How much biodiversity is concealed in the word “biodiversity”?

Stefano Mammola[1,2,3,4*],Caroline S. Fukushima[2], Girolama Biondo[4,5], Lucia Bongiorni[3,4,6], Fabio Cianferoni[4,7,8], Paolo Domenici[3,4,9,10], Carmelo Fruciano[3,4,11], Angelina Lo Giudice[4,12], Nuria Macías-Hernández[2,13], Jagoba Malumbres-Olarte[2,14], Marija Miličić[2,15], Michelangelo Morganti[3,4,16], Emiliano Mori[3,4,7], Ana Munévar[2,17], Paola Pollegioni[3,4,18], Ilaria Rosati[4,19], Simone Tenan[4,6], Fernando Urbano-Tenorio[2], Diego Fontaneto[1,3,4,†], Pedro Cardoso[2,14,†]

1. Molecular Ecology Group (MEG), Water Research Institute (CNR-IRSA), National Research Council Verbania Pallanza, Italy

2. Laboratory for Integrative Biodiversity Research (LIBRe), Finnish Museum of Natural History (LUOMUS), University of Helsinki, Helsinki, Finland

3. National Biodiversity Future Center, Palermo, Italy

4. Biodiversity Working Group (GDL Biodiversità), National Research Council, Rome, Italy

5. Istituto per lo studio degli impatti Antropici e Sostenibilità in ambiente marino (CNR-IAS), National Research Council, Campobello di Mazara, Italy

6. Institute of Marine Sciences (CNR-ISMAR), National Research Council, Venice, Italy

7. Research Institute on Terrestrial Ecosystems (CNR-IRET), National Research Council, Sesto Fiorentino (Florence), Italy

8. ‘La Specola’, Natural History Museum, University of Florence, Florence, Italy

9. Istituto di Biofisica (CNR-IBF), National Research Council, Pisa, Italy

10. Istituto per lo studio degli impatti Antropici e Sostenibilità in ambiente marino (CNR-IAS), National Research Council, Oristano, Italy

11. Institute for Marine Biological Resources and Biotechnology (CNR-IRBIM), National Research Council, Messina, Italy

12. Institute of Polar Sciences (CNR-ISP), National Research Council, Messina, Italy

13. Department of Animal Biology, Edaphology and Geology, University of Laguna, La Laguna, Tenerife 38206, Canary Islands, Spain

14. CE3C—Centre for Ecology, Evolution and Environmental Changes / Azorean Biodiversity Group and Universidade dos Açores, Angra do Heroísmo, Azores, Portugal

15. BioSense Institute—Research Institute for Information Technologies in Biosystems, University of Novi Sad, Novi Sad, Serbia

16. Water Research Institute (CNR-IRSA), National Research Council, Brugherio, Italy

17. IBS-Instituto de Biología Subtropical (UNaMCONICET), Puerto Iguazú, Misiones, Argentina

18. Research Institute on Terrestrial Ecosystems (CNR-IRET), National Research Council, Porano, Terni, Italy

19. Research Institute on Terrestrial Ecosystems (CNR-IRET), National Research Council, Lecce, Italy

* corresponding author: stefano.mammola@cnr.it ; stefano.mammola@helsinki.fi

† Shared last authors

Email & ORCID

Stefano Mammola: stefano.mammola@cnr.it ; ORCID: 0000-0002-4471-9055

Caroline S. Fukushima: carolinesayuri@gmail.com ; ORCID: 0000-0001-7909-0173

Girolama Biondo: girolama.biondo@ias.cnr.it; ORCID: 0000-0003-0437-6825

Lucia Bongiorni: lucia.bongiorni@cnr.it ; ORCID: 0000-0001-9033-4992

Fabio Cianferoni: fabio.cianferoni@cnr.it ; ORCID: 0000-0003-3170-0774

Paolo Domenici: paolo.domenici@cnr.it; ORCID: 0000-0003-3182-2579

Carmelo Fruciano: carmelo.fruciano@cnr.it ; ORCID: 0000-0002-1659-9746

Angelina Lo Giudice: angelina.logiudice@cnr.it; ORCID: 0000-0002-8842-083X

Nuria Macías-Hernández: nemacias@ull.edu.es; ORCID: 0000-0003-4759-3619

Jagoba Malumbres-Olarte: jagoba.malumbres.olarte@gmail.com ; ORCID: 0000-0002-6878-5719

Marija Miličić: marija.milicic@biosense.rs ; ORCID: 0000-0002-3154-660X

Michelangelo Morganti: michelangelo.morganti@cnr.it ; ORCID: 0000-0002-8047-0429

Emiliano Mori: emiliano.mori@cnr.it; ORCID: 0000-0001-8108-7950

Ana Munévar: a.munevar@conicet.gov.ar ; ORCID:0000-0003-2418-1160

Paola Pollegioni: paola.pollegioni@cnr.it ; ORCID: 0000-0001-6388-1931

Ilaria Rosati: ilaria.rosati@cnr.it; ORCID: 0000-0003-3422-7230

Simone Tenan: simone.tenan@cnr.it ; ORCID: 0000-0001-5055-9193

Fernando Urbano-Tenorio: fernandohurbanot@gmail.com ; ORCID: 0000-0002-4292-7651

Diego Fontaneto: diego.fontaneto@cnr.it; ORCID: 0000-0002-5770-0353

Pedro Cardoso: pedro.cardoso@helsinki.fi ; ORCID: 0000-0001-8119-9960

ABSTRACT

Amidst a global biodiversity crisis, the word “biodiversity” has become indispensable for practical conservation, including as a normative term. Yet, biodiversity is often used as a buzzword in scientific literature. Resonant titles promoting to have studied “global biodiversity” may then be used to oversell research that is narrow-focused on a limited sample of taxonomic groups, regions, or habitats. We selected a random sample of ~900 papers with the word “biodiversity” in their title to take a long view of the use and misuse of this term. We analyzed the degree to which studies actually consider different taxonomic groups and biodiversity facets and how all of this translates to the impact of a paper. As many as 22% of the articles used the term biodiversity in the title but did not measure it at any level. Among the articles sampling biodiversity directly, the proportion of biodiversity investigated was systematically low. We documented a decrease in the taxonomic scope of articles in recent years, especially those relying on big data. This is in stark contrast with the parallel advances in analytical tools, monitoring technologies, and the availability of data. Importantly, studies with general titles (i.e., using the word “biodiversity” without mentioning any taxa, habitat, or region) attract more citations and online attention (Altmetric), but only when they also have a wider taxonomic scope. Our results have broad ramifications for understanding how the extrapolation from studies with narrow taxonomic scope shapes our view of global biodiversity patterns and poorly informs conservation practices.

Keywords: Biological conservation, Research bias, Scientometrics, Scientific writing, Sixth mass extinction

INTRODUCTION

Global biodiversity is disappearing at an accelerating pace, not only from the physical world (1–3) but also from our minds (4, 5). Insofar as the long-term survival of humanity is intertwined with the natural world (6), preserving biodiversity in all its forms and functions is a central imperative of the 21st century (2, 7, 8). Consequently, the word "biodiversity" (a contraction of "biological diversity") has become indispensable for practical conservation, including as a normative term (9, 10). However, biodiversity remains an elusive concept, a constant matter of debate for biologists, ecologists, philosophers, economists, and conservation practitioners alike (9, 11–13). The Convention on Biological Diversity (14) states that biodiversity means "*the variability among living organisms from all sources […] this includes diversity within species, between species and of ecosystems.*" Although this and other definitions are inclusive—from gene to ecosystem services—a multifaceted view of biodiversity is rarely realized in scientific literature. As a result, science and policy intending to preserve biodiversity are often based on a fraction of its broad scope (15).

Indeed, notwithstanding the recent impulse in the digitalization of natural history data (16–18), we are still far from fully documenting the taxonomic diversity, phylogenetic diversity, and functions of all species on Earth (19). Growing evidence exists that research on biodiversity and its conservation is systematically biased in taxonomic, habitat, and geographic coverage, with similar biases operating hierarchically. At the organism level, research interests are often skewed toward vertebrate animals rather than invertebrates (20, 21), plants (22, 23), or fungi (24, 25). Furthermore, for all these groups, research and conservation efforts often correlate with aesthetic features (26–28), organismal complexity (29), cultural salience (30, 31), and phylogenetic proximity to humans (27), rather than extinction risk or ecological and socio-economic importance. At the habitat level, important blind spots for 'out of sight' systems exist (32)—for example, subterranean ecosystems like soils, aquifers, and caves are largely overlooked in global biodiversity and conservation agendas (33–35). Lastly, regions such as the Palearctic and Nearctic receive far more studies on extinction risk than others, biasing global patterns of conservation knowledge (15, 36, 37).

Importantly, the existence of similar biases is not always manifest or fully disclosed. Resonant titles promoting to have studied "global biodiversity" or invoking their comprehensiveness "across the Tree of Life" are sometimes used for overselling research that is actually narrow-focused on a limited sample of taxonomic groups, regions, or

habitats (38). This tendency may be problematic: in the long run, it leads to extrapolating results outside the target systems/taxa of these studies with direct consequences for our understanding of the ecology of life on Earth and its practical conservation (39, 40).

Here, we took a long view of how researchers have used the term “biodiversity” in scientific literature to understand its meaning and the consequences of its use. We gathered all articles listed in the *Web of Science* database that used the word "Biodiversity" in their title (Figure 1A). We randomly sampled ~10% of these papers (N = 916) and extracted detailed information on their geographical focus, methodologies, types of biodiversity facets considered (taxonomic, phylogenetic, functional, and other types of diversity), and the total number of organisms considered [at the level of Phylum/Division or higher categories for microorganisms (hereinafter “Phyla”)]. We used the latter information to calculate the number of Phyla that were considered in each study (“Observed biodiversity”) out of the total possible biodiversity (“Expected biodiversity”) (Figure 1B). Using regression-like analyses (41), we explored a number of interrelated questions:

i) How many papers using the word biodiversity in their title do actually measure biodiversity?

ii) How much biodiversity is sampled, on average, by these studies, and across what biodiversity facets?

iii) How does the sampled biodiversity vary over time and by regions?

iv) To which extent, when the sampled biodiversity is low, is this clarified in the title?

v) How do these factors affect the reach and impact of a given paper?

Owing to recent gigantic advances in analytical tools, monitoring technologies, and the availability of biodiversity variables and data (17, 42–45), we hypothesized that the biodiversity scope of papers should increase over time and should be higher in studies focusing on more biodiverse regions. If authors are not overselling their results, we also expect that papers with a narrow biodiversity scope should make it explicit in their title by descriptor terms specifying taxa, geographic regions or habitats of focus. Finally, we expected to observe a direct relationship between biodiversity scope and article impact (Figure 1C).

RESULTS

What proportion of biodiversity is sampled and across which biodiversity facets?

We managed to extract metadata for 851 full texts—we could not find and/or download 65 articles. The sampled proportion of biodiversity was systematically low across this random sample of literature. Of all papers, 22% did not consider any organism and were therefore excluded from further analyses—these were mostly theoretical papers. Across papers that actually considered or sampled biodiversity (N = 661), the proportion of biodiversity investigated by each study showed a highly skewed distribution, with most studies sampling a small proportion of biodiversity and a long tail of comparatively few studies sampling higher proportions of biodiversity (mean ± S.E.: 3.86% ± 0.15%; mode: 1.78%; range: 1.78–44.64%). Most of these studies did not distinguish between the different components of biodiversity, with only 19% of studies considering two, 4% three, and 0.6% four biodiversity facets. Taxonomic diversity was the most considered facet (84% of studies), followed by functional diversity (16%), (phylo)genetic diversity (12%), and other forms of diversity (8%)

How did the sampled biodiversity change through time?

We found no evidence of an increase in the sampled proportion of biodiversity in recent years (quasibinomial GLM; estimated β ± SE: 0.004 ± 0.007, p = 0.587), both considering the whole Tree of Life (Figure 2A) or calculating the proportion within major realms of life (Figure S1). The trend remain non-significant even considering only the edge of the data in the 75–100th percentile (Figure 2A; quasibinomial GLM; estimated β ± SE: 0.008 ± 0.009, p = 0.339).

How does the sampled proportion of biodiversity vary depending on different factors?

Next, we investigated the role of 11 factors in explaining the biodiversity sampled by each paper (Figure 3A). The model explained 23% of the variance. Sampled biodiversity was lower in studies focusing on the terrestrial realm compared to aquatic or multiple systems. At the biogeographic level, the lowest proportion of biodiversity was found in studies set in the Antarctic, Afrotropical, Indomalayan, and Nearctic regions (Figure 2B). Furthermore, low sampled biodiversity was associated with studies focusing on big data, studies focusing on phylogenetic diversity, and whenever a study mentioned the name of a taxon in the title alongside the word “biodiversity”. All other factors in the model exerted a

negligible effect on observed biodiversity. Across all these studies, the most commonly sampled taxa were Chordata (dominated by vertebrates) and arthropods, whereas microorganisms and fungi were the least studied (Figure 3B).

Next, we repeated the model on the subset of articles (N = 90) that mentioned no descriptors in the title. This second model explained 39% of the variance and largely confirmed the directions of the effects observed for the full model (*cf.* Figure 3A and Figure 3C). The key difference was a stronger effect for studies set in North America (Nearctic), which systematically sampled less biodiversity (Figure 3C). Furthermore, in this subsample of papers, the number of studies considering vertebrates was proportionally higher, suggesting that research on vertebrates is frequently oversold under the word "biodiversity" (Figure 3D).

How do all factors affect the reach and impact of articles?

We derived two measures of article impact—number of citations and Altmetric score—and tested how sampled biodiversity and the use of descriptors in the title affect impact (Figure 4). We interpreted the number of citations as a *proxy* measure for the impact of a given paper across the scientific community (46), and the Altmetric score as a *proxy* measure of the societal impact (47). In the model, we controlled for the number of countries of the coauthors and the Impact factor as confounding factors. In general, not mentioning descriptors led to more citations (Figure 4A) and societal attention (Figure 4C). All other things being equal, there was a positive effect of sampled biodiversity in interaction with the use of descriptors on impact. Whereas the impact of articles with more than one descriptor in the title was generally low, we found that articles with no descriptor or just one in the title attained greater impact when they sampled more biodiversity (Figure 4B, 4D).

DISCUSSION

Describing biological variation is central to natural sciences and beyond. Even though the concept of biodiversity has its origin in conservation biology (9), it is nowadays used in biological and non-biological disciplines and it crossed the academic walls to permeate political, management, and mass media discourses (13, 48). However, the term biodiversity has the curious quality of being widely used but rarely defined precisely (48). Here, we took a deep dive into the use of the term “biodiversity” in scientific literature, analyzing the degree to which studies consider different taxonomic groups and biodiversity facets in their analyses, and how all of this translates to the impact of a publication. Although our investigation has a coarse taxonomic resolution, we were able to identify clear trends and highlight important biases, with broad ramifications for understanding how the extrapolation from studies with narrow taxonomic scope may shape our view of global biodiversity patterns and inform conservation practices.

The first general finding of our study is that as many as 22% of the papers considered here used the term "biodiversity" in the title but did not measure it at any level. This suggests that biodiversity is often referred to as a normative and/or theoretical concept rather than a measurable phenomenon (10), and may in part be related to a publication strategy based on the use of trendy keywords (43, 49).

In the sample of studies that considered biodiversity (i.e., having a proportion of biodiversity > 0), we found no evidence of an increase in the sampled biodiversity in recent years. Some explanations for this pattern can be related to the fact that a few taxa are more likely to receive research funds and attention than others (31), to the point of hindering our ability to infer general biodiversity trends and patterns (e.g., IUCN red list data; (2, 20)). Furthermore, some taxa are easier to study due to their intrinsic or extrinsic characteristics (e.g., macroscopic size, large geographic range, and ease of sampling) and due to the greater availability of data. It is very likely that an unprecedented data baseline for certain groups—e.g., we now possess a complete phylogeny and compiled functional traits for all bird species (18)—will influence biodiversity research trajectories in the next decades, especially for broad-scale analyses.

The availability of large databases highlights another unexpected finding: studies using big data have narrower taxonomic scopes. One might expect that the availability of big data would allow us to study an increasing number of organisms, but in fact the opposite is true. There are multiple possible explanations for this pattern. First, the growing

availability of digital data does not immediately enable data synthesis (44). For example, a recent review of twelve major biodiversity databases suggested that variability in taxonomic and geographic scopes and potential incompatibility of metadata are major barriers to data integration (45). Second, many biases are intrinsic to biodiversity databases (17, 38, 50), suggesting that some of the data we need still await to be collected from the field (51), existing collections (52), or even "grey" literature (53, 54), all requiring gigantic human effort. Ultimately, it seems we are being flooded by repeated data on the same taxa [e.g., in 2020, vertebrates accounted for 68% of GBIF-available data; (17)], increasing biases in inadvertent ways (2). Current efforts to gather massive amounts of data should strive to explicitly decrease biases in what type and how data are collected and funding should be directed towards the knowledge gaps still remaining (55, 56), including undersampled taxa, regions, and habitats.

Studies on the Nearctic were also narrower in taxonomic scope, a pattern which emerged quite strikingly when restricting the analysis to studies without descriptors in the title. It is unclear why this is the case, but might be related to the fact that Nearctic species are overrepresented in databases focusing on certain taxa (15). Under the reasonable assumption that the geographical provenance of authors varies across biogeographical regions, cultural differences in writing style across countries may be also driving this pattern. That is, authors keen to extrapolate results from limited amounts of information may be overrepresented among those working on the Nearctic biota. Interestingly, studies focusing on some of the most biodiverse regions on the Earth, such as the Afrotropic and Indomalayan, were likewise narrow in their biodiversity scope.

Finally, studies dealing with vertebrates were more prone to use the word "biodiversity" in the titles, even without descriptors (Figure 5). As already discussed, vertebrates are often overrepresented in distribution, phylogenetic and functional datasets, being easier to obtain vast amounts of data and, in turn, publish comprehensive studies focusing on their taxa. It should be noted, however, that the use of vertebrate surrogates or umbrellas for inference to the whole spectrum of life forms is rarely justified and might lead to biased conclusions with implications for the conservation of broad biodiversity (2, 57, 58).

Ultimately, our results emphasize the need to think carefully when extrapolating from a few taxa, regions, or habitats to the full spectrum of living forms. Amidst the "publish or perish" academic culture, overselling results is demonstrably a good strategy to get

traction of one's own work. Indeed, while we do document a "reward" in the form of an increase in citation and Altmetric score when studies sampled a larger proportion of biodiversity, this occurs only when at most one descriptor is present in their titles. Even more strikingly, this modest reward pales in terms of costs/benefits when compared to simply omitting descriptors from the title. This is hardly surprising given the current landscape of ever-increasing new articles (59), fast and short communication [e.g., advertising through tweets; (60)], and the decreasing attention span of readers (61). Search algorithms giving more weight to information prominently placed in a paper combined with the need for authors to find seemingly "general" references for their own articles may have a further confounding effect.

Importantly, overselling might come at the expense of conservationists or policymakers, which may assume that results are generalizable beyond what they intend to or what they really reveal about patterns and processes. In the long run, this generalization may turn out detrimental for the majority of species and even ecosystem services on which we depend. This practice can misinform and misdirect conservation policies and actions by governments, organizations, and conservation practitioners at local or global scales, misallocating resources (23, 31) and perpetuating known biodiversity shortfalls (19). Notably, current biases are not being mitigated by new approaches using big data. If some parts of the biodiversity research can now largely be automated (42, 62, 63), others are still far from it, namely those that build on basic natural history knowledge in the most underexplored regions of the world, which often harbor the vast majority of biodiversity.

MOVING FORWARD

Here, we have pointed out many problems and knowledge gaps, leaving unanswered an important question: What is the way forward? Pragmatically, we see four main avenues warranting constant reflection and consideration:

i) We all, as scientists who routinely review manuscripts and/or handle manuscripts as editors, can play an active role in reducing blatant instances of "overselling", including in calling out manuscripts with too broad titles. These, as we document, may well produce a short-term positive effect on both the authors and the journal which publishes them in terms of citations and societal attention. However, it is doubtful that "overselling" would

serve the long-term goals of prestige and authoritativeness which any journal should strive for.

ii) The ability and expertise to recognize species are also fundamental for increasing the taxonomic scope of biodiversity databases and analyses based on them. This study adds quantitative evidence to a body of literature (64, 65) suggesting that it is time to reconsider the role of taxonomy and taxonomists within academia. As emphasized several times, the lack of a sufficient number of trained taxonomists, funds dedicated to this type of activity, and recognition of the importance of taxonomy all create a strong barrier to increasing the taxonomic coverage of studies.

iii) In most of the most biodiverse biogeographic regions the sampled proportion of biodiversity is systematically low. The underlying factors driving this pattern by region exceed the hypotheses proposed in this article. We can only reaffirm that constant support to local researchers and journals (66) and increased international scientific collaboration (67) are essential steps to improve the situation. This would enable a faster assessment of biodiversity unique for each region, including across less considered biodiversity facets (e.g., ecosystem services, cultural diversity and practices).

iv) The planning of big data generation, collection, storage, and sharing is still problematic, and possibly represents one of the major impediments towards integrated biodiversity research. Although tools for generating, aggregating and analyzing big datasets are increasingly available for the scientific communities, the nature of a complex multidisciplinary field such as biodiversity science imposes to define a common life cycle of data (45). This integration would enable the analysis of massive, multi-taxa datasets, a key step to achieving a global understanding of biodiversity patterns and comprehensive conservation strategies across all the branches of the Tree of Life (Figure 5).

METHODS

Data collection

On 22 May 2021, we queried the *Web of Science* core collection database for articles using the word “biodiversity” in their titles. We restricted the search to titles only given that they are the “hook” to readers (68, 69), the element of a paper that is most commonly assessed by scientists while screening for relevant papers. Indeed, it is estimated that a researcher, on average, skims 1,100 titles a year but will then go on to read 97 full texts

only (70).

We used the search string *TI = "Biodiversity" AND DT = "Article" AND WC = "Ecology" OR "Soil Science" OR "Environmental Studies" OR "Environmental Sciences" OR "Marine & Freshwater Biology" OR "Multidisciplinary Sciences" OR "Paleontology" AND PY = 1986–2020*. Note that we restricted the search to general Web of Science categories (WC) pertaining to biodiversity, avoiding taxon-specific categories (e.g., “Entomology”, “Fisheries”, “Ornithology”) which would have biased the search toward articles dealing with restricted samples of organisms. We selected the year 1986 as a lower boundary for the search because the term “BioDiversity” was coined in 1986 by Walter G. Rosen during the organization of the "*National Forum on BioDiversity*" (Washington, D.C).

The initial search yielded 10,170 hits. From this database, we randomly sampled 916 articles to be analyzed.

Metadata extraction

We inspected the full text of the sampled articles to extract the relevant data for our analyses. Note that we could not access the full-text for 65 articles, resulting in a final sample size of 851.

For each study, we first scored the year of publication, the year(s) the study occurred, and the biogeographic region (“Global”, “Nearctic”, “Neotropical”, “Afrotropical”, “Palearctic”, “Indomalayan”, “Oceanian”, “Australasian”, and/or “Antarctic”) and ecological domain (“Terrestrial”, “Saltwater”, and/or “Freshwater”) of focus. We also noted the approach(es) taken by the authors to study biodiversity [“Field sampling” (data collected in the field), “Big data” (use of pre-collected data, e.g., from online databases such as GBIF), “Review/Opinion” (theoretical studies or reviews), and “Other” (none of the previous)]. A single study may include multiple biogeographic regions, domains, and methods.

For each study’s title, we marked (“yes” or “no”) whether, alongside the word “biodiversity”, it mentioned: i) taxa or organisms (e.g., “biodiversity of dragonflies“, “biodiversity of wildflowers”, “biodiversity of zooplankton”); ii) locality or geographic regions (e.g., “Indo-Pacific biodiversity”, “tropical biodiversity”); and iii) habitats (e.g., “biodiversity of deserts“, “biodiversity of coral reefs”, “benthic biodiversity”). We interpreted these variables as the “descriptors” of the title (Figure 1B, 1C). Therefore, for each study, the

number of descriptors varied from 0 to 3.

Concerning the facets of biodiversity, we marked (“yes” or “no”) whether a study considered: i) taxonomic diversity; ii) (phylo)genetic diversity; iii) functional diversity; and iv) other forms of diversity (e.g., cultural diversity). Next, we noted the different organisms considered in the study at the Phylum (for animals, plants, and fungi) or higher-order (for microorganisms) level. Given the frequently changing taxonomy of microorganisms, we simply scored whether a study considered “Protista” (an artificial category used in several studies), “Bacteria”, “Archaea”, and/or “Viruses”, showing that we are already biased towards larger organisms even when trying to disentangle such biases. As a backbone taxonomy for Phyla/Divisions, we followed (71) for Metazoa and (72) for Fungi. Regarding land plants, we adopted the traditional division into Bryophyta, Pteridophyta, Gymnospermae, and Angiospermae. Regarding algae, given the number of classifications adopted by different authors and the number of *incertae sedis* taxa, we decided to group it in just one category (“Algae”). We also included three generic categories for animals, fungi, and plants (“Animal_generic”, “Plant_generic”, “Fungi_generic”) to be used for general studies when taxa were not explicitly named.

We calculated the number of groups that were considered in the article as “Observed biodiversity”. We considered the sum of Phyla (56 groups) to be our reference pool of biodiversity (“Expected biodiversity”) and used this reference to calculate the sampled proportion of biodiversity for each article.

Scientometric factors

To explore the relationship between sampled biodiversity, use of descriptors, and impact of a given paper, we extracted two measures of article impact: i) the number of citations received by each paper on the Web of Science; and ii) the Altmetric score, a measure of the general attention that a scholarly article has received online. Furthermore, we selected three confounding factors that are well-known correlates of these measures of impact (47, 73–75): i) Journal Impact Factor at the year of publication, based on annual Journal Citation Reports by Clarivate Analytics; ii) the number of coauthors in a given paper; and iii) the diversity of countries represented in the author's list (i.e., the number of unique countries based on the author’s affiliations).

Data analysis

We carried out all analyses in R version 4.1.0 (76) and used the package 'stats' version 4.1.0 for modelling and 'ggplot2' version 3.3.4 (77) for visualizations. In all regression-type analyses, we followed the general protocol by Zuur & Ieno (41). For data exploration, we visually inspected variable distribution and presence of outliers, multicollinearity among predictors, and balance of factor levels (78). In regression models, we scaled continuous variables to facilitate convergence; we assessed differences between pairs of levels for categorical variables with *posthoc* tests using the R package 'emmeans' version 1.5.3 (79). In discussing results, we adopted an evidence-based language (80), whereby we focused on effect sizes, directions of effects, and explained variance rather than significance (for the sake of tradition, p-values are marked in the figures and exact model estimates can be found in Table S1–S4).

Predictors of sampled biodiversity

To evaluate whether studies are increasing their taxonomic scope in recent years, we modeled the relationship between the proportion of biodiversity and the year of publication with a quasibinomial regression. By visually inspecting the data, we noticed they presented a "triangular" distribution with most data concentrating around zero (i.e., low sampled biodiversity), and a minor fraction of outliers that visually seemed to increase in recent years (Figure 2A). Thus, we repeated the model by fitting two quantile regressions, one with the full set of data and another with the data in the 75–100th percentile.

Next, we explored the role of different factors in explaining the observed biodiversity (dependent variable). As a result of data exploration, we removed two extreme outliers from the dependent variable Observed biodiversity. These were two studies with sampled biodiversity of 22 and 25, alone defining the 25–100% percentiles of the variable and thus able to strongly inflate the regression coefficient estimation. We decided to exclude these observations rather than transforming the data because the response variable was our primary interest (78). No collinearity was detected among predictors. Finally, in the categorical variable "Domain", we created a new level "Aquatic" to balance the factor levels, merging the levels "Saltwater" and "Freshwater".

We fitted an initial model assuming a Poisson error structure and a log link function to achieve positive fitted values. The model had the formula (in R notation):

(eq. 1) Observed biodiversity ~ Publication year + Domain + Biogeography + Method + Phylogenetic diversity + Functional diversity + Other diversities + Mention of location in title + Mention of habitat in title + Mention of taxon/a in title

The model was overdispersed (dispersion ratio = 1.611; Pearson's Chi^2 = 918.199, $p < 0.001$). Therefore, we fitted a new model assuming a negative binomial distribution—i.e., a generalization of Poisson distribution which loosens the assumption that the variance should be equal to the mean.

To further explore if patterns were different for articles using general titles, we fitted a second Poisson model including only those articles that mentioned no descriptors in the title, using the same formula as *eq. 1* except for the variables related to descriptors, which were not included. In this case, the model was not overdispersed (dispersion ratio = 0.652; Pearson's Chi^2 = 48.238, p = 0.991).

Drivers of article impact

We tested for relationships between article impact (citation or Altmetric counts) and seven article-level predictors. As a result of data exploration, we excluded the number of coauthors as this variable was correlated with the number of coauthors' countries (Pearson's r = 0.63). Furthermore, we log-transformed observed biodiversity, Impact factor, and the number of countries of the coauthors to homogenize their distributions and deal with a few outliers. Given that old papers had more time to attract citations and Altmetric attention than recent ones, we obtained a measurement of citation and Altmetric counts unaffected by age. Following Mammola et al. (81), we fitted two Poisson generalized additive models, exploring the relationship between the measure of article impact and the age of the paper. We then extracted the Pearson residuals from the two models, and used the age-residual values for citations and Altmetric scores as the response variables in two linear mixed models with the following formula (in R notation):

(eq. 2) Article impact ~ Impact Factor + Number of countries of coauthors + Mention of location in title + Mention of habitat in title + Mention of taxon/a in title + Observed

biodiversity : N° of descriptors in title

Note that, in the model, we tested for the interaction between observed biodiversity and the number of descriptors used in the title (see Figure 1C). The Impact Factor and Number of countries of the co-authors were included as confounding factors. Specifically, by the design of the study, we assumed that articles with a greater number of coauthors and published in high-impact factor venues will, on average, achieve a greater impact (73–75).

AUTHOR CONTRIBUTION

Conceptualization: SM, CSF, PC

Data collection: all authors except SM and PC

Analysis: SM

Interpretation: all authors

Writing, first draft: SM

Writing, contributions: all authors

ACKNOWLEDGEMENTS

The author would like to acknowledge the Biodiversity Working Group of CNR and the Italian National Biodiversity Future Center for support.

CONFLICT OF INTEREST

None declared

DATA AND CODE AVAILABILITY

The database and R code to generate analyses and figures will be published in an Open repository upon acceptance in a peer-reviewed journal.

SUPPORTING INFORMATION

Table S1–S4

Figure S1

LITERATURE CITED

1. A. D. Barnosky, *et al.*, Has the Earth's sixth mass extinction already arrived? *Nature* **471**, 51 (2011).
2. R. H. Cowie, P. Bouchet, B. Fontaine, The Sixth Mass Extinction: fact, fiction or speculation? *Biol. Rev.* **97**, 640–663 (2022).
3. S. H. M. Butchart, *et al.*, Global biodiversity: Indicators of recent declines. *Science (80-.).* **328**, 1164–1168 (2010).
4. I. Jarić, *et al.*, Societal extinction of species. *Trends Ecol. Evol.* **37**, 411–419 (2022).
5. M. Soga, K. J. Gaston, Extinction of experience: the loss of human–nature interactions. *Front. Ecol. Environ.* **14**, 94–101 (2016).
6. M. Loreau, *et al.*, Biodiversity as insurance: from concept to measurement and application. *Biol. Rev.* **96**, 2333–2354 (2021).
7. L. J. Pollock, *et al.*, Protecting Biodiversity (in All Its Complexity): New Models and Methods. *Trends Ecol. Evol.* **35**, 1119–1128 (2020).
8. H. M. Pereira, *et al.*, Scenarios for global biodiversity in the 21st century. *Science (80-.).* **330**, 1496–1501 (2010).
9. C. H. Lean, Biodiversity Realism: Preserving the tree of life. *Biol. Philos.* **32**, 1083–1103 (2017).
10. S. Sarkar, "What Should 'Biodiversity' Be?" in *History, Philosophy and Theory of the Life Sciences*, E. Casetta, J. Marques da Silva, D. Vecchi, Eds. (Springer International Publishing, 2019), pp. 375–399.
11. S. Sarkar, Defining "Biodiversity"; Assessing Biodiversity. *Monist* **85**, 131–155 (2002).
12. H. M. Pereira, *et al.*, Essential biodiversity variables. *Science (80-.).* **339**, 277–278 (2013).
13. S. Díaz, Y. Malhi, Biodiversity: Concepts, Patterns, Trends, and Perspectives. *Annu. Rev. Environ. Resour.* (2022) https:/doi.org/10.1146/annurev-environ-120120-054300.
14. Convention on Biological Diversity, Article 2. Use of Terms. (2022).

15. M. A. Titley, J. L. Snaddon, E. C. Turner, Scientific research on animal biodiversity is systematically biased towards vertebrates and temperate regions. *PLoS One* **12**, e0189577 (2017).

16. J. E. Ball-Damerow, *et al.*, Research applications of primary biodiversity databases in the digital age. *PLoS One* **14**, e0215794 (2019).

17. H. J. Mason, M. J. T., N. Daniel, W. S. B., S. Dmitry, Data integration enables global biodiversity synthesis. *Proc. Natl. Acad. Sci.* **118**, e2018093118 (2021).

18. J. A. Tobias, *et al.*, AVONET: morphological, ecological and geographical data for all birds. *Ecol. Lett.* **25**, 581–597 (2022).

19. J. Hortal, *et al.*, Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annu. Rev. Ecol. Evol. Syst.* **46**, 523–549 (2015).

20. P. Cardoso, T. L. Erwin, P. A. V. Borges, T. R. New, The seven impediments in invertebrate conservation and how to overcome them. *Biol. Conserv.* **144**, 2647–2655 (2011).

21. S. R. Leather, Institutional vertebratism hampers insect conservation generally; not just saproxylic beetle conservation. *Anim. Conserv.* **16**, 379–380 (2013).

22. M. Balding, K. J. H. Williams, Plant blindness and the implications for plant conservation. *Conserv. Biol.* **30**, 1192–1199 (2016).

23. M. Adamo, *et al.*, Dimension and impact of biases in funding for species and habitat conservation. *Biol. Conserv.* **272**, 109636 (2022).

24. S. Jennifer, G. S. C., H. Danny, F. Giuliana, M. G. M., Include all fungi in biodiversity goals. *Science (80-.).* **373**, 403 (2021).

25. R. Oyanedel, A. Hinsley, B. T. M. Dentinger, E. J. Milner-Gulland, G. Furci, A way forward for wild fungi in international sustainability policy. *Conserv. Lett.* **22**, e12882 (2022).

26. P. Cardoso, Habitats Directive species lists: Urgent need of revision. *Insect Conserv. Divers.* **5**, 169–174 (2012).

27. A. Miralles, M. Raymond, G. Lecointre, Empathy and compassion toward other species decrease with evolutionary divergence time. *Sci. Rep.* **9**, 19555 (2019).

28. M. Adamo, *et al.*, Plant scientists' research attention is skewed towards colourful, conspicuous and broadly distributed flowers. *Nat. Plants* **7**, 574–578 (2021).

29. J. R. U. Wilson, P. Şerban, B. Braschler, E. S. Dixon, D. M. Richardson, The authors reply. *Front. Ecol. Environ.* **6**, 299–300 (2008).

30. J. R. U. Wilson, S. Proches, B. Braschler, E. S. Dixon, D. M. Richardson, The (Bio)diversity of Science Reflects the Interests of Society. *Front. Ecol. Environ.* **5**, 409–414 (2007).

31. S. Mammola, *et al.*, Towards a taxonomically unbiased European Union biodiversity strategy for 2030. *Proc. R. Soc. B Biol. Sci.* **287**, 20202166 (2020).

32. Á. Delso, J. Fajardo, J. Muñoz, Protected area networks do not represent unseen biodiversity. *Sci. Rep.* **11**, 12275 (2021).

33. C. A. Guerra, *et al.*, Tracking, targeting, and conserving soil biodiversity. *Science (80-.).* **371**, 239–241 (2021).

34. D. Sánchez-Fernández, D. M. P. Galassi, J. J. Wynne, P. Cardoso, S. Mammola, Don't forget subterranean ecosystems in climate change agendas. *Nat. Clim. Chang.* **11**, 458–459 (2021).

35. C. Fišer, *et al.*, The European Green Deal misses Europe's subterranean biodiversity hotspots. *Nat. Ecol. Evol.* (2022) https:/doi.org/10.1038/s41559-022-01859-z.

36. M. Di Marco, *et al.*, Changing trends and persisting biases in three decades of conservation science. *Glob. Ecol. Conserv.* **10**, 32–42 (2017).

37. F. Chichorro, A. Juslén, P. Cardoso, A review of the relation between species traits and extinction risk. *Biol. Conserv.* **237**, 220–229 (2019).

38. C. S. Fukushima, S. Mammola, P. Cardoso, Global wildlife trade permeates the Tree of Life. *Biol. Conserv.* **247**, 108503 (2020).

39. A. C. Hughes, *et al.*, Sampling biases shape our view of the natural world. *Ecography (Cop.).* **44**, 1259–1269 (2021).

40. K. A. Wilson, *et al.*, Conserving biodiversity efficiently: What to do, where, and when. *PLOS Biol.* **5**, e223 (2007).

41. A. F. Zuur, E. N. Ieno, A protocol for conducting and presenting results of regression-type analyses. *Methods Ecol. Evol.* **7**, 636–645 (2016).

42. M. I. Tosa, *et al.*, The rapid rise of next-generation natural history. *Front. Ecol. Evol.* **9**, 480 (2021).

43. S. C. Anderson, *et al.*, Trends in ecology and conservation over eight decades. *Front. Ecol. Environ.* **19**, 274–282 (2021).

44. J. Cavender-Bares, *et al.*, Integrating remote sensing with ecology and evolution to advance biodiversity conservation. *Nat. Ecol. Evol.* (2022) https:/doi.org/10.1038/s41559-022-01702-5.

45. X. Feng, *et al.*, A review of the heterogeneous landscape of biodiversity databases: Opportunities and challenges for a synthesized biodiversity knowledge base. *Glob. Ecol. Biogeogr.* **31**, 1242–1260 (2022).

46. D. W. Aksnes, L. Langfeldt, P. Wouters, Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories. *SAGE Open* **9**, 2158244019829575 (2019).

47. A. C. Araujo, A. A. Vanin, D. P. Nascimento, G. Z. Gonzalez, L. O. P. Costa, What are the variables associated with Altmetric scores? *Syst. Rev.* **10**, 193 (2021).

48. A. J. Hamilton, Species diversity or biodiversity? *J. Environ. Manage.* **75**, 89–92 (2005).

49. S. Mammola, On deepest caves, extreme habitats, and ecological superlatives. *Trends Ecol. Evol.* **35**, 469–472 (2020).

50. J. Troudet, P. Grandcolas, A. Blin, R. Vignes-Lebbe, F. Legendre, Taxonomic bias in biodiversity data and societal preferences. *Sci. Rep.* **7**, 9132 (2017).

51. C. A. Ríos-Saldaña, M. Delibes-Mateos, C. C. Ferreira, Are fieldwork studies being relegated to second place in conservation science? *Glob. Ecol. Conserv.* **14**, e00389 (2018).

52. E. K. Meineke, T. J. Davies, B. H. Daru, C. C. Davis, Biological collections for understanding biodiversity in the Anthropocene. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **374**, 20170386 (2018).

53. N. R. Haddaway, H. R. Bayliss, Shades of grey: Two forms of grey literature important for reviews in conservation. *Biol. Conserv.* **191**, 827–829 (2015).

54. S. Chowdhury, *et al.*, Growth of non-English-language literature on biodiversity conservation. *Conserv. Biol.* **36**, e13883 (2022).

55. F. Ronquist, *et al.*, Completing Linnaeus's inventory of the Swedish insect fauna: Only 5,000 species left? *PLoS One* **15**, e0228561 (2020).

56. M. A. Bologna, *et al.*, Towards the new Checklist of the Italian Fauna 2022. *Biogeogr. – J. Integr. Biogeogr.* **37**, ucl001 (2022).

57. D. R. Stewart, Z. E. Underwood, F. J. Rahel, A. W. Walters, The effectiveness of surrogate taxa to conserve freshwater biodiversity. *Conserv. Biol.* **32**, 183–194 (2018).

58. S. K. Oberprieler, A. N. Andersen, G. R. Gillespie, L. D. Einoder, Vertebrates are poor umbrellas for invertebrates: cross-taxon congruence in an Australian tropical savanna. *Ecosphere* **10**, e02755 (2019).

59. E. Landhuis, Scientific literature: information overload. *Nature* **535**, 457–458 (2016).

60. Z. Fang, R. Costas, W. Tian, X. Wang, P. Wouters, How is science clicked on Twitter? Click metrics for Bitly short links to scientific publications. *J. Assoc. Inf. Sci. Technol.* **72**, 918–932 (2021).

61. P. D. B. Parolo, *et al.*, Attention decay in science. *J. Informetr.* **9**, 734–745 (2015).

62. T. T. Høye, *et al.*, Deep learning and computer vision will transform entomology. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2002545117 (2021).

63. P. Cardoso, *et al.*, Automated Discovery of Relationships, Models, and Principles in Ecology. *Front. Ecol. Evol.* **8**, 454 (2020).

64. S. Bacher, Still not enough taxonomists: reply to Joppa et al.. *Trends Ecol. Evol.* **27**, 65–66 (2012).

65. M. S. Engel, *et al.*, The taxonomic impediment: a shortage of taxonomists, not the lack of technical approaches. *Zool. J. Linn. Soc.* **193**, 381–387 (2021).

66. P. V Stefanoudis, *et al.*, Turning the tide of parachute science. *Curr. Biol.* **31**, R184–R185 (2021).

67. P. Cardoso, C. S. Fukushima, S. Mammola, Quantifying the internationalization and representativeness in research. *Trends Ecol. Evol.* **37**, 725–728 (2022).

68. S. M. Murphy, *et al.*, Does this title bug (Hemiptera) you? How to write a title that increases your citations. *Ecol. Entomol.* **44**, 593–600 (2019).

69. S. B. Heard, C. A. Cull, E. R. White, If this title is funny, will you cite me? Citation impacts of humour and other features of article titles in ecology and evolution. *bioRxiv*, 2022.03.18.484880 (2022).

70. M. A. Mabe, M. Amin, Dr Jekyll and Dr Hyde: author□reader asymmetries in scholarly publishing. *Aslib Proc.* **54**, 149–157 (2002).

71. C. Nielsen, *Animal Evolution: Interrelationships of the Living Phyla*, 3rd editio (Oxford University Press, UK, 2012).

72. M. A. Naranjo-Ortiz, T. Gabaldón, Fungal evolution: diversity, taxonomy and phylogeny of the Fungi. *Biol. Rev.* **94**, 2101–2137 (2019).

73. I. Tahamtan, L. Bornmann, What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018. *Scientometrics* **121**, 1635–1684 (2019).

74. I. Tahamtan, A. Safipour Afshar, K. Ahamdzadeh, Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics* **107**, 1195–1225 (2016).

75. S. Mammola, E. Piano, A. Doretto, E. Caprio, D. Chamberlain, Measuring the influence of non-scientific features on citations. *Scientometrics* **127**, 4123–4137 (2022).

76. R Core Team, R: A Language and Environment for Statistical Computing (2021).

77. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis.* (Springer-Verlag, 2016).

78. A. F. Zuur, E. N. Ieno, C. S. Elphick, A protocol for data exploration to avoid common statistical problems. *Methods Ecol. Evol.* **1**, 3–14 (2009).

79. R. V. Lenth, emmeans: Estimated Marginal Means, aka Least-Squares Means. (2021).

80. S. Muff, E. B. Nilsen, R. B. O'Hara, C. R. Nater, Rewriting results sections in the language of evidence. *Trends Ecol. Evol.* **37**, 203–210 (2021).

81. S. Mammola, D. Fontaneto, A. Martínez, F. Chichorro, Impact of the reference list features on the number of citations. *Scientometrics* **126**, 785–799 (2021).

FIGURES

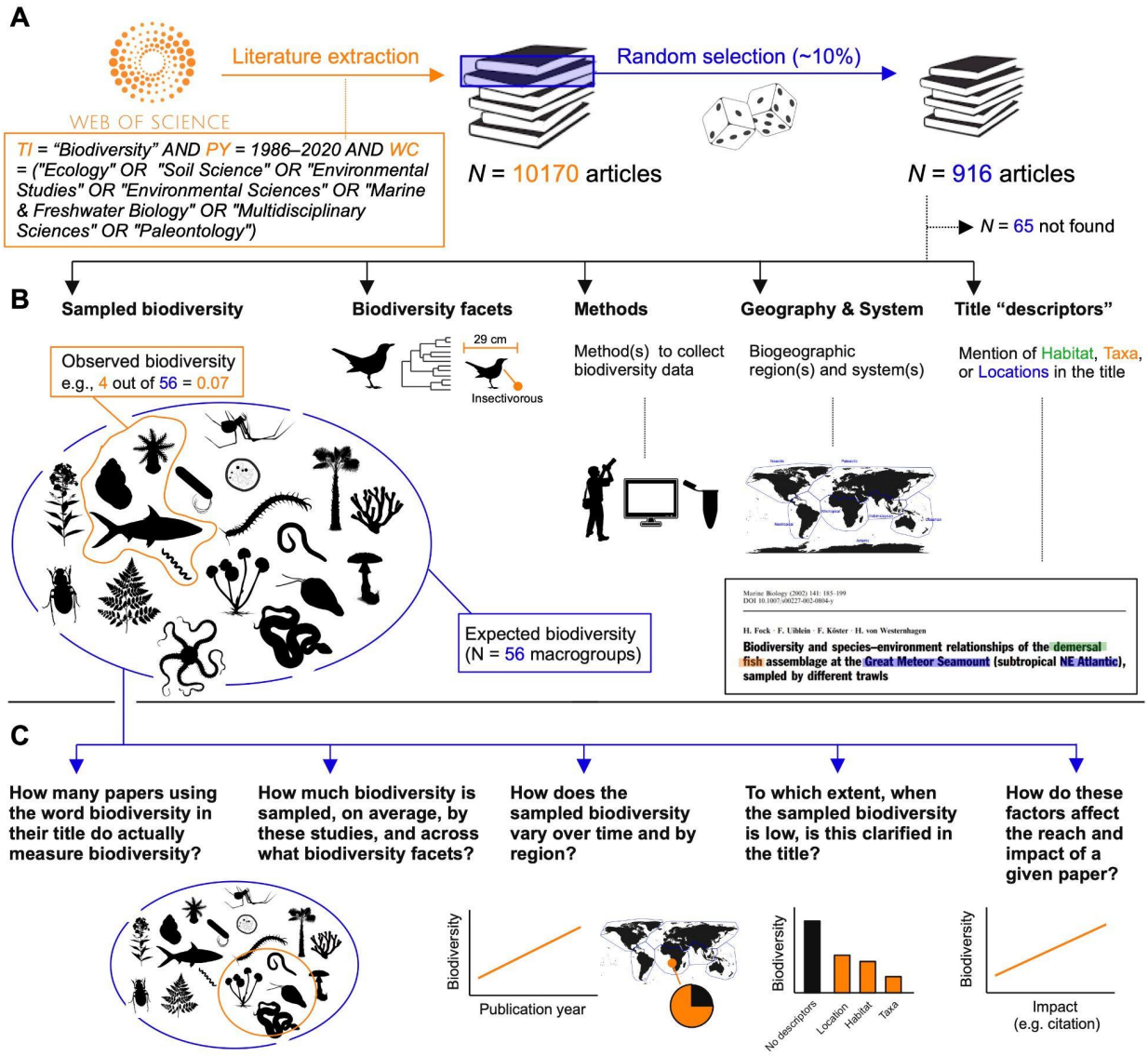

Figure 1. Infographic summarizing the study design. **A**) Literature sampling; **B**) Summary of the main variables extracted from each paper; **C**) Summary of the research questions and hypotheses.

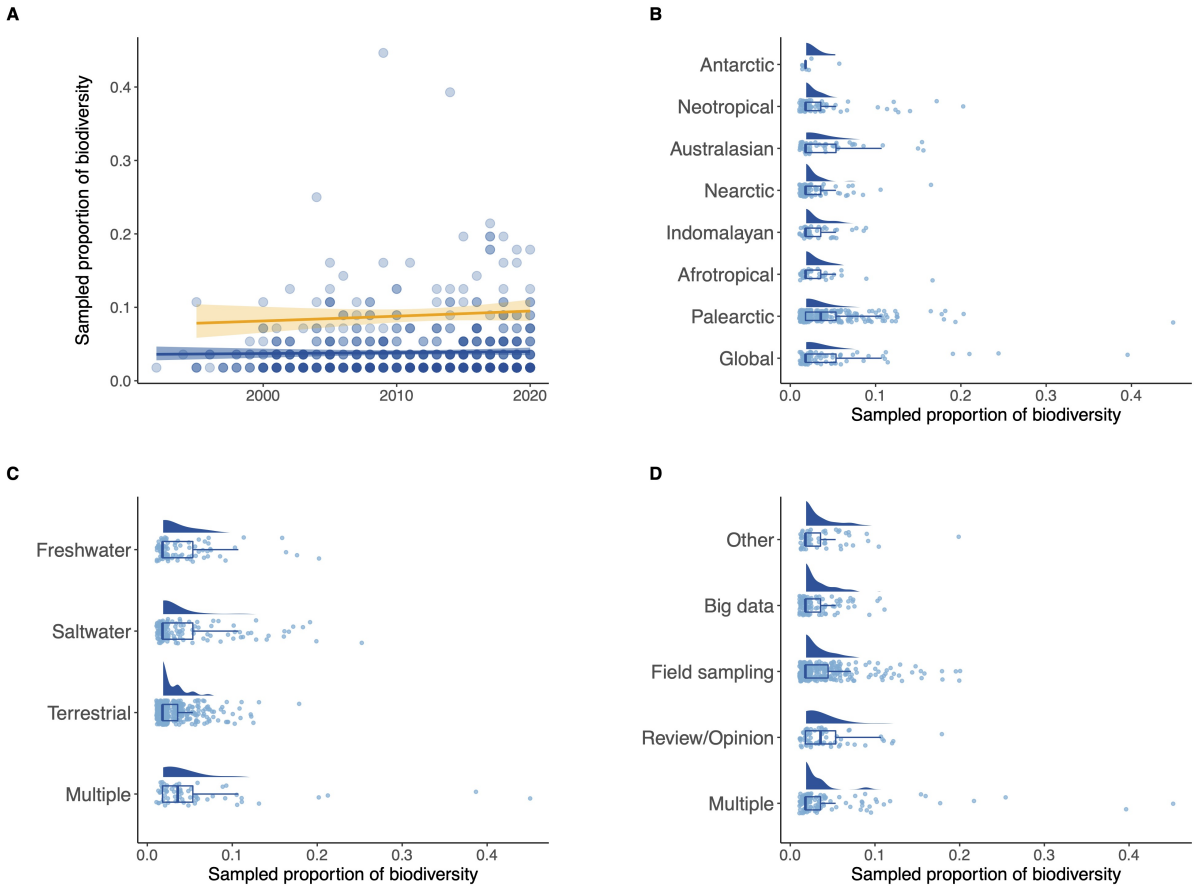

Figure 2. Change in biodiversity proportion over time and across regions, systems, and methodologies in articles with sampled biodiversity > 0. Due to the proximity of several values, most points are superimposed. **A**) Annual variations in the proportion of biodiversity considered in each study. Regression lines: in blue, full data; in orange, only data in the 75–100th percentile. **B–D**) Breakdown of biodiversity proportion by biogeographic regions, systems, and research methods. Jittered points are the actual values, boxplots summarize median and quantiles, and density plots summarize data distribution.

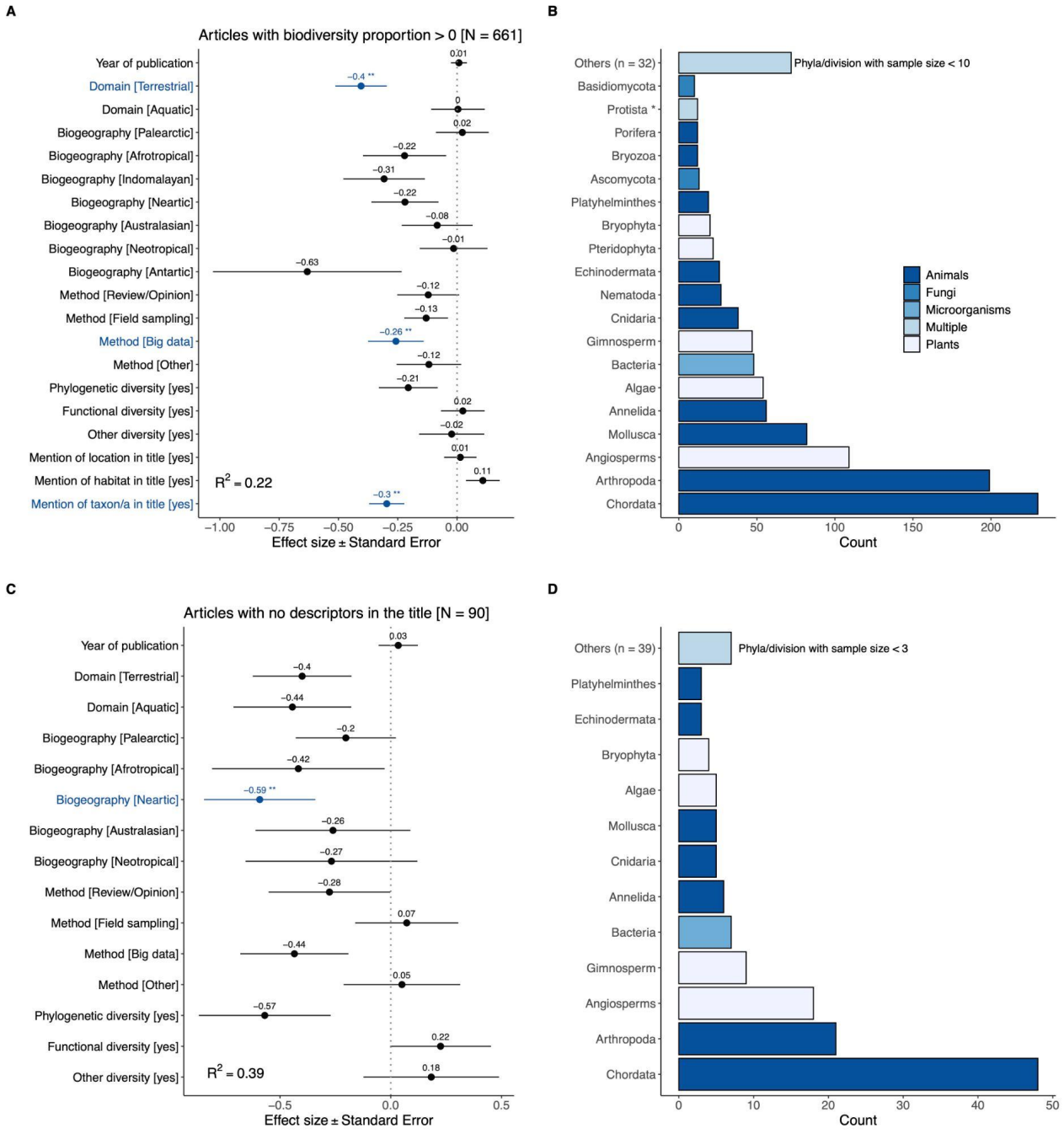

Figure 3. Predictors of sampled biodiversity across studies. **A**) Estimated parameters for a negative binomial generalized linear model testing the relationship between sampled biodiversity and different predictors (Table S1). **B**) Most commonly investigated biodiversity groups across studies. **C**) Estimated parameters for a Poisson generalized linear model testing the relationship between sampled biodiversity and different predictors across studies that mentioned no descriptors in the title (Table S2). **D**) Most commonly investigated biodiversity groups across studies that mentioned no descriptors in the title. In **A** and **C**, models are based on studies with sampled biodiversity > 0. Error bars indicate standard errors. Significant values (*: < 0.05; **: < 0.01) are highlighted in blue. Reference categories: Domain [Multiple]; Biogeography [Global]; Method [Multiple]; Biodiversity facets / Mention of descriptors [no]. In **B** and **D**, scarcely sampled taxa are grouped in the category “Others”.

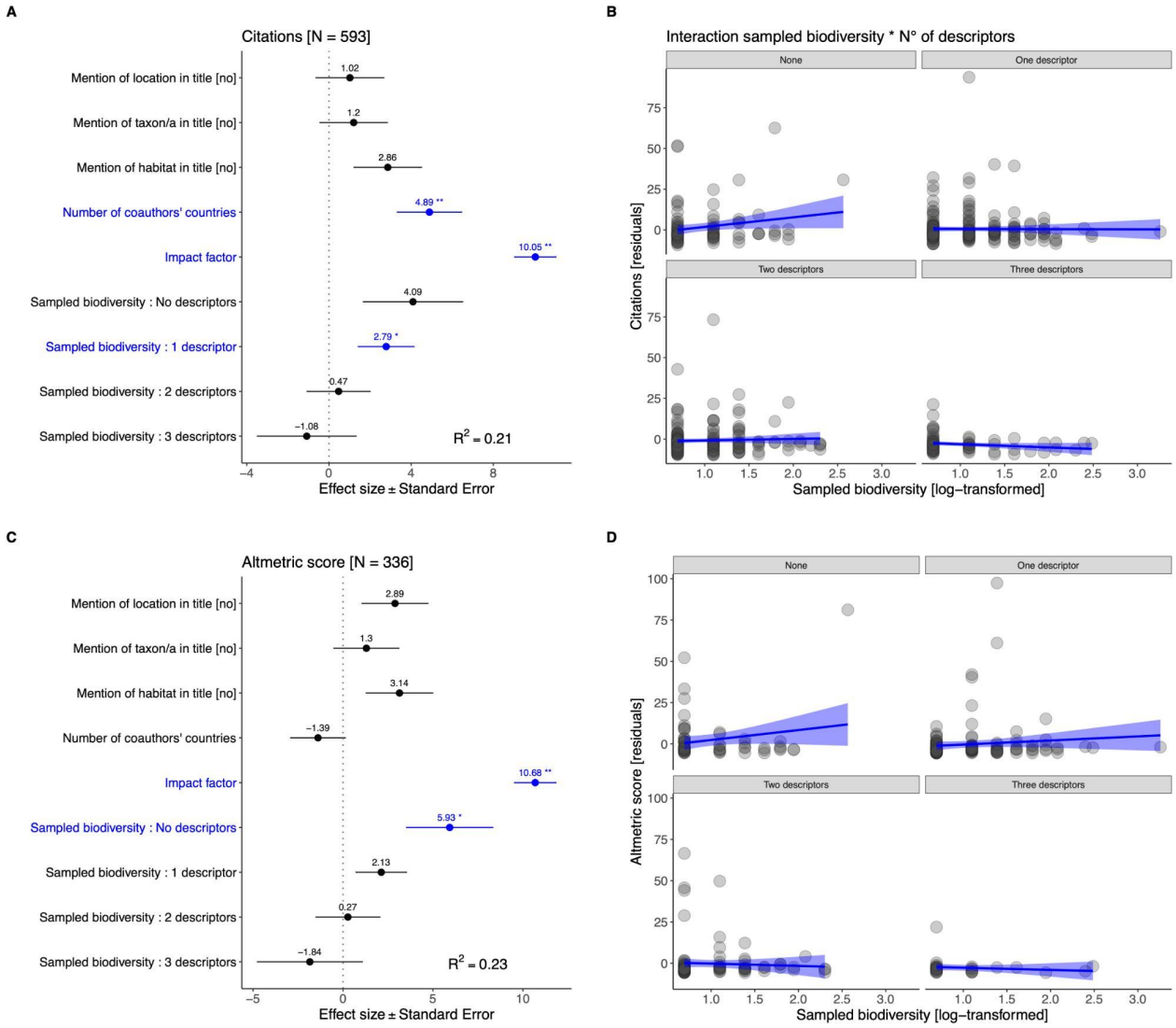

Figure 4. Drivers of article impact. **A**) Estimated parameters for a linear model testing the relationship between citations and different article-level predictors (Table S3). **B**) Visualization of the influence of the interaction between the number of descriptors and sampled biodiversity on citation counts. **C**) Estimated parameters for a linear model testing the relationship between Altmetric score and different article-level predictors (Table S4). **D**) Visualization of the influence of the interaction between the number of descriptors and sampled biodiversity on Altmetric scores. In **A** and **C**, error bars indicate standard errors. Significant values (*: < 0.05; **: < 0.01) are highlighted in blue. Reference categories: Mention of descriptors [yes]. In **B** and **D**, the blue lines are fitted values and shaded surfaces represent 95% confidence intervals.

Figure 5. There is a great disparity in the taxonomic scope of studies focusing on biodiversity, with vertebrates being the dominant component (especially in studies that do not use descriptors in the title; Figure 3D). This may be problematic if the extrapolation from studies with narrow taxonomic scope may bias and/or limit our view of the ecology of life on Earth and inform its conservation. Original illustration by Jagoba Malumbres-Olarte.