Japan* (pp. 629–649). Springer.
430 Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245),
431 261–266.
432 Iknayan, K. J., Tingley, M. W., Furnas, B. J., & Beissinger, S. R. (2014). Detecting diversity: Emerging
433 methods to estimate species diversity. *Trends in Ecology & Evolution*, *29*(2), 97–106.
434 Jablonski, D., Belanger, C. L., Berke, S. K., Huang, S., Krug, A. Z., Roy, K., … Valentine, J. W. (2013). Out
435 of the tropics, but how? Fossils, bridge species, and thermal ranges in the dynamics of the
436 marine latitudinal diversity gradient. *Proceedings of the National Academy of Sciences*,
437 *110*(26), 10487–10494.
438 Jablonski, D., Roy, K., & Valentine, J. W. (2006). Out of the tropics: Evolutionary dynamics of the
439 latitudinal diversity gradient. *Science*, *314*(5796), 102–106.
440 Jackson, J. B. C., & Cheetham, A. H. (1990). Evolutionary Significance of Morphospecies: A Test with
441 Cheilostome Bryozoa. *Science*, *248*(4955), 579–583.
442 https://doi.org/10.1126/science.248.4955.579

James, N. P., & Clarke, J. A. (Eds.). (1997). *Cool-water carbonates*. Tulsa, Oklahoma: Society for Sedimentary Geology.

Kim, J., Kim, J., & Lee, H. (2019). DigChem: Identification of disease-gene-chemical relationships from Medline abstracts. *PLoS Computational Biology*, *15*(5), e1007022.

Klein, E., Appeltans, W., Provoost, P., Saeedi, H., Benson, A., Bajona, L., … Bristol, R. (2019). OBIS Infrastructure, Lessons Learned, and Vision for the Future. *Frontiers in Marine Science*, *6*, 588.

Kopperud, B. T., Lidgard, S., & Liow, L. H. (2019). Text-mined fossil biodiversity dynamics using machine learning. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1901), 20190022. https://doi.org/10.1098/rspb.2019.0022

Krug, A. Z., Jablonski, D., & Valentine, J. W. (2007). Contrarian clade confirms the ubiquity of spatial origination patterns in the production of latitudinal diversity gradients. *Proceedings of the National Academy of Sciences*, *104*(46), 18129–18134.

Lidgard, S. (2008). Predation in marine bryozoan colonies: Taxa, traits and trophic groups. *Marine Ecology Progress Series*, *359*, 117–131.

Lidgard, S., & Buckley, G. A. (1994). Toward a morphological species concept in cheilostomates: Phenotypic variation in Adeonellopsis yarraensis (Waters). In P. J. Hayward, J. S. Ryland, & P. D. Taylor (Eds.), *Biology and Palaeobiology of Bryozoans* (pp. 101–105). Fredensborg: Olsen & Olsen.

Lindsay, D. J., Grossmann, M. M., Bentlage, B., Collins, A. G., Minemizu, R., Hopcroft, R. R., … Nishikawa, J. (2017). The perils of online biogeographic databases: A case study with the 'monospecific' genus *Aegina* (Cnidaria, Hydrozoa, Narcomedusae). *Marine Biology Research*, *13*(5), 494–512. https://doi.org/10.1080/17451000.2016.1268261

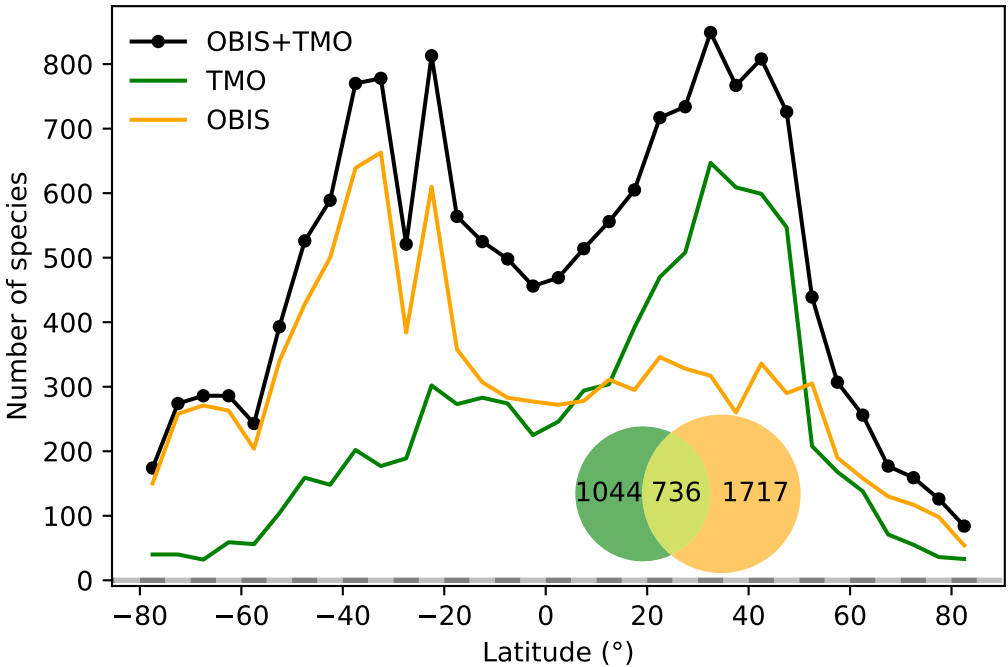
Liow, L. H., Reitan, T., & Harnik, P. G. (2015). Ecological interactions on macroevolutionary time scales: Clams and brachiopods are more than ships that pass in the night. *Ecology Letters*, *18*(10), 1030–1039.

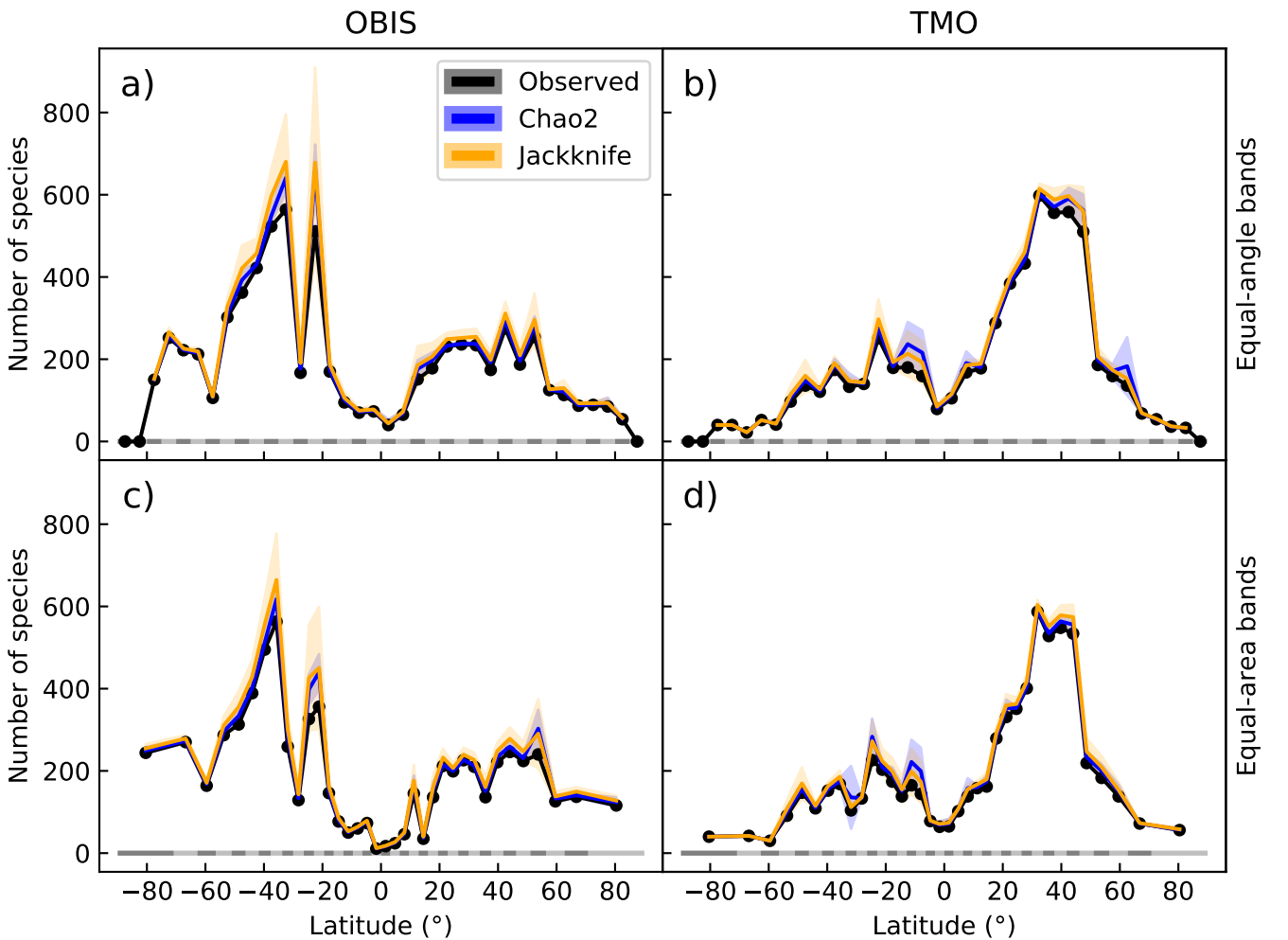
469 Liu, X., & Liu, H. (2008). Phylum Bryozoa [In Chinese]. In R. Liu (Ed.), *Checklist of Marine Biota of*
470 *Chinese Seas* (pp. 812–840). Qingdao: Science EP.
471 López Gappa, J. J. (2000). Species richness of marine Bryozoa in the continental shelf and slope off
472 Argentina (south-west Atlantic). *Diversity and Distributions*, *6*(1), 15–27.
473 Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford
474 CoreNLP natural language processing toolkit. *Association for Computational Linguistics (ACL)*
475 *System Demonstrations*, 55–60. Retrieved from
476 http://www.aclweb.org/anthology/P/P14/P14-5010
477 McGill, B. J. (2019). The what, how and why of doing macroecology. *Global Ecology and*
478 *Biogeography*, *28*(1), 6–17.
479 Menegotto, A., Kurtz, M. N., & Lana, P. C. (2019). Benthic habitats do show a significant latitudinal
480 diversity gradient: A comment on Kinlock et al. (2018). *Global Ecology and Biogeography*,
481 *28*(11), 1712–1717. https://doi.org/10.1111/geb.12970
482 Menegotto, A., & Rangel, T. F. (2018). Mapping knowledge gaps in marine diversity reveals a
483 latitudinal gradient of missing species richness. *Nature Communications*, *9*(1).
484 https://doi.org/10.1038/s41467-018-07217-7
485 Moyano, H. I. (1991). Bryozoa marinos chilenos VIII. Una síntesis zoogeográfica con consideraciones
486 sistemáticas y la descripción de diez especies y dos géneros nuevos. *Gayana Zoologia*, *55*(4),
487 305–389.
488 Okada, Y., & Mawatari, S. (1958). Distributional provinces of marine Bryozoa in the Indo-Pacific
489 region. In *Proceedings of the 8th Pacific Science Congress of the Pacific Science Association*
490 *1953* (Vol. 3, pp. 391–402). Quezon City: National Research Council of the Philippines.
491 Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., … Wagner, H. (2019).
492 *vegan: Community Ecology Package*. Retrieved from https://CRAN.R-
493 project.org/package=vegan

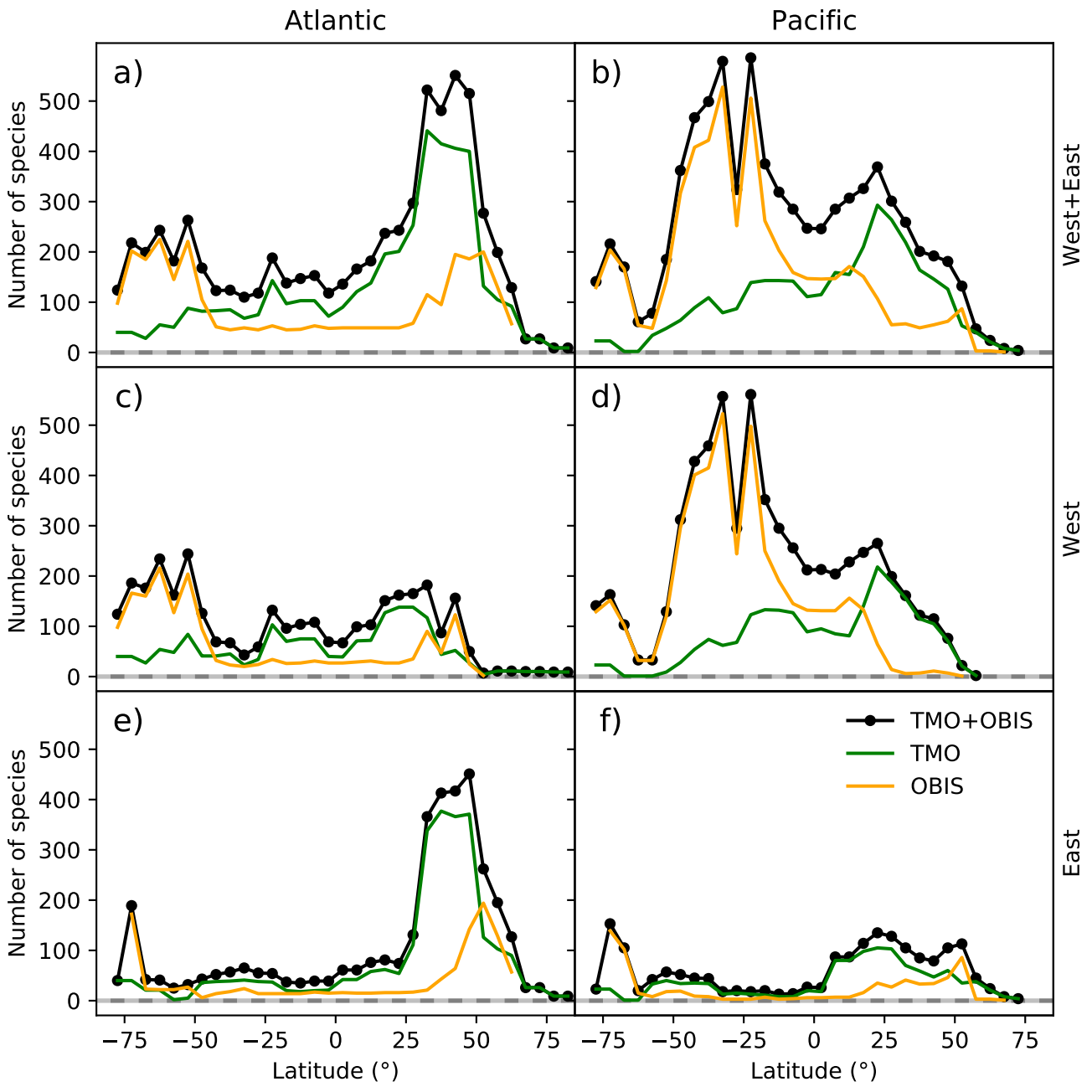
494 Orr, R. J., Haugen, M. N., Berning, B., Bock, P., Cumming, R. L., Florence, W. K., … others. (2019). A
495 genome-skimmed phylogeny of a widespread bryozoan family, Adeonidae. *BMC Evolutionary*
496 *Biology*, *19*(1), 1–10.
497 Page, R. D. M. (2019). Ozymandias: A biodiversity knowledge graph. *PeerJ*, *7*, e6739.
498 https://doi.org/10.7717/peerj.6739
499 Patterson, T. (2019). Free vector and raster map data. Retrieved November 27, 2019, from
500 www.naturalearthdata.com
501 Percha, B., Garten, Y., & Altman, R. B. (2012). Discovery and explanation of drug-drug interactions via
502 text mining. In *Biocomputing 2012* (pp. 410–421). World Scientific.
503 Peters, S. E., Husson, J. M., & Wilcots, J. (2017). The rise and fall of stromatolites in shallow marine
504 environments. *Geology*, *45*(6), 487–490. https://doi.org/10.1130/G38931.1
505 Provoost, P., & Bosch, S. (2020). *R client for the OBIS API*. Retrieved from
506 https://github.com/iobis/robis
507 Rabosky, D. L., Chang, J., Title, P. O., Cowman, P. F., Sallan, L., Friedman, M., … Alfaro, M. E. (2018).
508 An inverse latitudinal gradient in speciation rate for marine fishes. *Nature*, *559*(7714), 392–
509 395. https://doi.org/10.1038/s41586-018-0273-1
510 Reimer, J. D., Biondi, P., Lau, Y. W., Masucci, G. D., Nguyen, X. H., Santos, M. E. A., & Wee, H. B.
511 (2019). Marine biodiversity research in the Ryukyu Islands, Japan: Current status and trends.
512 *PeerJ*, *7*, e6532. https://doi.org/10.7717/peerj.6532
513 Rutherford, S., D'Hondt, S., & Prell, W. (1999). Environmental controls on the geographic distribution
514 of zooplankton diversity. *Nature*, *400*(6746), 749–753. https://doi.org/10.1038/23449
515 Saeedi, H., Dennis, T. E., & Costello, M. J. (2017). Bimodal latitudinal species richness and high
516 endemicity of razor clams (Mollusca). *Journal of Biogeography*, *44*(3), 592–604.
517 https://doi.org/10.1111/jbi.12903

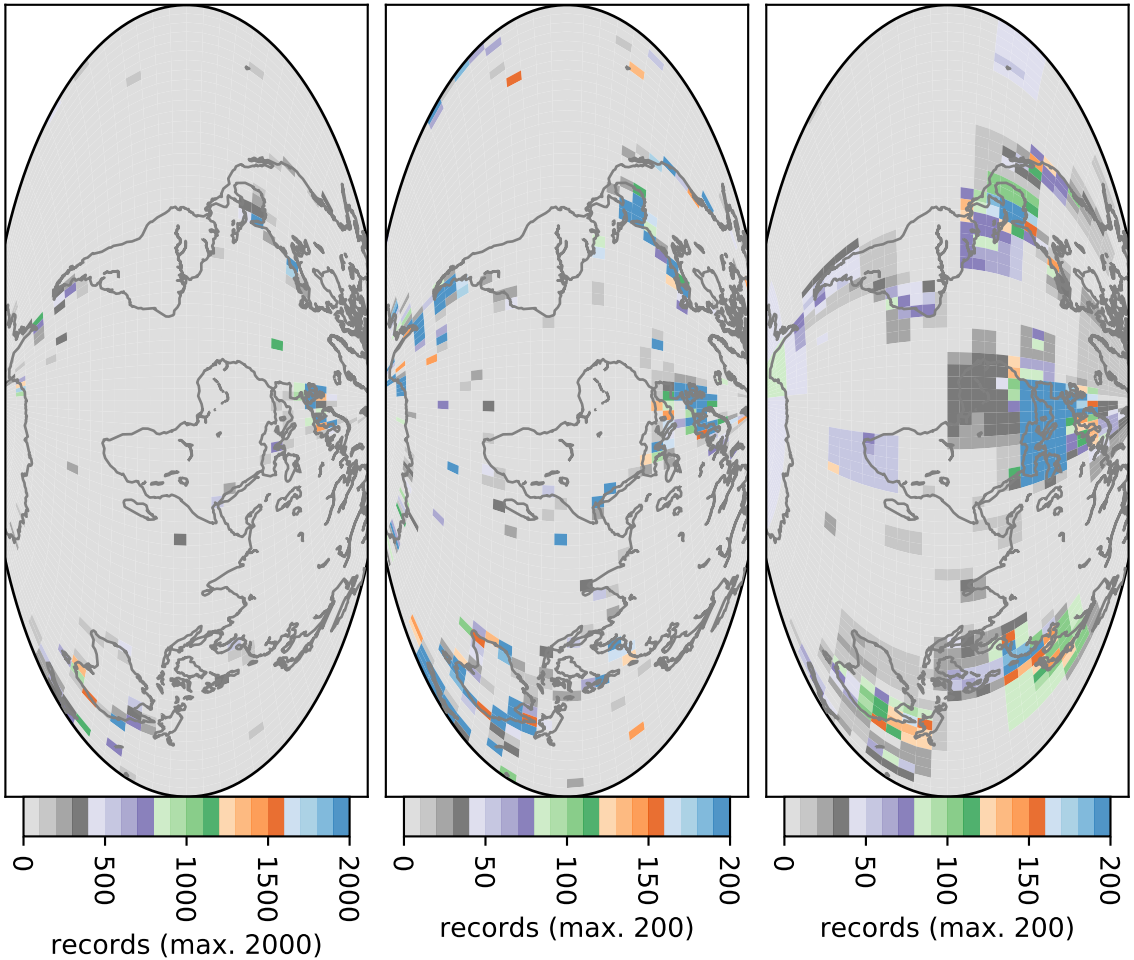
518 Saeedi, H., Reimer, J. D., Brandt, M. I., Dumais, P.-O., Jażdżewska, A. M., Jeffery, N. W., … Costello, M.
519 J. (2019). Global marine biodiversity in the context of achieving the Aichi Targets: Ways
520 forward and addressing data gaps. *PeerJ*, *7*, e7221.
521 Schlanger, S., & Konishi, K. (1975). The geographic boundary between the coral-algal and the
522 bryozoan-algal limestone facies: A paleolatitude indicator. *9th International Geological*
523 *Congress of Sedimentology, Nice, Theme*, *1*, 187–190.
524 Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den Berg, R., Titov, I., & Welling, M. (2018). Modeling
525 relational data with graph convolutional networks. *European Semantic Web Conference*,
526 593–607. Springer.
527 Schopf, T. J. M. (1970). Taxonomic diversity gradients of ectoprocts and bivalves and their geologic
528 implications. *Geological Society of America Bulletin*, *81*, 3765–3768.
529 Sepkoski, J. J. (1993). Ten years in the library: New data confirm paleontological patterns.
530 *Paleobiology*, *19*(1), 43–51.
531 Taylor, P. D., & Allison, P. A. (1998). Bryozoan carbonates through time and space. *Geology*, *26*(5),
532 459–462.
533 Taylor, P. D., & Di Martino, E. (2014). Why is the tropical Cenozoic fossil record so poor for
534 bryozoans? *Studi Trentini Di Scienze Naturali*, *94*, 249–257.
535 Tilbrook, K. J., & De Grave, S. (2005). A biogeographical analysis of Indo-West Pacific cheilostome
536 bryozoan faunas. In H. I. Moyano, J. M. Cancino, & P. N. Wyse Jackson (Eds.), *Bryozoan*
537 *Studies 2004* (pp. 341–349). Leiden: Balkema.
538 Tilbrook, K. J., Hayward, P. J., & Gordon, D. P. (2001). Cheilostomatous Bryozoa from Vanuatu.
539 *Zoological Journal of the Linnean Society*, *131*, 35–109.
540 Vieira, L. M., Migotto, A. E., & Winston, J. E. (2008). Synopsis and annotated checklist of Recent
541 marine Bryozoa from Brazil. *Zootaxa*, *1810*, 1–39.

542 Walther, B. A., & Moore, J. L. (2005). The concepts of bias, precision and accuracy, and their use in
543 testing the performance of species richness estimators, with a literature review of estimator
544 performance. *Ecography*, *28*(6), 815–829.
545 Winston, J. E. (1986). An annotated checklist of coral-associated bryozoans. *American Museum*
546 *Novitates*, *2859*, 1–39.
547 Wood, A. C. L., Rowden, A. A., Compton, T. J., Gordon, D. P., & Probert, P. K. (2013). Habitat-forming
548 bryozoans in New Zealand: Their known and predicted distribution in relation to broad-scale
549 environmental variables and fishing effort. *PLoS ONE*, *8*(9), e75160.
550 Woolley, S. N. C., Tittensor, D. P., Dunstan, P. K., Guillera-Arroita, G., Lahoz-Monfort, J. J., Wintle, B.
551 A., … O'Hara, T. D. (2016). Deep-sea diversity patterns are shaped by energy availability.
552 *Nature*, *533*(7603), 393–396. https://doi.org/10.1038/nature17937
553 WoRMS Editorial Board. (2020). *World Register of Marine Species*. Retrieved from
554 http://www.marinespecies.org/aphia.php?p=taxdetails&id=146142
555


OBIS+TMO
TMO
OBIS
Number of species
0
100
200
300
400
500
600
700
800
−80
−60
−40
−20
0
20
40
60
80
Latitude (°)
1044
736
1717


OBIS
TMO
a)
b)
c)
d)
Observed
Chao2
Jackknife
Number of species
Equal-angle bands
Equal-area bands
0
200
400
600
800
−80 −60 −40 −20 0 20 40 60 80
Latitude (°)


Atlantic
Pacific
a)
b)
c)
d)
e)
f)
Number of species
0
100
200
300
400
500
West+East
West
East
TMO+OBIS
TMO
OBIS
−75
−50
−25
0
25
50
75
Latitude (°)


c) OBIS
b) OBIS
a) TMO
0
500
1000
1500
2000
records (max. 2000)
0
50
100
150
200
records (max. 200)
0
50
100
150
200
records (max. 200)