nature sustainability

Article

https://doi.org/10.1038/s41893-024-01392-w

Growing prominence of deep-sea life in marine bioprospecting

Received: 3 July 2023

Accepted: 25 June 2024

Published online: 8 August 2024

Check for updates

Erik Zhivkoplias[1], Jean-Baptiste Jouffray[1,2], Paul Dunshirn[3], Agnes Pranindita[1,4] & Robert Blasiak[1,5]

Marine bioprospecting, which involves the exploration of genetic and biochemical material from marine organisms, can be used towards addressing a broad range of public and environmental health applications such as disease treatment, diagnostics and bioremediation. Marine genetic resources are important reservoirs for such bioprospecting efforts; however, the extent to which they are used commercially for natural product discovery and the marine sources from which they are derived are not well understood. Here we introduce a comprehensive database of marine genes referenced in patent filings, the Marine Bioprospecting Patent database. It includes 92,550 protein-coding sequences associated with 4,779 patent filings, identified by analysing all relevant records from genetic sequence databases. Three companies alone—BASF, IFF and DuPont—included sequences from 949 species (more than half of referenced species with identified marine origin). Microbial life in the deep sea, a vast and remote biome predominantly beyond national jurisdiction, is already attracting substantial economic interest; the top ten patent holders have all filed marine gene patents referencing sequences from deep-sea life. Our findings provide an updated understanding of the marine bioprospecting landscape, contribute to the sustainable use of marine biodiversity and underscore the need for policymakers to ensure stewardship of deep-sea ecosystems.

Biodiscovery—the exploration and use of genetic and biochemical properties of biological materials—has a long and rich history. For instance, centuries before the discovery of penicillin from mould in a laboratory, skin diseases were already being treated in the Kingdom of Jordan via red soils with potent antibacterial properties that have only recently been confirmed[1]. Other examples include traditional medicines extracted from evergreen shrubs for cancer treatment[2], derivatives of the foxglove plant used to treat heart problems[3], anti-malarial quinine[4] and fungi-extracted podophyllotoxin to treat sexually transmitted diseases[5]. However, recent advances in genetics and sequencing innovations have spurred an unprecedented growth in the scale of discoveries. Today, bioprospecting—the search for potential products with scientific and industrial value derived from biological resources such as animals, plants and microorganisms—often involves large-scale screening, analysis and prediction of prospective biological compounds through the exploration of databases with sequencing data, including DNA extracted directly from environmental samples[6].

In this context, the ocean is considered a promising but largely untapped frontier for biodiscovery[7]. Marine organisms have evolved over millions of years to adapt to extreme conditions of temperature, salinity, light, pressure and water flow[8]. These conditions as well as a far longer evolutionary history have contributed to substantially greater taxonomic and functional diversity in marine habitats than in other biomes[9]. Nearly one million eukaryotic species are believed to inhabit the ocean[10], and the number of archaea and bacteria may be ten thousand times higher[11], yet most remain undescribed by science.

[1]Stockholm Resilience Centre, Stockholm University, Stockholm, Sweden. [2]Stanford Center for Ocean Solutions, Stanford, CA, USA. [3]Research Platform Governance of Digital Practices, University of Vienna, Vienna, Austria. [4]Bolin Centre for Climate Research, Stockholm University, Stockholm, Sweden. [5]Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo, Japan. e-mail: erik.zhivkoplias@su.se

Despite these knowledge gaps, marine biotechnology—the use of marine organisms and their compounds for a wide range of applications in industrial sectors—has managed to distinguish itself from the broader biotechnology landscape. For instance, while nearly half of the approved pharmaceuticals are based on biological compounds produced by living organisms, success rates are two to four times higher for compounds from marine organisms[7,12]. Annual sales and licensing revenues from marine drugs have exceeded US$1 billion annually since 2011[13], and prospects for greater commercial growth are substantial: in 2020 alone, more than 1,400 new compounds were isolated from marine species[14]. Biomolecules extracted from marine bacteria and other products developed from sequences of larger marine organisms are widely used in food production, diagnostics, bioremediation and disease treatment[15]. Some notable examples include the discovery of a thermostable enzyme required for the production of lactose-free milk in Archaea *Pyrococcus furiosus*[16], seawater cyanobacteria toxins developed into anticancer treatment products[17] and the extensive use of green fluorescent protein found in jellyfish *Aequorea victoria*[18] as a molecular marker, both in medical and diagnostic contexts and fundamental research.

Establishing a regulatory landscape that keeps pace with rapid advances in biotechnology, while also promoting transparency, equitable access and benefit-sharing mechanisms, has proven challenging[19]. The adoption of the Convention on Biological Diversity (CBD) in 1993 was a crucial milestone, as it defined genetic resources as 'any material of marine plant, animal, microbial or other origin containing functional units of heredity of actual or potential value', and established the fair and equitable sharing of benefits from their use as one of the convention's three core objectives[20]. In 2014, the convention's Nagoya Protocol provided a framework to regulate the access and benefit sharing of marine genetic resources (MGR) sampled in national jurisdictions[21]. Yet, some two-thirds of the ocean lies beyond national jurisdiction, and it was not until 2023, following protracted negotiations, that the 'High Seas Treaty' was agreed upon, including provisions to address MGR from areas beyond national jurisdiction (ABNJ)[22].

Despite these encouraging developments, the actual and potential value of MGR for marine bioprospecting remains poorly understood. Studies have focused on counting referenced marine species in patents[23] or GenBank[24], examining sequences in international patent applications[25–28] or exploring biological compounds for natural product discovery[29,30]. A common aspect to all these studies, however, is their lack of focus on the connection between the actors involved in the use of MGR and the potential sources for natural product discovery. They also suffer from limited information in patent and GenBank records about the geographical origin of gene sequences, which in many cases are referenced without naming the source species. The unevenness of these data presents a challenge for interpreting the true scale, scope and trajectory of marine bioprospecting.

Here we address these gaps by creating a comprehensive database of genetic sequences and related patent applications from 1989 to 2022 in marine bioprospecting. In addition to systematically compiling and presenting key data about the sequences, coded proteins, date of deposition and patent holders, we also address significant data gaps by developing and applying a BlastX sequence similarity model to consider sequences from unnamed species. We also assess the biodiversity data of species currently considered unique to ABNJ and highlight the special importance of deep-sea conservation for future biotechnology focused on the innovation and development of naturally derived products.

Results

Our analysis of patent filings revealed 29,065 nucleotide sequences from 1,474 disclosed marine species across 3,636 unique patents, representing approximately 1% of all gene patents submitted to the International Nucleotide Sequence Database Collaboration (INSDC).

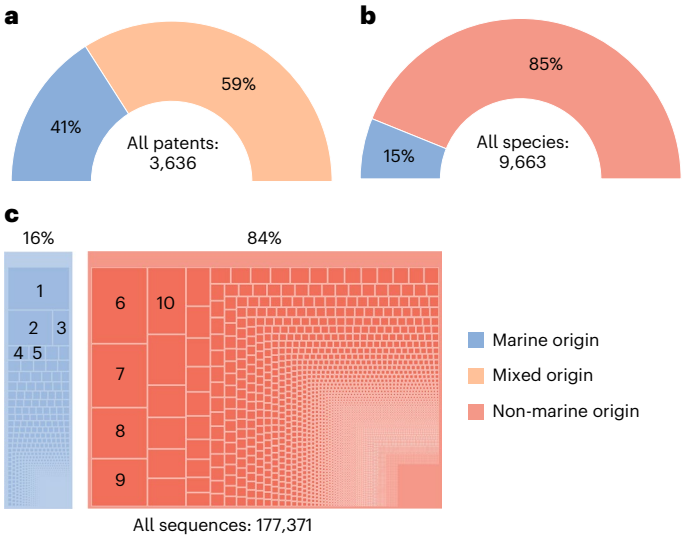

Fig. 1 | Patent applications associated with marine species. a, Number of patents that contain sequences associated with species of marine origin only, and patents with species from both marine and mixed origins. **b**, Number of unique marine and non-marine species referenced in patent applications that include at least one sequence of marine origin. **c**, Sequence frequency aggregated by species they originate from. Only 16% of all sequences attached to patent applications originate from marine species. The top five species with the highest frequency in each class are indicated. Marine species: (1) *C. intestinalis* (2.96%), (2) *Gadus morhua* (1.76%), (3) *Anguilla japonica* (0.67%), (4) *Salmo salar* (0.34%) and (5) *Oncorhynchus mykiss* (0.26%). Non-marine species: (6) *Arabidopsis thaliana* (4.31%), (7) *Zea mays* (3.61%), (8) *Glycine max* (2.87%), (9) *Homo sapiens* (2.71%) and (10) *Oryza sativa* (2.60%).

Many patents referenced multiple sequences, with a majority including both marine and non-marine sequences (Fig. 1a). Overall, marine sequences and species represented only 16% and 15%, respectively, of all sequences and species identified within the 3,636 patents (Fig. 1b,c). For comparison, approximately 242,000 marine species have been described to date (World Register of Marine Species (WoRMS), 2022), corresponding to roughly 10% of the 2.1 million species described by science[31]. This suggests considerable untapped potential of marine bioprospecting (Fig. 1b and Supplementary Table 1).

Types of sequence in marine gene patents

The patent applicants who referenced the highest number of unique genetic sequences included both protein-coding and non-coding sequences, with the former having a higher potential for natural product discovery (Fig. 2). Most of the companies with a large number of applications referenced protein-coding genes that originate from multiple species, with an average length between 500 and 2,000 nucleotides. Some applicants specifically focused on MGR from a single species and predominately referenced non-coding sequences. For instance, the Fisheries Research Agency of the National Research and Development Agency in Japan included 1,179 sequences in their patent applications, mostly originating from Japanese eel (*Anguilla japonica)*, yet only 127 are protein-coding sequences. Similarly, the Japan Science and Technology Agency has referenced 5,190 sequences from the sea vase tunicate (*Ciona intestinalis*), only 150 of which are protein-coding genes.

Most short non-coding sequences of identical length, originating from the same species, exhibit a wide range of GC content (that is, the percentage of two DNA basic building blocks), which is typical for artificially modified sequences used in amplification or as probes for detecting specific sequences of DNA or RNA (Supplementary Fig. 1). Out of all the patents that include at least one sequence from disclosed marine species, 71% contain nucleotide sequences that are potentially

https://doi.org/10.1038/s41893-024-01392-w

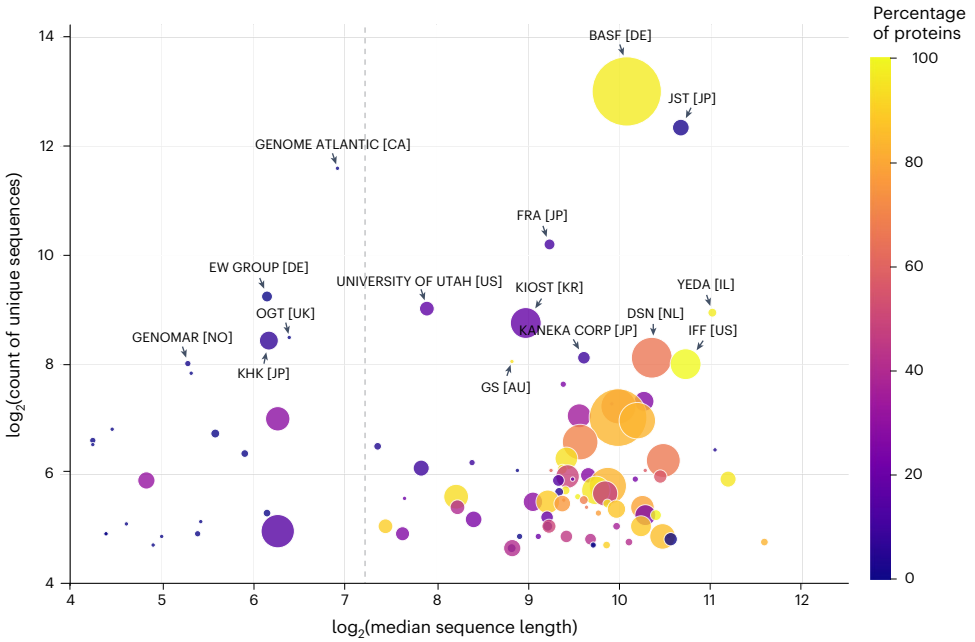

Fig. 2 | Key actors in marine biotechnology. Applicants that submitted at least 25 nucleotide sequences in their patent claims (81% of all sequences) are shown. Companies that submitted at least 250 sequences are indicated. The size of the dots represents the number of patents submitted by each applicant. The dotted grey line indicates the shortest protein length estimation (150 base pairs). The continuous colour bar indicates the percentage of protein-coding sequences submitted in applicant claims. BASF [DE], BASF; JST [JP], Japan Science and Technology Agency; GENOME ATLANTIC [CA], Genome Atlantic; FRA [JP], Fisheries Research Agency; EW GROUP [DE], EW Group GmbH; UNIVERSITY OF UTAH [US], The University of Utah; YEDA [IL], Yeda Research and Development Company Ltd.; KIOST [KR], The Korea Institute of Ocean Science and Technology; DU PONT [US], DuPont; OGT [UK], Oxford Gene Technology Ltd.; KHK [JP], Kyowa Kirin Co., Ltd.; DSM [NL], DSM N.V.; KANEKA CORP [JP], Kaneka Corporation; GS [AU], Gene Stream Pty Ltd.; GENOMAR [NO], GenoMar.

protein-coding genes. This suggests that most MGR are used in bioprospecting (Fig. 3a). For sequences of particular interest (that is, those submitted to all patent systems), we provide examples illustrating the conversion of DNA molecules into products of value (Box 1).

Marine Bioprospecting Patent database

While INSDC records provide considerable insight into the genes referenced in patents, only 37.3% of records include the name of source species, primarily filed under the World Intellectual Property Organization (WIPO), the European Patent Office, the Patent Office of Japan and the Korean Intellectual Property Office. Most of the remaining records are from the US Patent and Trademark Office, which does not share species names in its records (Supplementary Fig. 2).

To address this gap, we developed a sequence similarity model and BlastX search tool to query all genetic sequences with unknown origins against the UniProtKB protein sequence database. This model retrieved an additional 60,636 sequences, which can be said with a high degree of certainty to originate from marine organisms. Together with the 31,914 protein-coding sequences of confirmed marine species, this resulted in a comprehensive database of 92,550 sequences, which form the basis for all subsequent analysis in this paper and were used to construct the Marine Bioprospecting Patent (MABPAT) database (https://mabpat.shinyapps.io/main/).

Key actors in marine biotechnology

We found that 100 applicants accounted for 58% of all patents that contain protein-coding sequences with identified marine origin (that is, bioprospecting patents). The remaining 42% were associated with applicants who filed fewer than two patents on average. For companies in the top 100 (Supplementary Table 3), the total number of patent applications would have been underestimated by at least one-third if we had not applied the sequence similarity model. Transnational corporations (1,675 applications) are the most frequent type of applicant, although roughly one-fifth of filings are from research institutes and their commercialization centres (634 applications) (Fig. 3b). In total, 78% of all bioprospecting patents filed by the top 100 were submitted by actors headquartered in the USA, Germany or Japan (Fig. 3c).

The number of patents registered by each applicant is correlated with the total count of unique species included in such patents ($r = 0.8168$, $P = 2.17552 \times 10^{-318}$). To illustrate how much biological diversity each of these applicants is drawing upon, we connected patent holders and unique species included in patent claims and aggregated on the domain (Fig. 4a) and phylum level of biological taxonomy (Supplementary Fig. 6). For each flow diagram, we also added information if the corresponding marine species had been observed in a deep-sea environment. The most active users of MGR are primarily dependent on sequences from bacteria and archaea (Fig. 4a). The ten largest actors, including eight multinational corporations and two public research bodies (Fig. 4a), collectively registered more than one-third of all patents in the top 100. Deep-sea marine species have attracted interest from all ten of the largest users of MGR.

The opacity of marine bioprospecting in ABNJ

Issues of access and benefit sharing related to genetic material from ABNJ are of particular interest as they fall outside the scope of the Nagoya Protocol of the Convention on Biological Diversity and were at the core of negotiations for the High Seas Treaty adopted in June 2023. It is therefore notable that among 1,639 species of identified marine origin referenced in INSDC patent records, 281 species have been observed in ABNJ, with only 5 of them being exclusive. This contrasts with the 5,889 species found exclusively in ABNJ, predominantly from the Arthropoda, Foraminifera and Nematoda phyla, according to our analysis of species observation data available in the Ocean Biodiversity Information System (OBIS), a global open-access database on marine biodiversity (https://obis.org). The complete taxonomic distribution is given in Supplementary Fig. 7. According to the records from the World

https://doi.org/10.1038/s41893-024-01392-w

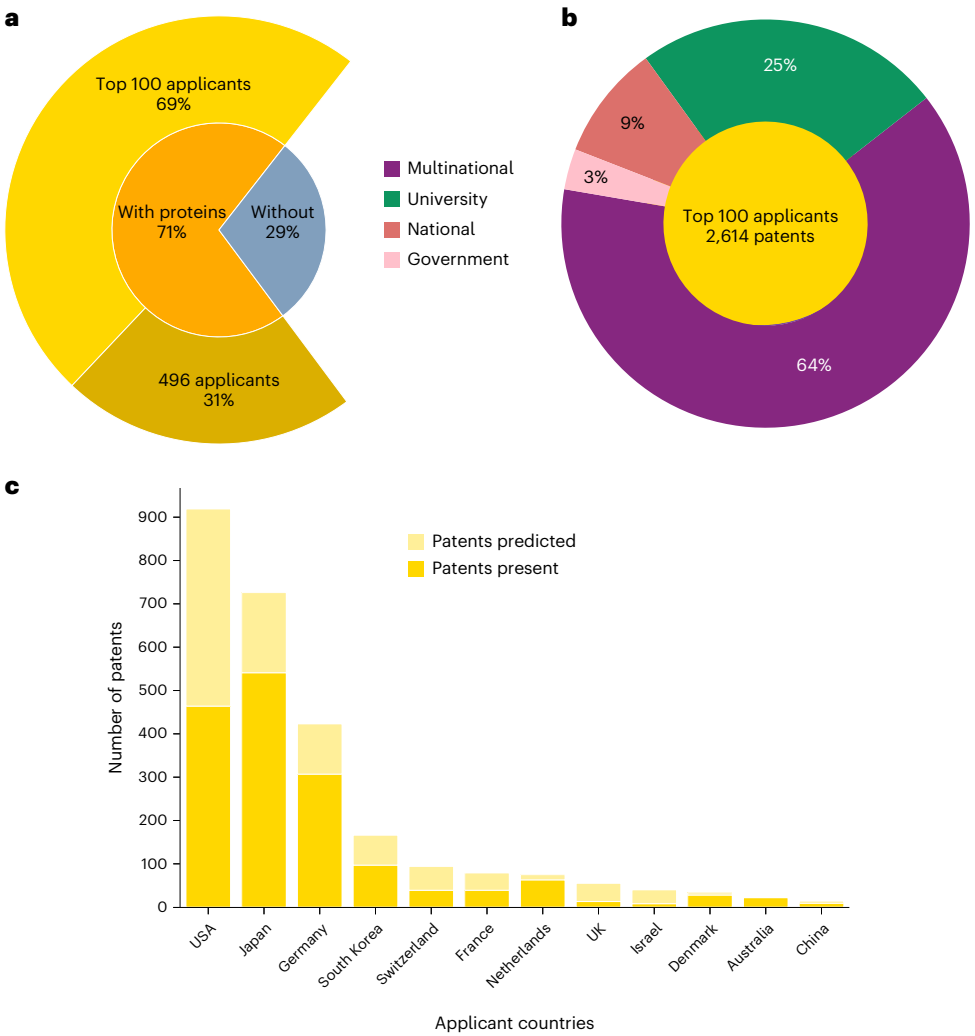

Fig. 3 | Patent applications in global marine bioprospecting. a, Share of companies that submitted patents with at least one protein-coding sequence (bioprospecting patents) and patents with non-coding sequences only. **b**, Top 100 largest patent applicants in marine bioprospecting, aggregated by applicant type. The terms 'multinational' and 'national' denote the extent of company presence in more than two countries or less, respectively. **c**, Top 100 largest patent applicants in marine bioprospecting, aggregated by country of origin (the country of headquarters).

Register of Deep-Sea Species (WoRDSS), 39% of marine species were exclusively found to inhabit deep-sea environments, in contrast to only 15% of all species listed in WoRMS (Fig. 4b). The spatial distribution of ABNJ-specific species (Supplementary Table 4) is predominantly in the sub-Antarctic and Antarctic latitudes (Supplementary Fig. 8).

ABNJ account for 64% of the ocean surface area and 95% of its volume. Once thought to be largely devoid of life, the deep-sea habitats and the water column have been found to harbour many marine species. While many of these species are thought to be considerably cosmopolitan, hotspots of endemism are found throughout the deep sea, perhaps most strikingly around hydrothermal vent systems[32]. According to geolocations of active hydrothermal vents (721 in total), more than half (363) are located in ABNJ.

Discussion

Marine biotechnology is mainly focused on species that serve as model organisms in basic research and as a backbone in genetic engineering, allowing the creation of new drugs and increasing the efficiency of biotechnological processes for food and energy production, plant agriculture or the invention of new materials[33]. Marine species currently represent a small, but important, share that is used as a source for natural product discovery[7,30]. Unravelling the global scope of economic interest in MGR is a crucial first step towards understanding the value that rests in the biological functions encoded in genetic sequences and pathways to fair and equitable sharing of benefits from its use.

Patent data are a valuable source of information in examining innovation and technological advancements, which are widely acknowledged as key drivers of firm performance and economic growth[34,35]. Aggregate patent application counts in particular are useful for studying national patenting activity[36]. Patent data also provide insights into the scope of 'pre-emptive patenting' to block competitors, to increase the market price of existing products or to ensure operational freedom[37]—strategies that biotechnology corporations are known to use[38]. While estimating the market value of patents or establishing links to commercialization is challenging[39], patent data are a useful indicator for gaining insights into the long-term economic interest of societal actors in MGR applications on a global level, in the form of either knowledge production or market control.

The MABPAT database offers a global catalogue of patent sequences derived from marine species over the past three decades. It includes in depth information on patent applications, the genetic sequences attached to them and the marine species from which the sequences were derived, effectively connecting the resources and users of marine bioprospecting. In doing so, the MABPAT database not only

BOX 1

Marine gene patents needing extra protection

Global actors often seek patent protection for their inventions in multiple countries. Filing fees for patent applications extend into the thousands of US dollars per application, and the inventor still has to pay additional fees for each filing. For instance, the average cost of filing in the USA, including attorney's fees, has been estimated at around US$50,000 (ref. 78). Therefore, it is more likely that protection will be sought for highly promising products, methods or associated biotechnological processes. A review of patent filings that include identical nucleotide sequences submitted to all national patent bureaus (13 sequences in total) reveals that the scope of patented commercial biotechnological applications is wide and usually involves transferring specific enzymes to cell metabolic pathways to maximize the production of a specific compound. Examples include applications of biotechnology in medicine (enzymes used for skin care), the food industry (enzymes used in baking and dairy products), agriculture (production of herbicide-tolerant transgenic crops), industrial production (metal nanomaterials used in products such as creams, shampoos, clothing, footwear and plastic containers) and the production of biofuel (isobutanol production and hydrocarbon biosynthesis) (Supplementary Table 2).

One sequence originating from the methanogenic marine archaea species *Methanococcus maripaludis* has been the subject of a series of lawsuits between Butamax (now a subsidiary of International Flavors & Fragrances (IFF)) and Gevo, a transnational biofuels company. The enzyme found in *M. maripaludis* was essential for isobutanol production, and these companies fought over the production of isobutanol instead of ethanol using yeast. After many years of patent disputes, the issue was finally settled through the splitting of the parties' licences in all fields of isobutanol production[79]. In 2021, all Butamax-owned patents were completely acquired by Gevo[80].

fills an important research gap but also contributes to the transparency and interoperability of MGR use. By making it publicly available, we hope to enable further research efforts to inform improved policymaking. The analysis that generated this database also resulted in three key insights that are addressed below.

Rapid technological advances and data governance

Scholars have suggested that the earliest form of a patent system can be traced back 2,500 years ago to ancient Greece and that the first modern patent law dates back to the year 1474[40]. Little surprise then that the patent system has struggled to keep pace with the rapid advances in genetics and genomics research of the past decades, as seen, for instance, in the considerable variation in ground rules for patenting genetic sequences across jurisdictions[27]. Key developments over the past 30 years have focused on jurisdictional norms and compliance standards. In 1998, international applications introduced a mandatory data element for sequence description ('organism'), which aimed to indicate biological origin[41]. Yet, current international standards[42] still allow the inclusion of custom organism names not listed in the Integrated Taxonomic Information System (https://www.itis.gov/), including 'unknown', 'unidentified' and 'artificial sequence'. The new requirements of INSDC[43], announced in November 2021, aim to ensure correct origin disclosure for all incoming sequences. But the effect on the 24.5 million patent sequences already stored in the databases as well as new depositions remains uncertain given that it is ultimately up to patent offices to define standards for the sequences attached to patent applications (https://www.ncbi.nlm.nih.gov/education/patent_and_ip_faqs/).

The analysis of patents therefore often depends on either accepting considerable data gaps or developing methods to reconstruct missing data. In this study, for instance, 17.2 million sequences would have been excluded from the analysis owing to the lack of species names (primarily from the US Patent and Trademark Office, the largest repository of biological sequences and patents). Instead, our sequence similarity model allowed us to reasonably and more comprehensively estimate the patent shares across national states and actor types. This reconstruction allowed us to identify marine origin, focusing on molecular similarities of biological molecules instead of relying on disclosed species names, and to confirm with higher confidence than previous work that Japan, the USA and Germany are the headquarters location for the world's primary MGR patent applicants[25,27]. The disproportionate importance of these three states suggests a corresponding responsibility to work towards innovative benefit-sharing and capacity-building mechanisms. These could include, for instance, the establishment of a multilateral fund for the equitable sharing of benefits between providers and users of digital sequence information (DSI), which has been agreed to be finalized at CBD COP16 (ref. 44).

Importance of microorganisms and deep-sea life for bioprospecting

Marine viruses, although having been recognized as being highly prevalent in ocean ecosystems, contributing to the largest pool of genetic diversity[45], have seen little commercial activity to date beyond a limited focus on those that affect commercial aquaculture production. However, the potential role of viruses in creating proteins of interest for marine bioprospecting could be bigger than we think. Viruses have shaped the majority of the genomes of Archaea and Bacteria via horizontal gene transfer, the exchange of genetic material between organisms that do not form parent–offspring relationships[46]. Bacterial and archaeal species often live in symbiosis and exchange genes with microbial eukaryotes, protists[47], and together constitute the vast majority of organisms used in marine bioprospecting. Importantly, many archaeal and bacterial species used in bioprospecting live in deep-sea habitats, most of which are located in ABNJ. The diversity of microbial marine species is still highly underrepresented in databases that document the distribution and abundance of marine life (Box 2). This underrepresentation may account for the lack of patenting interest in species found exclusively in ABNJ. However, even with limited data, our findings show that ABNJ-specific species are 2.5 times more likely to inhabit the deep ocean compared with marine species in general.

Our analysis of the past three decades of global gene patents indicates that deep-sea species have become an important source for marine bioprospecting. All of the ten largest actors in marine bioprospecting are already using deep-sea species. As a result, there is a logic for benefit sharing from MGR utilization to flow into conservation projects aimed at protecting at-risk deep-sea habitats[48], not least as a vital source for future biotechnology focused on innovation and development of naturally derived products. More advanced biodiversity models that put emphasis on safeguarding entire communities with unique functional roles, including microbial species, should also be better integrated into conservation plans[49].

With the successful conclusion of the High Seas Treaty and the recognition of DSI in the legally binding agreement, MGR used for bioprospecting and product discovery opens a new opportunity to protect biodiversity in deep-sea habitats. However, the INSDC database, the largest data repository of DSI, is currently missing from the biodiversity informatics landscape[50]; therefore, genetic diversity and information

https://doi.org/10.1038/s41893-024-01392-w

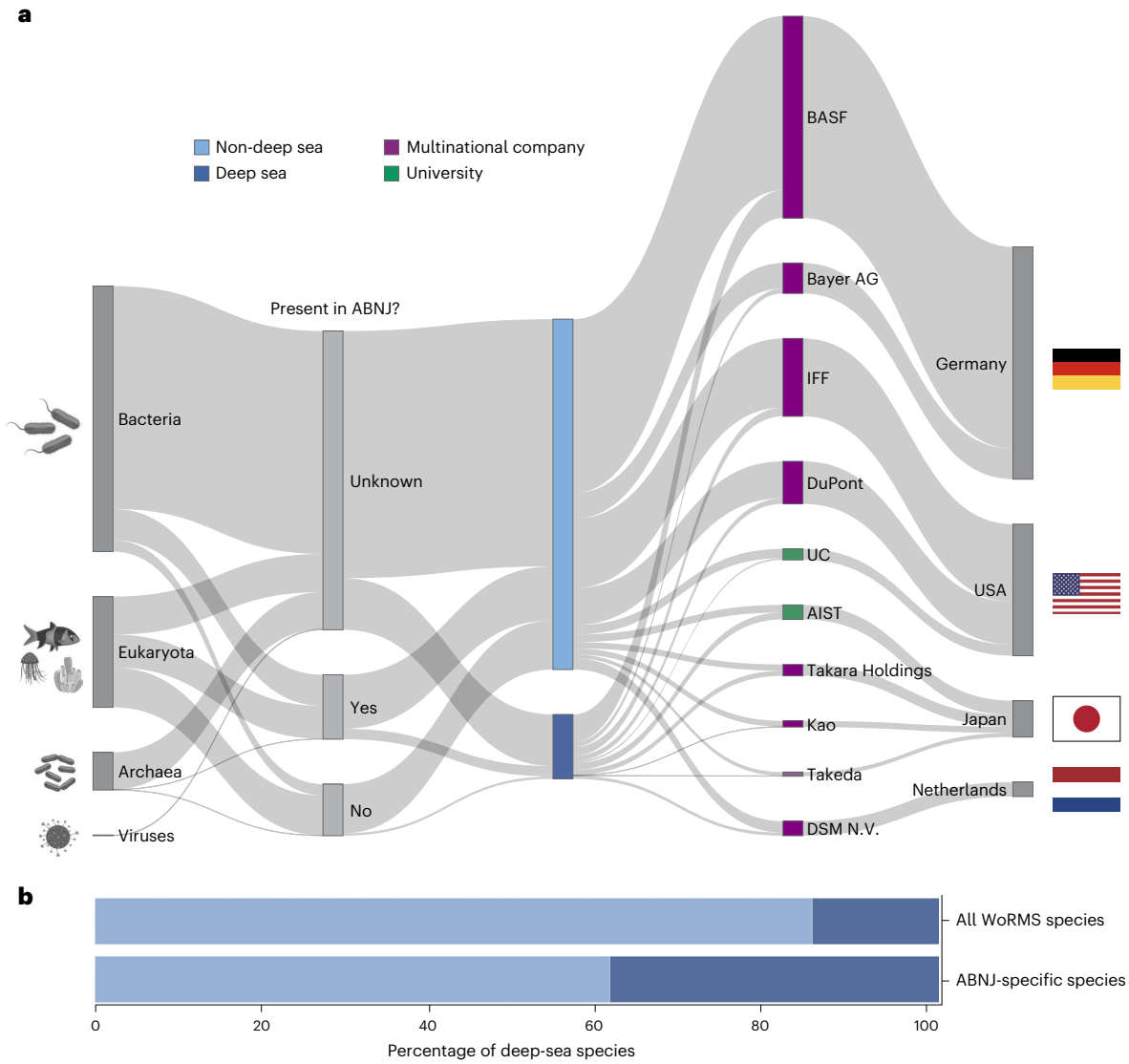

Fig. 4 | Species of interest to marine bioprospecting. a, Species of interest in bioprospecting connected to a company of reference (top 10 largest patent applicants) in patent application grouped by domain and potential presence in ABNJ and deep-sea habitats. AIST, The National Institute of Advanced Industrial Science and Technology; UC, The University of California. The interactive version of this plot is available at https://mabpat.shinyapps.io/main/. **b**, Percentage of deep-sea species present exclusively in ABNJ and all species living in the ocean according to WoRMS. Credit: flags in **a**, flagpedia.net; icons in **a**, FreePik.com.

on the spatial origin of genetic information are not available on a global level. Adoption of the principles of Open and Responsible Data Governance and the development of MGR data repositories[51] will be a necessary step to overcome the lack of information on MGR in ABNJ.

Intellectual property questions are not discussed within the High Seas Treaty, yet commercial sensitivities and national patent regulations are important for benefit sharing related to MGR sourced from the deep sea in ABNJ. While the agreed text of the treaty includes a voluntary mechanism to ensure traceability of MGR collected from ABNJ to end product, the treaty implementation will not affect sequences already used in marine bioprospecting up to date. As there are no legal requirements for patent holders to disclose commercialization of their patents, the scale of commercial products developed and marketed from deep-sea organisms will remain poorly understood. A continued increase in corporate interest along current trajectories would lead to unequal opportunities for new developments in biotechnology.

Multi-stakeholder collaboration in MGR protection

Analysis of bioprospecting patents yielded an asymmetrical distribution of patent registrations, consistent with previous findings[25,27]. The sector is dominated by transnational corporations, which have a higher capacity to undertake genomic research. One-third of all patents were held by the ten largest actors, eight of which are large multinational corporations and none of which conduct marine research themselves but instead rely on public gene databases for sequences with potential commercial applications. While many multinational pharmaceutical companies have marine biology departments[52], their total share of bioprospecting patents is modest (Supplementary Table 3). Still, a fair estimate of corporate engagement in marine species discovery is hard to calculate. Marine scientists who study microbial diversity often engage in collaboration with the oil and gas industry for the collection of samples in deep-sea oil wells[53,54]. With the rising popularity of using remotely operating vehicles for the inspection and maintenance of offshore oil and gas development sites, it is likely that more science–industry partnerships will emerge to support collection of biological data in the deep sea[55].

The disproportionate role of a small number of actors also suggests the potential for science–industry collaboration in the spirit of previous efforts with so-called keystone actors, which consists in engaging the largest companies in a given sector to enable transformative

BOX 2

A vast sea of unknown microbial diversity in the ocean

Microbial species from all domains of life account for more than 95% of total marine biomass and play a pivotal role in the functioning of marine ecosystems, as a foundational level of food webs, climate regulation elements and the backbone of immense genetic diversity[81]. However, as of November 2022, the OBIS database contains only 499 bacterial (https://obis.org/taxon/6) and 7 archaeal species (https://obis.org/taxon/8). The total diversity of microbial marine species, including those uniquely present in ABNJ, is highly underrepresented. Similarly, the WoRDSS list contains almost no bacteria (18 species) or archaea (3 species) as both databases use WoRMS taxonomy as a backbone, and INSDC and National Center for Biotechnology Information (NCBI) taxonomy is not supported as one of the data sources in OBIS.

Thanks to the continuous effort of TARA Oceans, Malaspina, BioGEOTRACES and other projects contributing towards a deeper understanding of microbial diversity, more data have become available to project world map microbial distribution in the global ocean[82]. Some regional diversity studies, including a study in the Clarion–Clipperton Fracture Zone, have already shown that the molecular diversity of deep-sea species is equivalent to levels found in coral reef ecosystems[83]. Ocean datasets are currently missing a global map of microbial functions[84]. As most marine species listed in patent claims are of microbial origin, our understanding of how many of them are uniquely present in ABNJ is far from complete[20,21,31,42–44,57,58,70–76,78–80].

change[56]. Constructive efforts to promote sustainable management in ABNJ have also been undertaken by partnerships such as the Deep Seas Project (https://www.deep-seas.eu) and the Common Oceans ABNJ Project[57], as well as regional bodies such as OSPAR Comission, the North East Atlantic Fisheries Comission and the Sargasso Sea Commission[58], which have addressed challenges related to illegal, unreported and unregulated fishing, and pollution, based on integrated and holistic approaches. The International Seabed Authority, empowered by UNCLOS (Supplementary Text 1) to manage the resources of the seabed in ABNJ, has begun to apply tools such as Regional Environmental Management Plans (REMPs) and designated associated Areas of Particular Ecological Interest (APEIs) aimed at conserving ecosystem function and biodiversity. The impact of such measures could be further amplified by seeking a coordinated approach in accordance with overarching environmental goals[59]. Such initiatives can foster cross-sectoral dialogue and capacity-building activities that improve the capacity of national governments and local communities to engage in sustainable resource use in ABNJ.

Corporate efforts to safeguard intellectual property rights, significant data gaps and the heterogeneity of data standards have contributed to the use of ambiguous terminology and a lack of precision in discussions concerning MGR and bioprospecting in ABNJ. This has shaped perceptions of the scale and nature of commercial interest in MGR from ABNJ, feeding expectations of a lucrative 'deep-sea gold rush' without adequate empirical support for such claims[60,61]. While the conclusion of the High Seas Treaty has laid the foundation for improved management in ABNJ, its entry into force and full implementation are a remote prospect and, in the meantime, voluntary collaborative efforts based on the best available science can help inform future binding mechanisms to ensure conservation and sustainable use. By filling the crucial knowledge gap in understanding the potential of MGR, the MABPAT database represents a first step in that direction.

Methods

Summary statistics of patents that include MGR

The GenBank patent division, the European Bioinformatics Institute database (EMBL-EBI) and the DNA DataBank of Japan (DDBJ) exchange their data daily and together form the INSDC. Genetic sequences associated with patents were retrieved from the Patent division of GenBank from the NCBI (GenBank database) on 10 November 2022; this included 24,600,503 annotated sequences. All files (from gbpat1.seq.gz to gbpat254.seq.gz) were downloaded and processed following the methodology of ref. 25 to create database entries with information on the nucleotide sequence of DNA, species name, patent number, patent data and the party registering the patent. This was done by splitting each file into individual sequences and by extracting the data in the 'origin' field (nucleotide sequence), 'organism' field (species name) and 'journal' field (patent application number, year of application, patent system and patent applicant name) for each sequence. Unlike previous studies[25,27] that restricted their analysis to sequences submitted in a given patent system, here we considered both patents submitted in national jurisdictions and those filed under the Patent Cooperation Treaty ('international' patents) of WIPO.

As of November 2022, sequences from a total of 14,708 different species were included in the GenBank database. To determine the subset of 'marine species' within the database, the taxon match tool of the WoRMS was used for all database entries, resulting in a filtered list of 4,000 species. Web searches were conducted for each of these species to verify the marine origin and to collect further information about the nature of each species. More than half of the matched species were subsequently excluded as non-unique to marine environments, resulting in the list of 1,474 marine species, which was used to select patent records associated with disclosed marine species. See ref. 27 for details of marine origin determination and criteria for filtering.

The taxonomy (domain and phylum) of 879 marine species was retrieved from the WoRMS database. In cases in which such taxonomic levels were not available, we obtained species taxonomy from the NCBI taxonomy database (https://www.ncbi.nlm.nih.gov/taxonomy) and Wikipedia (https://en.wikipedia.org/wiki/) (220 and 356 species, respectively). We did not succeed in matching 19 of the marine species (predominantly marine bacterial strains) into related taxonomic groups owing to lack of certainty in organism names. The complete list of marine species selected for this study is given in Supplementary Table 5a.

MABPAT construction

Marine biotechnology pipelines usually focus on the search for biological compounds that encode a new functionality[62]. There are two types of nucleotide sequences encoded in DNA: protein-coding sequences and non-coding sequences. The latter could have either a functional or a non-functional role in genome regulation, including DNA fragments that code for proteins involved in all cell functions. Except for short peptides like cone snail peptide toxins[63], most natural products are derived from proteins, which are polypeptide chains of a certain length. While identifying the shortest polypeptide chain length to form a protein is still controversial, it is currently estimated at 50 (ref. 64) to 100 (ref. 65) amino acids or 150 to 300 DNA base pairs, respectively.

Another important metric widely used to analyse genome composition variation in molecular biology and genomics is nucleotide usage, which is normally calculated as GC content—the percentage of certain nucleotide bases (guanine and cytosine) that form stronger chemical bonds in DNA strings. Modern genetic engineering techniques such as CRISPR[66] have proven to be very useful at enhancing important functions of proteins by altering DNA makeup. This could involve

changing individual nucleotides or introducing short sequences that control gene regulation and protein synthesis. Hence, GC content for modified proteins with similar functionality remains the same. Short DNA sequences, below the shortest DNA length required for protein formation, have various functions, including in the amplification of a specific gene sequence (as PCR primers), and usually have a wide range of GC content.

To predict whether genetic sequences are protein coding or not, we applied two filtering criteria: sequence length threshold and the presence of an open reading frame (ORF)—a gene region that has the potential to be transcribed into RNA and, after, translated into proteins. Sequences with an ORF longer than 150 base pairs have been considered protein-coding sequences. As most natural products are derived from proteins, we reason that at least one protein-coding sequence has to be included in a patent application, to be related to marine bioprospecting. Following that, we selected 31,914 protein-coding sequences associated with 1,039 marine species together with 112,115 of other sequences that have been submitted as a part of the same application.

For all companies that have registered patents associated with MGR, we counted the total number of nucleotide sequences and calculated the average sequence length (Fig. 2). Based on the shortest protein length estimation, the number of protein-coding or non-coding sequences for each company was identified. In each category, for the ten companies with the highest counts of genetic sequences attached to patent claims, we calculated the length and DNA composition (GC content) of each sequence, and coloured by distinct species origin (Supplementary Fig. 1).

For each sequence that was included in patent applications submitted in national jurisdictions as well as 'international' patents (sequences of special patenting interest), we collected the description of the invention and the protein function, if a nucleotide sequence search (BlastX) resulted in a significant match to a protein with annotated function. Web searches were conducted for each of these proteins to collect further information about protein function and potential application. The resulting information about the sequences of special patenting interest is available in Supplementary Table 2.

Patents owned by subsidiaries were replaced with ultimate owner names of controlled subsidiaries as stated in the Orbis company database, which contains information on around 400 million companies worldwide (Orbis; https://orbis.bvdinfo.com/). For jointly owned patents, the ownership was assigned to the first company on the list. After filtering and removing duplicate entity names and aggregating subsidiaries, we identified a total of 1,125 applicants and collected information about each through web searches, including the country where it is headquartered and the type of entity that it represents. Our classification resulted in five major entity types: multinational (presence in more than two countries) or national companies, universities and their commercialization centres, governmental agencies and 'other' (predominantly applications submitted by private individuals). We also included patent applications from 201 entities that contained protein-coding sequences with identified marine origins, which we were unable to classify under any specific entity type ('none').

Each record in the MABPAT database includes the following: (1) patent applicant name, (2) type of applicant, (3) country where it is headquartered, (4) year of application, (5) patent application number, (6) patent system, (7) genetic sequence identification, (8) marine species name associated with the sequence, (9 and 10) species taxonomy, (11) taxonomic source, (12) whether species can be classified as 'deep-sea' species, (13) source of deep-sea presence, (14) whether species were observed in ABNJ, (15) genetic sequence, (16) GC content, (17) sequence length, (18) whether the sequence originated from a marine organism, (19) whether the marine origin of the sequence was disclosed by the patent applicant or bioinformatically predicted, (20) whether the sequence contains protein-coding information and (21) sequence prediction source. If the marine origin was predicted, the following information about the most similar protein entry in the reference database is provided: (22) protein entry header, (23) protein entry sequence identification, (24) protein entry title, (25) *E*-value, (26) hit identity and (27) query coverage.

Deep-sea presence of marine species

The search for presence of species in deep-sea habitats was conducted based on multiple sources. For species in the Eukarya domain of life, we used the WoRDSS, a taxonomic database of deep-sea species. As Bacteria and Archaea species are not present in WoRDSS, we used web search based on the PubMed (https://pubmed.ncbi.nlm.nih.gov/) and Integrated Microbial Genomes and Microbiomes (https://img.jgi.doe.gov/) databases to establish their potential presence in deep-sea habitats, whether within or beyond national jurisdiction. Samples of species collected from deep-sea environments that have already been found to be associated with international patent applications[27] are also marked as 'deep-sea' species. For the definition of deep-sea marine species, we followed the inclusion criteria in WoRDSS, that is, that the biological material was sampled in depths greater than 500 m.

BlastX sequence similarity model and patent share estimation

Sequence similarity models are widely used to identify newly sequenced data or unknown species[67]. To conduct sequence similarity BlastX searches (translated nucleotide versus protein) against the database of annotated protein sequences, we created the reference database of all proteins belonging to 627 genera of previously confirmed marine species in Supplementary Table 5a. A total of 24,024,531 proteins from all species within those genera were selected from UniProt Knowledgebase (UniProtKB/Swiss-Prot; UniProt Consortium 2023) which included Swiss-Prot (the expertly curated protein records) and TrEMBL (bioinformatically predicted proteins).

BlastX searches with a specific set of search parameters (E-value $\leq 10^{-5}$, query coverage $\geq 80\%$, hit identity $\geq 99\%$) were used to verify that marine sequences could be identified to a genus level with at least 95% confidence (correct hit) (Supplementary Fig. 3a). We also tested whether correct hits and searches with confidence below 95% tend to be included in certain patent applications, patented by certain actors or in certain patent systems, but did not find any preference (Supplementary Fig. 3b,c). Using the sequence search tool DIAMOND[68], we queried 12,716 protein-coding sequences with disclosed marine origin against the selected records from UniProtKB, which resulted in 10,514 correct hits (82.68% recovery rate).

We then queried 7,467,396 sequences with unknown taxonomic origin ('unknown', 'unidentified' and 'synthetic construct' species tag)—62.7% of all GenBank records—against the selected records from UniProtKB, and found 234,836 sequences originating from 1,368 species not previously disclosed in patent records. All matched species were subsequently verified to be exclusively present in marine habitats, resulting in a final list of 561 additional marine species (Supplementary Table 5b). Overall, we have recovered 60,636 previously unknown protein-coding sequences with marine origin and 144,545 other sequences that have been submitted as a part of the same patent application (2,257 patent applications in total).

Finally, we compared summary statistics (number of sequences, number of patents and median year of application) for the top 10 largest patent applicants that referenced sequences with disclosed marine origin and top 10 applicants that referenced sequences with predicted marine origin (Supplementary Figs. 4 and 5, respectively), and found that both lists contained the two largest patent applicants (Bayer and BASF, respectively).

Hydrothermal vent presence and ABNJ-unique species counts

The geolocation of hydrothermal vents was collected from the InterRidge Vents Database. The maritime boundary map of World High Seas was downloaded from Marine Regions (https://marineregions.org/).

Each set of hydrothermal vent coordinates was checked for presence within any of the High Seas polygons. Spatial vector data were analysed with the R package `sf` version 1.0-9 (ref. 69).

To establish the list of species uniquely present in ABNJ, we used species geographical abundance data from OBIS. We first retrieved all 28,375 species with at least one occurrence record in ABNJ (https://obis.org/area/1). For each ABNJ-present species, we checked if it was also observed in the territorial waters of any country. Species with at least one occurrence record were excluded. Data were obtained from the OBIS database (2022) using the R package `robis` version 2.11.0. (ref. 70) and `parallel` version 3.6.2. (ref. 71).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data were collected from publicly available data sources: INSDC (ftp.ncbi.nih.gov/genbank/)[72], UniProtKB[73], WoRMS (https://www.marinespecies.org)[74], WoRDSS[75], PubMed (https://pubmed.ncbi.nlm.nih.gov) and the Integrated Microbial Genomes and Microbiomes database (https://img.jgi.doe.gov). Species observation records were obtained from OBIS (https://obis.org). The geolocations of hydrothermal vents were collected from the InterRidge Vents database (http://vents-data.interridge.org)[76]. The maritime boundaries map of World High Seas was downloaded from Marine Regions (World EEZ v.11) (https://marineregions.org). The resulting MABPAT database is available at https://mabpat.shinyapps.io/main and via figshare at https://doi.org/10.6084/m9.figshare.25289404.v3 (ref. 77).

Code availability

Analysis scripts are available via GitHub at https://github.com/zhivkoplias/mabpat.

References

1. Falkinham, J. O. et al. Proliferation of antibiotic-producing bacteria and concomitant antibiotic production as the basis for the antibiotic activity of Jordan's red soils. *Appl. Environ. Microbiol.* **75**, 2735–2741 (2009).
2. Cragg, G. M. & Pezzuto, M. Natural products as a vital source for the discovery of cancer chemotherapeutic and chemopreventive agents. *Med Princ. Pract.* **25**, 41–59 (2016).
3. Hournan, P. C. H., Hertog, M. G. L. & Katanc, M. B. Analysis and health effects of flavonoids. *Food Chem.* **57**, 43–46 (1996).
4. Achan, J. et al. Quinine, an old anti-malarial drug in a modern world: role in the treatment of malaria. *Malar. J.* **10**, 144 (2011).
5. Shah, Z. et al. Podophyllotoxin: history, recent advances and future prospects. *Biomolecules* **11**, 603 (2021).
6. Atanasov, A. G. et al. Natural products in drug discovery: advances and opportunities. *Nat. Rev. Drug Discov.* **20**, 200–216 (2021).
7. Sigwart, J. D. et al. Unlocking the potential of marine biodiscovery. *Nat. Prod. Rep.* **38**, 1235–1242 (2021).
8. Beraldi-Campesi, H. Early life on land and the first terrestrial ecosystems. *Ecol. Process.* **2**, 1 (2013).
9. Román-Palacios, C., Moraga-López, D. & Wiens, J. J. The origins of global biodiversity on land, sea and freshwater. *Ecol. Lett.* **25**, 1376–1386 (2022).
10. Appeltans, W. et al. The magnitude of global marine species diversity. *Curr. Biol.* **22**, 2189–2202 (2012).
11. Eguíluz, V. M. et al. Scaling of species distribution explains the vast potential marine prokaryote diversity. *Sci. Rep.* **9**, 18710 (2019).
12. Gerwick, W. H. & Moore, B. S. Lessons from the past and charting the future of marine natural products drug discovery and chemical biology. *Chem. Biol.* **19**, 85–98 (2012).
13. Blasiak, R. et al. A forgotten element of the blue economy: marine biomimetics and inspiration from the deep sea. *PNAS Nexus* **1**, pgac196 (2022).
14. Carroll, A. R., Copp, B. R., Davis, R. A., Keyzers, R. A. & Prinsep, M. R. Marine natural products. *Nat. Prod. Rep.* **38**, 362–413 (2021).
15. Blasiak, R. et al. Making marine biotechnology work for people and nature. *Nat. Ecol. Evol.* **7**, 482–485 (2023).
16. Li, B. et al. Preparation of lactose-free pasteurized milk with a recombinant thermostable β-glucosidase from *Pyrococcus furiosus*. *BMC Biotechnol.* **13**, 73 (2013).
17. Aesoy, R. & Herfindal, L. in *Principles of Cancer Treatment and Anticancer Drug Development* (ed. Link, W.) 137–139 (Springer International Publishing, 2022).
18. Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W. & Prasher, D. C. Green fluorescent protein as a marker for gene expression. *Science* **263**, 802–805 (1994).
19. Wynberg, R. & Laird, S. A. Fast science and sluggish policy: the Herculean task of regulating biodiscovery. *Trends Biotechnol.* **36**, 1–3 (2018).
20. *Convention on Biological Diversity* (Secretariat of the CBD, UN Environment Programme, 2011); https://www.cbd.int/convention/text
21. *Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity* (Secretariat of the CBD, UN Environment Programme, 2011); https://www.cbd.int/abs/text
22. UN General Assembly. *Draft Agreement under the United Nations Convention on the Law of the Sea on the Conservation and Sustainable Use of Marine Biological Diversity of Areas Beyond National Jurisdiction*; https://www.un.org/bbnj/sites/www.un.org.bbnj/files/draft_agreement_advanced_unedited_for_posting_v1.pdf (2023).
23. Oldham, P., Hall, S. & Barnes, C. *Patent Landscape Report on Animal Genetic Resources* (World Intellectual Property Organization, 2014); https://www.wipo.int/edocs/pubdocs/en/wipo_pub_947_3.pdf
24. Scholz, A. H. et al. Myth-busting the provider-user relationship for digital sequence information. *Gigascience* **10**, giab085 (2021).
25. Arnaud-Haond, S., Arrieta, J. M. & Duarte, C. M. Marine biodiversity and gene patents. *Science* **331**, 1521–1522 (2011).
26. Arrieta, J. M., Arnaud-Haond, S. & Duarte, C. M. What lies underneath: conserving the oceans' genetic resources. *Proc. Natl Acad. Sci. USA* **107**, 18318–18324 (2010).
27. Blasiak, R., Jouffray, J.-B., Wabnitz, C. C. C., Sundström, E. & Österblom, H. Corporate control and global governance of marine genetic resources. *Sci. Adv.* **4**, eaar5237 (2018).
28. Blasiak, R., Jouffray, J.-B., Wabnitz, C. C. C. & Österblom, H. Scientists should disclose origin in marine gene patents. *Trends Ecol. Evol.* **34**, 392–395 (2019).
29. Katz, L. & Baltz, R. H. Natural product discovery: past, present, and future. *J. Ind. Microbiol. Biotechnol.* **43**, 155–176 (2016).
30. Jaspars, M. et al. The marine biodiscovery pipeline and ocean medicines of tomorrow. *J. Mar. Biol. Assoc.* **96**, 151–158 (2016).
31. *The IUCN Red List of Threatened Species* Version 2024-1 (IUCN, 2022).
32. Van Dover, C. L. et al. Scientific rationale and international obligations for protection of active hydrothermal vent ecosystems from deep-sea mining. *Mar. Policy* **90**, 20–28 (2018).
33. Khan, I., Akmal, K. F., Chong, W. S., Maran, B. A. V. & Shah, M. D. in *Marine Biotechnology: Applications in Food, Drugs and Energy* (eds Shah, M. D. et al.) Ch. 1 (Springer Nature, 2023).
34. Hasan, I. et al. The innovation–economic growth nexus: global evidence. *Res. Policy* **39**, 1264–1276 (2010).
35. Lara-Lopez, A., Valdés, L., de Pinho, R. & Enevoldsen, H. In *Global Ocean Science Report 2020: Charting Capacity for Ocean Sustainability* (ed. Isensee, K.) 135–173 (UNESCO Publishing, 2020).

36. Haščič, I. et al. *Public Interventions and Private Climate Finance Flows: Empirical Evidence from Renewable Energy Financing;* OECD Environment Working Papers no. 80 (2015).
37. Guellec, D., Martinez, C. & Zuniga, M. P. Pre-emptive patenting: securing market exclusion and freedom of operation. *Econ. Innov. New Technol.* **21**, 1–29 (2012).
38. Gurgula, O. Strategic patenting by pharmaceutical companies—should competition law intervene? *IIC Int. Rev. Ind. Prop. Copyr. Law* **51**, 1062–1085 (2020).
39. Hall, B. H., Jaffe, A. & Trajtenberg, M. Market value and patent citations. *RAND J. Econ.* **36**, 16–38 (2005).
40. Adams, J. N. in *Research Handbook on Patent Law and Theory* (ed. Takenaka, T.) Ch. 1, 2–26 (Edward Elgar Publishing, 2019).
41. Jefferson, O. A., Köllhofer, D., Ajjikuttira, P. & Jefferson, R. A. Public disclosure of biological sequences in global patent practice. *World Pat. Inf.* **43**, 12–24 (2015).
42. *Standard ST.26: Recommended Standard for the Presentation of Nucleotide and Amino Acid Sequence Listings Using XML (Extensible Markup Language)* (WIPO, 2023); https://www.wipo.int/export/sites/www/standards/en/pdf/03-26-01.pdf
43. *Spatio-Temporal Annotation Policy* (INSDC, 2021); https://www.insdc.org/news/spatio-temporal-annotation-policy-18-11-2021/
44. *COP15: Nations Adopt Four Goals, 23 Targets for 2030 in Landmark UN Biodiversity Agreement* (Secretariat of the CBD, UN Environment Programme, 2022); https://www.cbd.int/article/cop15-cbd-press-release-final-19dec2022
45. Suttle, C. Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–812 (2007).
46. Sobecky, P. A. & Hazen, T. H. Horizontal gene transfer and mobile genetic elements in marine systems. *Methods Mol. Biol.* **532**, 435–53, (2009).
47. Husnik, F. et al. Bacterial and archaeal symbioses with protists. *Curr. Biol.* **31**, 862–877 (2021).
48. Cordes, E. E. & Levin, L. A. Exploration before exploitation. *Science* **359**, 719 (2018).
49. Pollock, L. J. et al. Protecting biodiversity (in all its complexity): new models and methods. *Trends Ecol. Evol.* **35**, 1119–1128 (2020).
50. Bingham, H. et al. The biodiversity informatics landscape: elements, connections and opportunities. *Res. Ideas Outcomes* **3**, e14059 (2017).
51. Oldham, P., Chiarolla, C. & Thambisetty, S. *Digital Sequence Information in the UN High Seas Treaty: Insights from the Global Biodiversity Framework-related Decisions;* LSE Law School Policy Briefing Series 53/2023. Available at *SSRN* https://doi.org/10.2139/ssrn.4343130 (2023).
52. Trevisanut, S. & Bonfanti, A. *Intellectual Property Rights Beyond National Jurisdiction: Outlining a Regime for Patenting Products Based on Marine Genetic Resources of the Deep-Sea Bed and High Sea.* Available at *SSRN* https://doi.org/10.2139/ssrn.1861020 (2011).
53. Alexander, J. B. et al. Complementary molecular and visual sampling of fish on oil and gas platforms provides superior biodiversity characterisation. *Mar. Environ. Res.* **179**, 105692 (2022).
54. Franco, N. R. et al. Bacterial composition and diversity in deep-sea sediments from the southern Colombian Caribbean Sea. *Diversity* **13**, 10 (2020).
55. McLean, D. L. et al. Enhancing the scientific value of industry remotely operated vehicles (ROVs) in our oceans. *Front. Mar. Sci.* **7**, 00220 (2020).
56. Österblom, H. et al. Scientific mobilization of keystone actors for biosphere stewardship. *Sci. Rep.* **12**, 3802 (2022).
57. *The Common Oceans: ABNJ Deep Seas Project* (FAO, 2018); https://www.fao.org/3/CA2245EN/ca2245en.pdf
58. Wright, G. & Rochette, J. *Regional Ocean Governance of Areas Beyond National Jurisdiction: Lessons Learnt and Ways Forward* (STRONG High Seas Project, 2019); https://www.prog-ocean.org/wp-content/uploads/2019/03/STRONG-HS_Lessons-Learnt-Report.pdf
59. Amon, D. J. et al. Assessment of scientific gaps related to the effective environmental management of deep-seabed mining. *Mar. Policy* **138**, 105006 (2022).
60. Leary, D. & Juniper, S. K. in *The Limits of Maritime Jurisdiction* (eds Schofield, C. et al.) Ch. 34, 769–785 (Martinus Nijhoff Publishers, 2014).
61. Leary, D. Marine genetic resources in areas beyond national jurisdiction: do we need to regulate them in a new agreement? *Marit. Saf. Secur. Law J.* **5**, 22–47 (2018).
62. Rotter, A. et al. The essentials of marine biotechnology. *Front. Mar. Sci.* **8**, 629629 (2021).
63. Terlau, H. & Olivera, B. M. *Conus* venoms: a rich source of novel ion channel-targeted peptides. *Physiol. Rev.* **84**, 41–68 (2004).
64. Woolfson, D. N., Baker, E. G. & Bartlett, G. J. How do miniproteins fold? *Science* **357**, 133–134 (2017).
65. Brunet, M. A., Leblanc, S. & Roucou, X. Reconsidering proteomic diversity with functional investigation of small ORFs and alternative ORFs. *Exp. Cell Res.* **393**, 112057 (2020).
66. Zhang, F., Wen, Y. & Guo, X. CRISPR/Cas9 for genome editing: progress, implications and challenges. *Hum. Mol. Genet.* **23**, 40–46 (2014).
67. Pearson, W. R. An introduction to sequence similarity ("homology") searching. *Curr. Protoc. Bioinformatics* **Chapter 3**, 3.1.1–3.1.8 (2013).
68. Buchfink, B., Xie, C. & Huson, D. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
69. Pebesma, E. Simple features for R: standardized support for spatial vector data. *R J.* **10**, 439 (2018).
70. Provoost, P., Bosch, S. & Best, B. iobis/robis: robis 2.11.0. *Zenodo* https://zenodo.org/doi/10.5281/zenodo.1489948 (2022).
71. parallel: support for Parallel computation in R. R version 3.6.2 https://rdocumentation.org/packages/parallel/versions/3.6.2
72. Arita, M. et al. The international nucleotide sequence database collaboration. *Nucleic Acids Res.* **49**, D121–D124 (2021).
73. The Uniprot Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
74. *World Register of Marine Species* (WoRMS Editorial Board, accessed 15 November 2022); https://doi.org/10.14284/170
75. Glover, A. G., Higgs, N. & Horton, T. *World Register of Deep-Sea Species (WoRDSS)* (accessed 15 November 2022); https://doi.org/10.14284/352
76. Beaulieu, S. E. & Szafranski, K. *InterRidge Global Database of Active Submarine Hydrothermal Vent Fields* Version 3.4 (InterRidge, accessed 1 February 2023).
77. Zhivkoplias, E. MArine Bioprospecting PATent dataset. *figshare* https://doi.org/10.6084/m9.figshare.25289404.v3 (2024).
78. How much does a patent cost? *BlueIron* (16 January 2022); https://blueironip.com/how-much-does-a-patent-cost
79. Butamax, Gevo settle patent dispute. *Biomass Magazine* (24 August 2015); https://biomassmagazine.com/articles/butamax-gevo-settle-patent-dispute-12339
80. Gevo acquires Butamax patent estate. *Yahoo Finance* (23 September 2021); https://finance.yahoo.com/news/gevo-acquires-butamax-patent-estate-130000249.html
81. Abida, H. et al. Bioprospecting marine plankton. *Mar. Drugs* **11**, 4594–4611 (2013).
82. Delmont, T. O. et al. Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genom.* **2**, 100123 (2022).

83. Haward, M. G. & Rogers, A. D. Marine genetic resources in areas beyond national jurisdiction: promoting marine scientific research and enabling equitable benefit sharing. *Front. Mar. Sci.* **8**, 667274 (2021).
84. Tara Ocean Foundation, T. O. & Oceans, T. Priorities for ocean microbiome research. *Nat. Microbiol.* **7**, 937–947 (2022).

Acknowledgements

We thank D. Khvostovetc for providing valuable consultancy in the Patent Cooperation Treaty and European Patent Convention. E.Z., A.P. and R.B. are funded by FORMAS, project number 2020-01048. A.P. is also funded by FORMAS, project number 2019-01220. P.D. is funded by the Research Platform Governance of Digital Practices at the University of Vienna. J.-B.J. is funded by the Knut and Alice Wallenberg Foundation (2021.0343).

Author contributions

E.Z. and A.P. collected the raw data. E.Z., J.-B.J. and R.B. designed the research and analysed the data. E.Z., P.D., J.-B.J. and R.B. wrote the paper.

Funding

Open access funding provided by Stockholm University.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41893-024-01392-w.

Correspondence and requests for materials should be addressed to Erik Zhivkoplias.

Peer review information *Nature Sustainability* thanks Peter McGarvey and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Erik Zhivkoplias

Last updated by author(s): May 16, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	Custom code used during data analysis is available through GitHub at https://github.com/zhivkoplias/mabpat. Statistical analysis was performed using pearsonr function in SciPy 1.24.4. Sequence similarity search was performed used blastx 2.9.0 and diamond v0.9.35.136. Spatial vector data were analysed with the R packages sf 1.0-9, robis 2.11.0., and parallel 3.6.2.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data was collected from publicly available data sources:
International Nucleotide Sequence Database Collaboration (INSDC) - Arita M, Karsch-Mizrachi I, Cochrane G. The international nucleotide sequence database

collaboration. Nucleic Acids Res. 2021 Jan 8;49(D1):D121-D124. https://doi:10.1093/nar/gkaa967 Accessed 2022-11-15.
UniProtKB - The UniProt Consortium et al. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research, 51, 523–531 (2023). https://doi.org/10.1093/nar/gkac1052 Accessed 2023-10-25.
World Register of Marine Species (WoRMS)- WoRMS Editorial Board (2023). World Register of Marine Species. Available from https://www.marinespecies.org at VLIZ. https://doi:10.14284/170 Accessed 2022-11-15.
World Register of Deep-Sea species (WoRDSS) - Glover, A.G., Higgs, N., and Horton, T. (2023). World Register of Deep-Sea species (WoRDSS). https://doi:10.14284/352 Accessed 2022-11-15.
Species observation records were obtained from Ocean Biodiversity Information System (OBIS).
The geolocation of hydrothermal vents was collected from the InterRidge Vents Database - Beaulieu, S.E., Szafranski, K. (2020) InterRidge Global Database of Active Submarine Hydrothermal Vent Fields, Version 3.4. World Wide Web electronic publication available from http://vents-data.interridge.org Accessed 2023-02-01.
The maritime boundaries map of World High Seas was downloaded from Marine Regions (https://marineregions.org/).

The resulting MABPAT database is available at https://mabpat.shinyapps.io/main/ as well as in Figshare repository at https://doi.org/10.6084/m9.figshare.25289404.v3.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	The study does not involve human participants
Reporting on race, ethnicity, or other socially relevant groupings	The study does not involve human participants
Population characteristics	The study does not involve human participants
Recruitment	The study does not involve human participants
Ethics oversight	The study does not involve human participants

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Our study collected all publicly available genetic sequences referenced in patent applications that include at least one sequence of marine origin. Total counts were calculated for different species types (marine and non-marine), sequences (protein-coding and non-protein coding), types of patent applicants (MULTINATIONAL, NATIONAL, UNIVERSITY, OTHER), and species ecological presence (deep-sea and non deep-sea). The average sequence length (mean) was calculated for all applicants that submitted at least 25 nucleotide sequences in their patent claims. We tested whether the total number of patents registered by each applicant was correlated with the total count of unique species included in such patents, and found a positive correlation: r = 0.8168, p-value = 2.17552e-318. The sequences with unknown origin identical to sequences derived from species with marine origin were identified by our sequence similarity model based on statistical significance: E-value: less or equal to 10^-5, query coverage: more or equal to 80%, hit identity: more or equal to 99%.
Data exclusions	No data were excluded from the analysis.
Replication	No replication was used for this study. All data collected were quantifiable and replication would not change any bias in data analysis.
Randomization	No randomization was used for this study. All data collected were quantifiable and randomization would not change any bias in data analysis.
Blinding	No blinding was used for this study. All data collected were quantifiable and blinding would not change any bias in data analysis.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*
Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*
Research sample	*Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.*
Sampling strategy	*Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data collection	*Describe the data collection procedure, including who recorded the data and how.*
Timing and spatial scale	*Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Reproducibility	*Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.*
Randomization	*Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.*
Blinding	*Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Did the study involve field work? ☐ Yes ☒ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*
Access & import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*
Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	*Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	*State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.*
Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*

Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	*For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Reporting on sex	*Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☒	☐	Public health
☒	☐	National security
☒	☐	Crops and/or livestock
☒	☐	Ecosystems
☒	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☒	☐	Demonstrate how to render a vaccine ineffective
☒	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	☐	Increase transmissibility of a pathogen
☒	☐	Alter the host range of a pathogen
☒	☐	Enable evasion of diagnostic/detection modalities
☒	☐	Enable the weaponization of a biological agent or toxin
☒	☐	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session
(e.g. UCSC)

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)	*Specify: functional, structural, diffusion, perfusion.*
Field strength	*Specify in Tesla*
Sequence & imaging parameters	*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
Area of acquisition	*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software	*Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
Normalization	*If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*
Normalization template	*Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*
Noise and artifact removal	*Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring | *Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings | *Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*

Effect(s) tested | *Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016) | *Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*

Correction | *Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☒	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity | *Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).*

Graph analysis | *Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*

Multivariate modeling and predictive analysis | *Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*