NAR Genomics and Bioinformatics, 2024, **6**, lqae075
https://doi.org/10.1093/nargab/lqae075
Advance access publication date: 17 July 2024
Standard Article

The Catalan initiative for the Earth BioGenome Project: contributing local data to global biodiversity genomics

Montserrat Corominas[1,2,3,*,†], Tomàs Marquès-Bonet[4,5,6,7,†], Miquel A. Arnedo[8,9], Mònica Bayés[6,10], Jordina Belmonte[11,12], Hector Escrivà[13], Rosa Fernández[4], Toni Gabaldón[5,14,15,16], Teresa Garnatje[17,18], Josep Germain[19], Manel Niell[20], Ferran Palero[21], Joan Pons[22], Pere Puigdomènech[3,23], The Catalan Initiative for the Earth BioGenome Project, Vanesa Arroyo[20], Cristian Cuevas-Caballé[1,9], Joan Ferrer Obiol[24], Ivo Gut[6,10], Marta Gut[6,10], Oriane Hidalgo[16,25], Guillem Izquierdo-Aranega[1,9], Laia Pérez-Sorribes[17,26], Emilio Righi[27], Marta Riutort[1,9], Joan Vallès[3,9,28], Julio Rozas[1,9], Tyler Alioto[6,10] and Roderic Guigó[3,27,29,*]

[1]Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona (UB), 08028 Barcelona, Catalonia, Spain
[2]Institut de Biomedicina (IBUB), Universitat de Barcelona (UB), 08028 Barcelona, Catalonia, Spain
[3]Institut d'Estudis Catalans (IEC), 08001 Barcelona, Catalonia, Spain
[4]Institute of Evolutionary Biology (IBE, UPF-CSIC), PRBB, 08003 Barcelona, Spain
[5]Catalan Institution of Research and Advanced Studies (ICREA), 08010 Barcelona, Spain
[6]Centre Nacional d'Anàlisi Genòmica (CNAG), 08028 Barcelona, Spain
[7]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain
[8]Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals, Facultat de Biologia, Universitat de Barcelona (UB), 08028 Barcelona, Catalonia, Spain
[9]Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona (UB), 08028 Barcelona, Catalonia, Spain
[10]Universitat de Barcelona (UB), 08028 Barcelona, Spain
[11]Departament de Biologia Animal, Biologia Vegetal i Ecologia, Facultat de Biociències, Universitat Autònoma de Barcelona (UAB), 08193 Bellaterra, Catalonia, Spain
[12]Institut de Ciència i Tecnologia Ambientals (ICTA-UAB), Universitat Autònoma de Barcelona (UAB), 08193 Bellaterra, Catalonia, Spain
[13]Sorbonne Université, CNRS, Biologie Intégrative des Organismes Marins, BIOM, F-66650, Banyuls-sur-Mer, France
[14]Barcelona Supercomputing Centre (BSC-CNS), 08034 Barcelona, Spain
[15]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08028 Barcelona, Spain
[16]CIBER de Enfermedades Infecciosas, Instituto de Salud Carlos III, Madrid, Spain
[17]Institut Botànic de Barcelona (IBB), CSIC-CMCNB, 08038 Barcelona, Catalonia, Spain
[18]Jardí Botànic Marimurtra - Fundació Carl Faust, 17300 Blanes, Catalonia, Spain
[19]Institució Catalana d'Història Natural, 08001 Barcelona, Catalonia, Spain
[20]Andorra Recerca + Innovació (ARI), AD600 Sant Julià de Lòria, Andorra
[21]Institut Cavanilles de Biodiversitat i Biologia Evolutiva (ICBIBE), Paterna, Valencia, Spain
[22]Departament de Biodiversitat Animal i Microbiana, Institut Mediterrani d'Estudis Avançats (CSIC-UIB), 07190 Esporles, Illes Balears, Spain
[23]Centre de Recerca en Agrigenòmica, CSIC/IRTA/UAB/UB, 08193 Bellaterra, Catalonia, Spain
[24]Department of Environmental Science and Policy, University of Milan, Milan, Italy
[25]Royal Botanic Gardens, Kew, TW9 3DS Richmond, UK
[26]Estación Biológica de Doñana, CSIC, 41092 Sevilla, Spain
[27]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, 08003 Barcelona, Catalonia, Spain
[28]Laboratori de Botànica (UB), Unitat Associada al CSIC, Facultat de Farmàcia i Ciències de l'Alimentació, Universitat de Barcelona, 08028 Barcelona, Catalonia, Spain
[29]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Catalonia, Spain

[*]To whom correspondence should be addressed. Tel: +34934037003; Email: mcorominas@ub.edu
Correspondence may also be addressed to Roderic Guigó. Tel: +34933160110; Email: roderic.guigo@crg.cat
[†]The first two authors should be regarded as Joint First Authors.

Abstract

The Catalan Initiative for the Earth BioGenome Project (CBP) is an EBP-affiliated project network aimed at sequencing the genome of the >40 000 eukaryotic species estimated to live in the Catalan-speaking territories (Catalan Linguistic Area, CLA). These territories represent a biodiversity hotspot. While covering less than 1% of Europe, they are home to about one fourth of all known European eukaryotic species. These include

Received: December 29, 2023. **Revised:** May 10, 2024. **Editorial Decision:** June 17, 2024. **Accepted:** June 19, 2024

a high proportion of endemisms, many of which are threatened. This trend is likely to get worse as the effects of global change are expected to be particularly severe across the Mediterranean Basin, particularly in freshwater ecosystems and mountain areas. Following the EBP model, the CBP is a networked organization that has been able to engage many scientific and non-scientific partners. In the pilot phase, the genomes of 52 species are being sequenced. As a case study in biodiversity conservation, we highlight the genome of the Balearic shearwater *Puffinus mauretanicus*, sequenced under the CBP umbrella.

Introduction

The Earth BioGenome Project (EBP, https://www.earthbiogenome.org) is a global network of networks that aims to sequence the genomes of the approximately 1.8 million eukaryotic species known to live on Earth (1–3). Within the EBP, the Catalan Initiative for the Earth BioGenome Project (CBP, https://www.biogenoma.cat/) is an affiliated project network whose main objective is to sequence the genomes of the >40 000 eukaryotic species estimated to live in the Catalan-speaking territories. These territories, spanning across Spain, France, the Principality of Andorra and the Sardinian city of Alguer, have historically shared a strong cultural tradition, primarily reflected in the use of the Catalan language. They lay at the intersection of European and African plates (4), and at the cross-road between the Euro-Siberian and the Mediterranean biogeographical regions. These territories, while covering less than 1% of the European territory (70 520 km^2, 2 500 km of seashore), harbor about one fourth of all Europe's known eukaryotic species. They are also characterized by an elevated number of endemisms. Many endemic species are threatened, a trend that will get worse in the future, as the impact of climate change will likely be intense in the Mediterranean Basin, particularly in mountain areas (5–8).

Here, we describe the CBP, including the project's rationale, its origins and history, its current structure and status, and future direction. Due to its global nature, the success of the EBP is highly dependent on the engagement of scientific communities worldwide and, given its implications beyond science, it represents a unique opportunity to engage society as a whole to promote awareness about the relevance of biodiversity to regional and global health. The CBP has been able to involve many scientific, academic and societal stakeholders, and, in this regard, is an example of bottom-up regional and transnational initiatives in the field of biodiversity genomics. We also report some features of the genome of the Balearic shearwater (*Puffinus mauretanicus*) –one of the first genomes sequenced by the CBP— as an example of the relevance of genomics in biodiversity conservation.

A Catalan version of the article can be found at: https://doi.org/10.5281/zenodo.12722864.

Materials and methods

To explore the patterns of population structure in Mediterranean shearwaters, we used whole-genome resequencing data of 36 individuals of Balearic (n = 30) and Yelkouan shearwaters (n = 6) (Izquierdo-Aranega, Cuevas-Caballé, *et al.* in preparation). Briefly, reads were mapped to the *P. mauretanicus* reference genome (9) using BWA mem v0.7.17 (10) with default parameters and variant calling was performed by combining HaplotypeCaller in GATK4 v4.1.9 (11) and Freebayes v1.2 (12). We further filtered the resulting SNP dataset with VCFtools v.0.1.15 (13) using quality (–minQ 30), coverage (–minDP 5 and –maxDP 200) and missingness (–max-missing 0.75) thresholds. We used the resulting VCF file to perform a Principal Component Analysis (PCA) implemented in PLINK v1.90 (14).

To build a SNP panel for the assignment of shearwaters to their source colony, we selected the most differentiated SNPs from pairwise F_{ST} comparisons across each colony pair using VCFtools. We selected a final set of 61 SNPs from the top differentiated SNPs within each pairwise comparison, after making sure to only include each SNP once. To validate our SNP panel, we performed a PCA implemented in PLINK and we assessed genomic assignment of each individual to their sampled colonies of origin using Assignpop v1.2.4 (15) with K-fold cross-validation using the LDA model.

Results

The Catalan Initiative for the Earth BioGenome project

The CBP is a networked organization launched in 2019 by the Catalan Society of Biology (SCB, https://scb.iec.cat), and the Catalan Institution of Natural History (ICHN, https://blogs.iec.cat/ichn), two societies affiliated with the Institute of Catalan Studies (IEC, https://www.iec.cat). IEC is the academy tasked with the promotion of science and culture in the Catalan territories. A white paper describing the background, the aims and a proposed outline of the CBP was released by the SCB in October 2018 (https://www.biogenoma.cat/wp-content/uploads/2023/04/2018CatalanBioGenomeProjectNov12b.pdf). Seed funding to initiate the CBP's activities came from the IEC and from the legacy of Leandre Cervera, a former president of the SCB (1935–1963) during the period when the Society activities were banned (16). Core funding for the first phase (2023–2025) is provided by the Department of Research and Universities from the Generalitat de Catalunya (Government of Catalonia) to the IEC. The CBP has also the support of the Andorran Research + Innovation (ARI, https://www.ari.ad/en).

The CBP is an affiliated member of the EBP and a regional partner project of the European Reference Genome Atlas (ERGA, https://www.erga-biodiversity.eu/). ERGA is the pan-European partner of the EBP and aims to coordinate the production of high-quality genomes representative of eukaryotic biodiversity in Europe (17,18). Organically, within the International Nucleotide Sequence Database Collaboration (INSDC), the CBP BioProject (accession number PRJEB49670) is, therefore, a direct child of the BioProjects EBP (PRJNA533106) and ERGA (PRJEB43510) (Figure 1).

The pilot phase of the CBP was launched in the summer of 2020, with an open call for sequencing projects, followed by a second open call in the summer of 2021. During this phase, the ICHN worked in parallel to create a digitized catalog of the eukaryotic species living in Catalonia, building on early catalogs (19). As of November 2023, this catalog contains

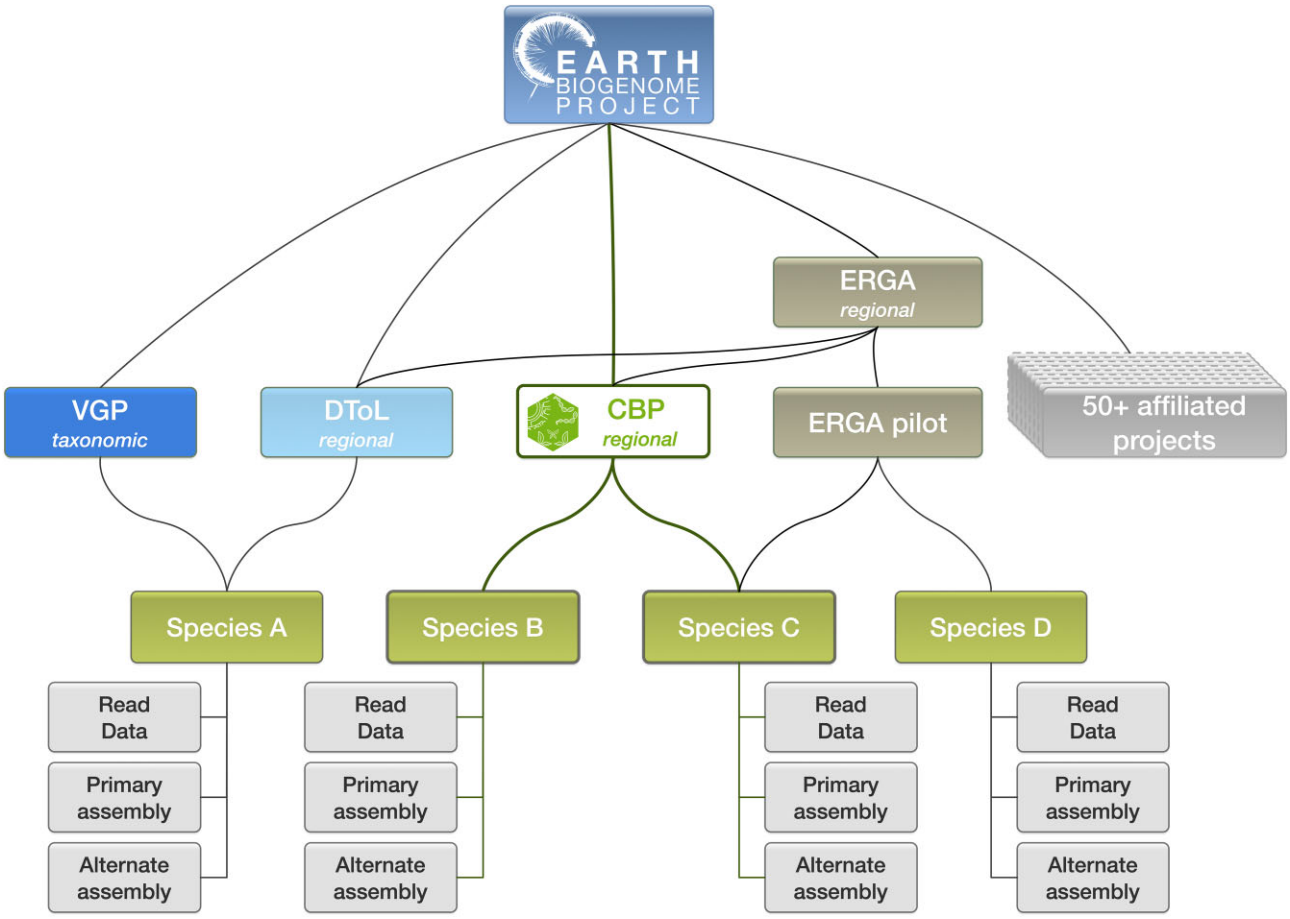

Figure 1. Data and genome assemblies submitted to the International Nucleotide Sequence Database Collaboration (INSDC) are grouped in umbrella projects at the species level, which in turn can be linked to one or more EBP-affiliated projects. Regional projects such as CBP and Darwin Tree of Life (DToL) (3) may themselves be grouped under larger regional initiatives, such as the European Reference Genome Atlas (ERGA).

38 429 species, but will likely contain >40 000 species when completed (Figure 2). At any rate, this catalog will serve as a reference for prioritizing genomes to be sequenced during the different phases of the CBP. Prioritization criteria include phylogenetic position and novelty, interest of local research groups, degree of endemism, conservation status and biomedical, agricultural or industrial interests, and adherence to the ERGA principles (18,20).

The overarching aim of the pilot phase has been, however, to seamlessly connect all the existing components to create the infrastructure required for streamlined production. Broadly following the model of the Darwin Tree of Life (2) and of ERGA (20), (Figure 3), samples from the species prioritized in each phase are collected, identified, annotated, barcoded and biobanked. The biobanking structure of the CBP is still being defined, but the main sites include the Museum of Natural Sciences of Barcelona, the biobank of the Barcelona Zoo, the Balearic Biodiversity Centre, and the Botanical Institute of Barcelona and the Center for Research in Agricultural Genomics for plant samples. The samples are then subjected to DNA and RNA sequencing. The primary sequencing center for the CBP is the National Center for Genomic Analysis (CNAG). At the CNAG, sequencing is generally performed using both short and long read technologies, as well as Hi-C for chromosome-level scaffolding, with the aim of meeting EBP standards (21), including a minimum contig N50 of 1Mb, chromosome-level scaffold N50, and consensus quality greater than QV40 when possible. After data quality control, genomes are assembled (https://github.com/cnag-aat/assembly_pipeline), manually curated (https://gitlab.com/wtsi-grit/rapid-curation) and then annotated. The CBP produces its own annotation, using the CNAG pipeline (https://github.com/cnag-aat/), which is distributed through the CPB portal (see below). This should be considered complementary to the annotation produced within the ERGA consortium, which can also be distributed through the portal. Raw data and assemblies are openly distributed to the designated public repositories.

Additional pipelines for extended gene annotation (22), downstream functional analysis (23) and phylogenomic analysis (24) already exist and are being reimplemented as scalable and reproducible workflows using NextFlow (25) and SnakeMake (26).

There is also significant effort within CBP toward outreach and public engagement. During the pilot phase, the project was presented to a number of international meetings and workshops. Numerous talks and seminars were held for the general public, including two DNA extraction workshops and an exhibition aimed at stimulating public interest in genomics (see www.biogenoma.cat/en/outcomes/). Street banners with

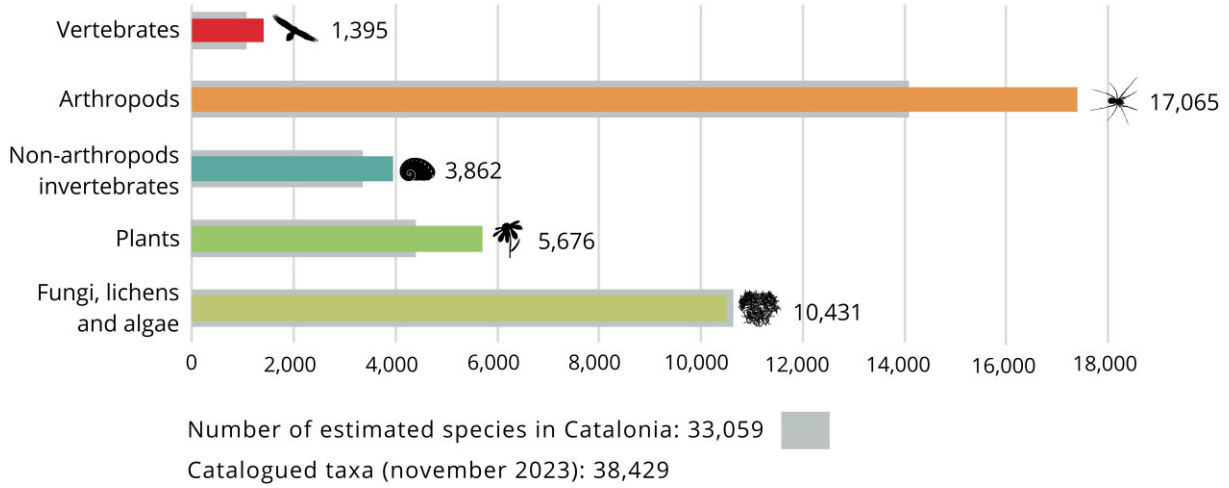

Figure 2. Catalonia's biodiversity catalog. Number of initially estimated (gray) and currently cataloged taxa from different taxonomic groups. The numbers refer exclusively to Catalonia, and do not include the rest of Catalan-speaking territories.

information about the activities of the Initiative have been distributed in Barcelona and other cities (Supplementary Figure S1). Informative leaflets can also be downloaded from the web portal (www.biogenoma.cat).

Current status

As of November 2023, 52 genomes representing species from a variety of taxa, habitats, and ecosystems are underway under the CBP umbrella (https://goat.genomehubs.org/projects/CBP), five of which have already been published (9,27–30) (Figure 4). The CBP has followed a bottom-up approach, being responsive to the interests of the local research community. Target species at the initial phase include those belonging to underexplored taxa, such as the freshwater flagellate *Singekia montserratensis*, an early divergent microbial eukaryote (31); Rare, endemic or difficult to locate species such as the Catalan blind scorpion (*Belisarius xambeui*), the monotypic Sensonatoridae groundwater amphipod (*Sensonator valentiensis*), the Catalan giant earthworm (*Norana najaformis*), the kangaroo shrimp (*Dugastella valentina*), the mustache shrimp (*Derocheilocaris remanei*), and the Balearic shearwater (*Puffinus mauretanicus*) (see below); Emergent model species such as the wall lizard (*Podarcis muralis*), and the sharp-snout sea bream (*Diplodus puntazzo*), are also targeted in this pilot phase. CBP's aim is also to provide genomes for endangered species, such as the Montseny brook newt (*Calotriton arnoldi*), the most threatened amphibian in Europe, or the iconic red coral (*Corallium rubrum*); Traditionally used medicinal species, such as the Pyrenean chamomile (*Achillea ptarmica subsp. pyrenaica*); and species of economic interest, such as the pearly razorfish (*Xyrichtys novacula*) or the tiger nut (*Cyperus esculentus*), among others.

A priority within the CBP is the development of bioinformatic methods to help address the data challenges of biodiversity genomics projects and the commitment to the rapid release of data. Thus, efforts have specifically focused on developing the BioGenome Portal (32) (https://ebp.biogenoma.cat/), a customizable, web-based tool designed to improve accessibility to data generated by the CBP or by another biodiversity genomics project.

The balearic shearwater genome: a case study in biodiversity conservation

The genome of the Balearic shearwater (*Puffinus mauretanicus*) is one of the first genomes published under the CBP umbrella (9). The genomic sequence of the Balearic shearwater and its related data can be accessed through the ENA (European Nucleotide Archive at https://www.ebi.ac.uk/ena/browser/view/PRJNA780920) and the CBP portal at (https://dades.biogenoma.cat/organisms/48681). This medium-sized pelagic seabird, endemic to the Balearic Islands, is listed as critically endangered by the IUCN and its populations are declining at an alarming annual rate of 7.4–14%. The main threats to the species' survival come from human activities, especially accidental bycatch in longline fisheries and predation in the breeding colonies by invasive alien species (33,34). Reducing bycatch rates of fisheries, ensuring predator-free breeding habitat and protecting foraging habitat are urgent conservation measures for avoiding the extinction of the species (35–37). The availability of a reference genome and population genomics data for the Balearic shearwater will enhance the proposal of science-based conservation policies that contribute to the implementation of the aforementioned measures.

Although limited, information from a single genome can provide knowledge of a species' evolutionary history, including the inference of potential events that have shaped its current genetic diversity. By using coalescent-based methods to analyze the genomic diversity, it is possible to determine the changes in effective population size throughout history. Specifically, it could be inferred that the Balearic shearwater suffered a sharp population decline during the last interglacial period (~120 Kya) (9). This severe population decline would be expected to reduce the species' genetic diversity, and ultimately increase its extinction risk through genetic mechanisms, including the accumulation of deleterious mutations, the loss of adaptive potential, and inbreeding depression (38,39). However, current genome-wide heterozygosity levels in the Balearic shearwater are roughly equivalent to those observed in other seabirds with larger population sizes, suggesting that: (i) this sharp demographic decline has not severely affected current levels of heterozygosity, (ii) that such levels were very high prior to the decline, or (iii) that admixture in-between glacial periods could have maintained high levels of heterozygosity

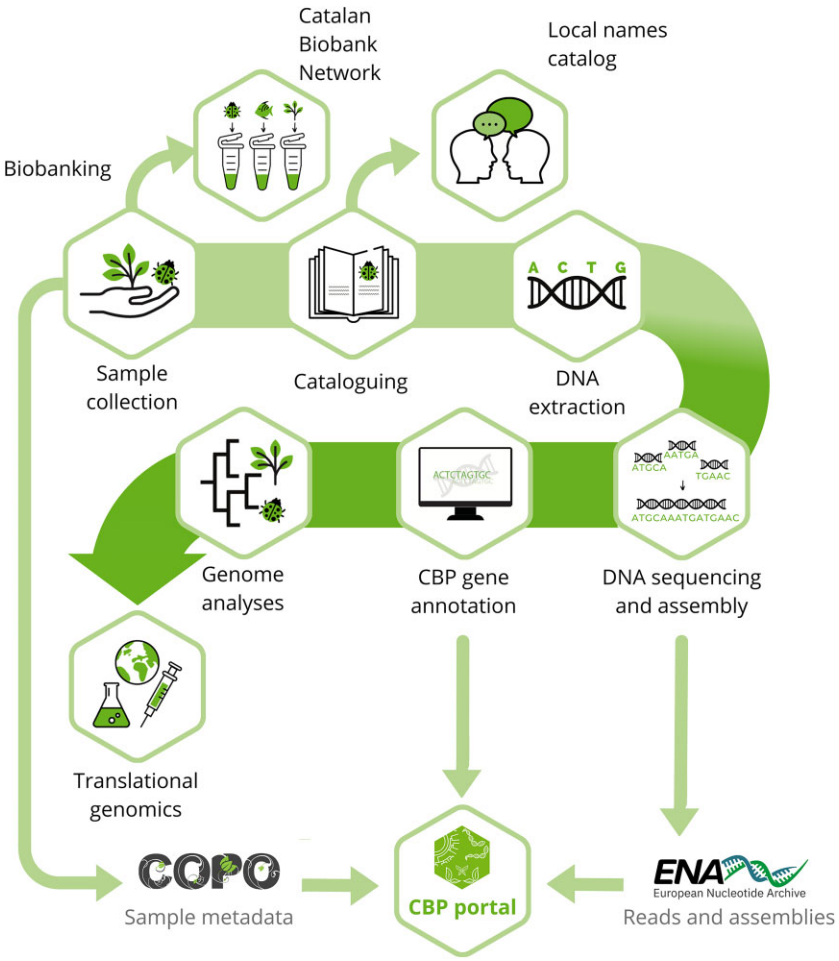

Figure 3. The CBP workflow (see main text for explanation).

(9). Nevertheless, the pressure that human activity is already placing on these specie's populations could eventually lead to a severe loss of diversity.

Reference genomes are also required to refine population genomic analyses, which allow obtaining insights into some critical aspects relevant to conservation. Under this context, whole-genome resequencing analysis of different populations of this species, together with that of the closely related Yelkouan shearwater (*P. yelkouan*) (Izquierdo-Arànega, Cuevas-Caballé, et al. in preparation) (Figure 5A), shows a fine characterization of the population structure across colonies, enabling genetic assignment of individuals to their natal colony (Figure 5B). As mentioned before, bycatch in longline fisheries is a main threat for the survival of the Balearic shearwater and one that is likely to drive the species to extinction unless bycatch rates are significantly reduced (33). Assigning Balearic shearwaters accidentally caught in fishing gear as bycatch back to their colony of origin would establish a connection between threats encountered at sea and specific colonies, aiding in management strategies to mitigate bycatch from affected colonies. With this purpose, an SNP panel comprising a smaller set of highly distinct SNPs is under development (Figure 5c). This panel will facilitate the identification of shearwaters caught as bycatch and trace them back to their original colony, thereby aiding in pinpointing potential bycatch hotspots. Furthermore, this tool will allow monitoring the potential decline in heterozygosity over time, which could compromise the species' evolutionary capacity.

Overall, to ensure the persistence of rare and endangered species, we need more science-based conservation actions. Projects like the EBP, and its affiliates, such as the CBP, are producing reference genomes at an unprecedented pace. These will enhance the use of genetic resources aimed at providing conservation-relevant information, and can be instrumental to implement effective conservation policies.

Linguistic, cultural and biological diversity

While the reasons are not well understood, there is strong evidence of geographical co-localization between biological and linguistic diversity (40,41). Unsurprisingly, threats to both living species and languages tend to occur in the same geographical regions. They could be both triggered by the same socioeconomic and political processes threatening the integrity and survival of local cultures and the environments in which they live (40). Extinction of languages is even more dramatic than the extinction of biodiversity, with the disappearance of 50–90% of the world's languages predicted by the end of the century (42). This will have a tremendous impact in our ability to translate knowledge gained from biodiversity genomics projects into applications in human health and in other areas. It has been demonstrated that the extinction of a language

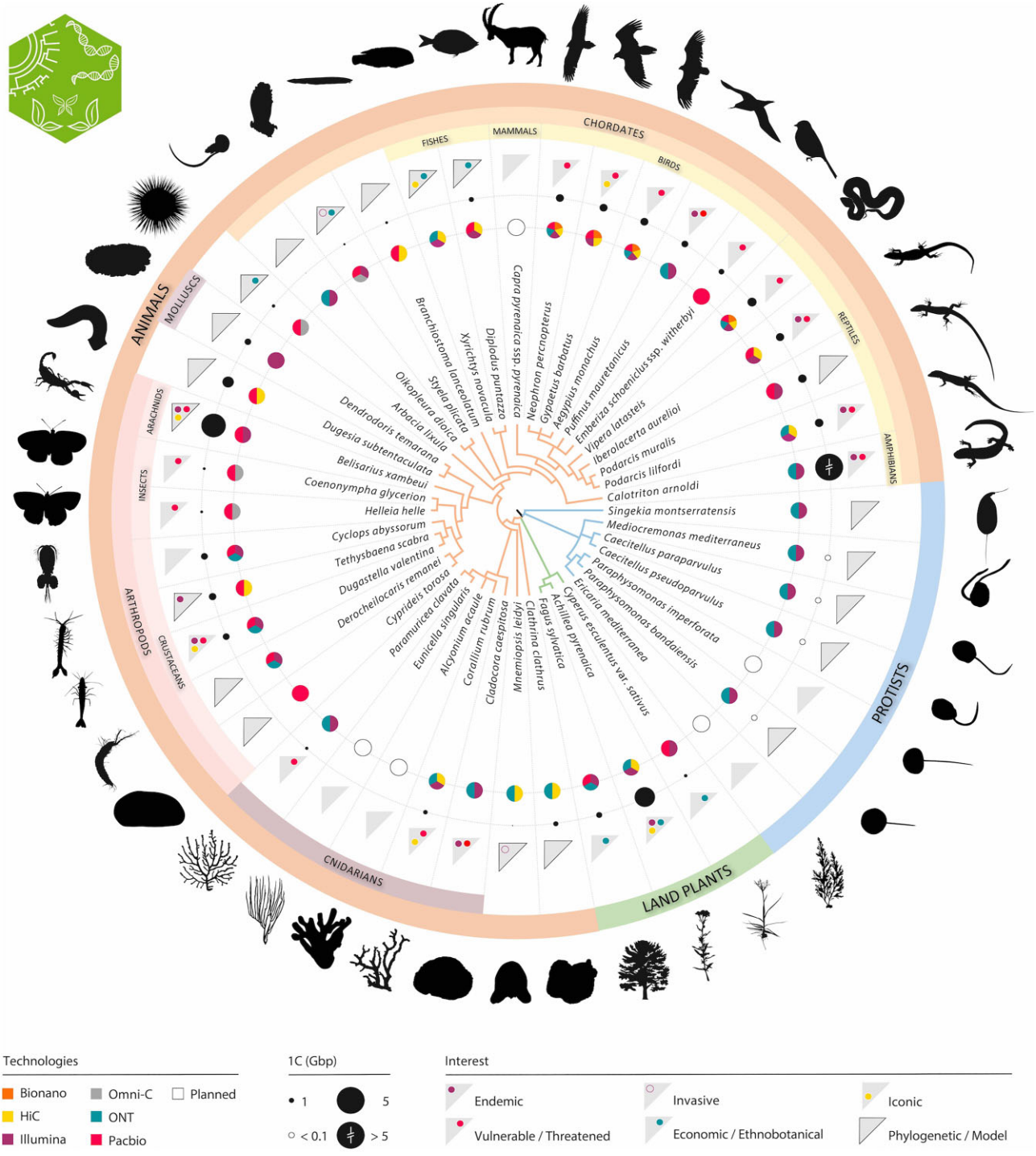

Figure 4. Genome projects under the umbrella of the Catalan Initiative for the Earth BioGenome Project (CBP) by June 2023. The figure captures the phylogenetic position of the species, the genome size when known, the technologies employed for sequencing, and rationale behind the species selection. Top left is the CBP logo.

contributes to the loss of traditional medical knowledge because the majority of medicinal plant uses are limited to a single language (43).

Beyond practical considerations, the common names of animals and plants in a given language often reflect the relationship of humans to the biodiversity inherent in the broader culture in which the language emerged. Naming objects that people know is one of the first linguistic activities of any human society, and this also applies to living organisms. At the same time, names remain longer than uses and other practices, even when traditional knowledge erodes or is lost.

The EBP and affiliated projects offer a unique opportunity to collect the common names in the local languages of the specimens targeted for sequencing in these projects. In this regard, there have been several efforts to collect the Catalan common names for, among others, birds (44,45), butterflies (46), marine mammals (47), fishes (48,49), fungi (50,51) and plants, in particular the latter, given the strong ethnobotanical

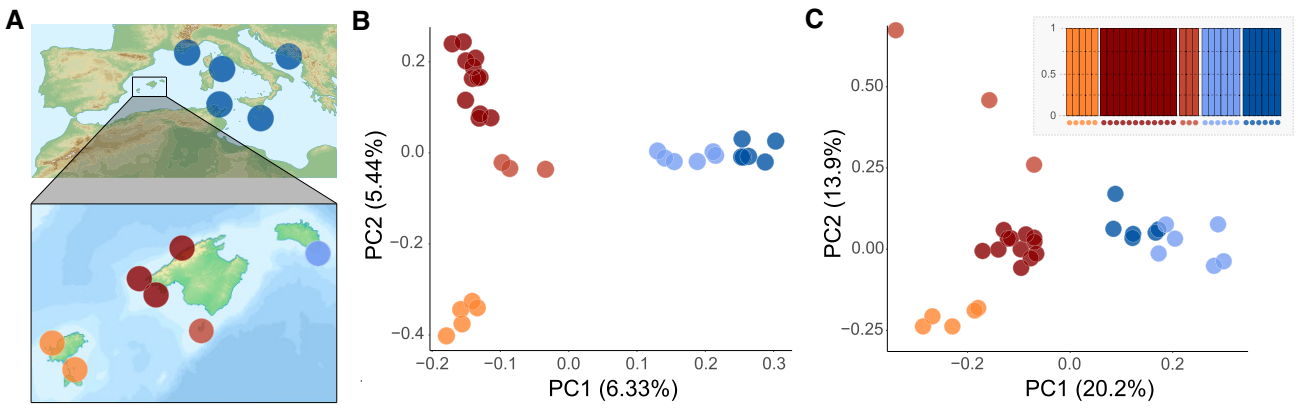

Figure 5. (**A**) Distribution map of the main colonies of *Puffinus mauretanicus* and *P. yelkouan* across the Western Mediterranean. (**B**) Principal component analysis (PCA) summarizing variation across genome-wide resequencing data from 32 *Puffinus* individuals, based on 1 304 832 autosomal SNPs. (**C**) PCA of the same individuals using only the 61 highly-differentiated SNPs. The inset in the top right corner depicts the assignment probability to each population (columns) of every individual (circles) using our SNP panel obtained with *assignPOP* (15). These figures are for illustrative purposes only. The description of the dataset and statistical methods used will be outlined in detail in a manuscript in preparation (Izquierdo-Arànega, Cuevas-Caballé, et al. in preparation).

tradition in the Catalan-speaking territories (52). In the case of plants, for example, there are 35 000 Catalan names corresponding to 6500 taxa (53), from traditional and academic sources, and the public website https://etnobotanica.iec.cat currently contains ca. 80 000 reports of local, traditional plant names for about 1600 taxa that have been quoted by nearly 2000 people. While it is believed that indigenous cultures share a connection between language, cultural identity and land (e.g. (54), we argue that this link can still be traced in post-industrial societies (in particular, in those in which linguistic diversity is also under threat). Thus, the CBP has created a working group to coordinate the different initiatives to compile the Catalan names of species, and to produce standardized descriptions of their meaning that could eventually be used by other EBP nodes.

Discussion

Sequencing, cataloging and characterizing the genomes of every species on Earth will lead to an unprecedented understanding of all phenomena of life as it will make it possible to identify the genetic events underlying the emergence of phenotypes (55). The successful completion of the EBP and associated projects requires, however, the engagement of scientific partners world-wide, and extends beyond scientific organizations to society as a whole. Thus, we believe that the role of structures, such as the CBP, built also to serve as an interface between science and citizenship, is crucial to the success of these projects. Other examples of similar initiatives with widely varying geographic scopes include the DtoL (2), the Portuguese Coalition for Biodiversity Genomics (56), the Africa BioGenome Project (57) or the California Conservation Genomics Project (58) among others.

Indeed, the CBP has triggered unprecedented local collaboration across research communities (from natural history to genomics). In the future, it can help enhance the natural history institutions with the state-of-the-art infrastructure and human resources required to guarantee cataloging, preservation of specimens, tissues and DNA, and curation of vouchers for future generations, as is already happening in many institutions worldwide, to reinvigorate taxonomic research at a time when it is at its historical lowest (e.g. (59)).

The CBP, on the other hand, seeks to involve the entire community based on recognition of strongly shared cultural backgrounds. For instance, given the heterogeneity and often remoteness of the Catalan landscapes, the CBP will need the contribution of an extensive network of naturalist associations familiar with the local ecosystems for streamlined access to biological samples. Other than just a scientific endeavor, we conceive the CBP as part of a world-wide transformative movement that raises social awareness of the threat that biodiversity loss poses to human well-being, and that globally engages the society toward a different and more balanced relationship with nature.

Beyond the borders of the Catalan-speaking territories, and within the EBP network of networks, the CBP aims to play a central role in the biodiversity genomic projects in Europe, currently organized under ERGA, and those of the Iberian Peninsula and the Mediterranean Basin, the geographical regions to which the Catalan territories belong. Therefore, the CBP maintains strong links with ERGA-Spain, ERGA-Andorra and ERGA-France. Nonetheless, the CBP is open to collaboration and sharing resources and expertise with partners all around the world. After the pilot phase of the project, the CBP has been guaranteed continued funding to initiate the first production phase during 2023–2025. Contributing both with data and resources, the CBP aims to engage scientists and citizens in perhaps one of the most important global projects in the history of biology.

Data availability

There is no data associated to this manuscript. All data produced by the CBP can be accessed through the CBP portal at https://dades.biogenoma.cat.

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

We would like to specially thank Harris Lewin, and the EBP leadership, and Mark Blaxter from the Darwin Tree of Life for their continuous and unconditional support. Figure 4 had original contributions from Mercè Rocadembosch, to whom we acknowledge her generous effort. We also thank the Executive Council of the Catalan Society of Biology (SCB) for their commitment to the initiative.

Funding

The CBP acknowledges funding from the Departament de Recerca i Universitats of the Generalitat de Catalunya through the Institut d'Estudis Catalans (IEC). The CBP is partially funded by the IEC, the Barcelona Zoo, the Departament d'Acció Climàtica, Alimentació i Agenda Rural of the Generalitat de Catalunya, the Euroregion Pirineus Mediterrània and the Universitat de Barcelona (UB). Institutional support to Centre Nacional d'Anàlisi Genòmica (CNAG) and Centre de Regulació Genòmica (CRG) was from the Spanish Ministry of Science and Innovation through the Instituto de Salud Carlos III and Generalitat de Catalunya through the Departament de Salut and the Departament de Recerca i Universitats.

Conflict of interest statement

None declared.

References

1. Lewin,H.A., Robinson,G.E., Kress,W.J., Baker,W.J., Coddington,J., Crandall,K.A., Durbin,R., Edwards,S.V., Forest,F., Gilbert,M.T.P., *et al.* (2018) Earth BioGenome Project: sequencing life for the future of life. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, 4325–4333.
2. Blaxter,M.L. (2022) Sequence locally, think globally: the Darwin Tree of Life Project. *Proc. Natl. Acad. Sci. U.S.A.*, **119**, e2115642118.
3. Lewin,H.A., Richards,S., Aiden,E.L., Allende,M.L., Archibald,J.M., Bálint,M., Barker,K.B., Baumgartner,B., Belov,K., Bertorelle,G., *et al.* (2022) The Earth BioGenome Project 2020: starting the clock. *Proc. Natl. Acad. Sci. U.S.A.*, **119**, e2115635118.
4. Casas-Sainz,A.M. and de Vicente,G. (2009) On the tectonic origin of Iberian topography. *Tectonophysics*, **474**, 214–235.
5. Tuel,A. and Eltahir,E.A.B. (2020) Why is the Mediterranean a climate change hot spot?*J. Clim.*, **33**, 5829–5843.
6. Hoegh-Guldberg,O., Jacob,D., Taylor,M., Guillén Bolaños,T., Bindi,M., Brown,S., Camilloni,I.A., Diedhiou,A., Djalante,R., Ebi,K., *et al.* (2019) The human imperative of stabilizing global climate change at 1.5°C. *Science*, **365**, eaaw6974.
7. Cramer,W., Guiot,J., Fader,M., Garrabou,J., Gattuso,J.P., Iglesias,A., Lange,M.A., Lionello,P., Llasat,M.C., Paz,S., *et al.* (2018) Climate change and interconnected risks to sustainable development in the Mediterranean. *Nat. Clim. Chang.*, **8**, 972–80.
8. Pepin,N., Bradley,R.S., Diaz,H.F., Baraer,M., Caceres,E.B., Forsythe,N., Fowler,H., Greenwood,G., Hashmi,M.Z., Liu,X.D., *et al.* (2015) Elevation-dependent warming in mountain regions of the world. *Nat. Clim. Chang.*, **5**, 424–430.
9. Cuevas-Caballé,C., Ferrer Obiol,J., Vizueta,J., Genovart,M., Gonzalez-Solís,J., Riutort,M. and Rozas,J. (2022) The first genome of the balearic shearwater (*Puffinus mauretanicus*) provides a valuable resource for conservation genomics and sheds light on adaptation to a pelagic lifestyle. *Genome Biol. Evol.*, **14**, evac067.
10. Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv doi: https://arxiv.org/abs/1303.3997, 26 May 2013, preprint: not peer reviewed.
11. Poplin,R., Ruano-Rubio,V., DePristo,M.A., Fennell,T.J., Carneiro,M.O., Van der Auwera,G.A., Kling,D.E., Gauthier,L.D., Levy-Moonshine,A., Roazen,D., *et al.* (2017) Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv doi: https://doi.org/10.1101/201178, 24 July 2018, preprint: not peer reviewed.
12. Garrison,E. and Marth,G. (2012) Haplotype-based variant detection from short-read sequencing – free bayes – variant calling – longranger. arXiv doi: https://arxiv.org/abs/1207.3907, 20 July 2012, preprint: not peer reviewed.
13. Danecek,P., Auton,A., Abecasis,G., Albers,C.A., Banks,E., DePristo,M.A., Handsaker,R.E., Lunter,G., Marth,G.T., Sherry,S.T., *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
14. Chang,C.C., Chow,C.C., Tellier,L.C.A.M., Vattikuti,S., Purcell,S.M. and Lee,J.J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, **4**, 7.
15. Chen,K.Y., Marschall,E.A., Sovic,M.G., Fries,A.C., Gibbs,H.L. and Ludsin,S.A. (2018) assignPOP: an r package for population assignment using genetic, non-genetic, or integrated data in a machine-learning framework. *Methods Ecol. Evol.*, **9**, 439–446.
16. Casassas i Simó,O., Camarasa i Castillo,J.M. and Junyent i Rodríguez,C. (2012) Cent anys de la Societat Catalana de Biologia, la primera societat filial de l'Institut d'Estudis Catalans. *Treballs De La Societat Catalana De Biologia*, **63**, 299–324.
17. Formenti,G., Theissinger,K., Fernandes,C., Bista,I., Bombarely,A., Bleidorn,C., Ciofi,C., Crottini,A., Godoy,J.A., Höglund,J., *et al.* (2022) The era of reference genomes in conservation genomics. *Trends Ecol. Evol.*, **37**, 197–202.
18. Mazzoni,C.J., Ciofi,C. and Waterhouse,R.M. (2023) Biodiversity: an atlas of European reference genomes. *Nature*, **619**, 252.
19. Folch i Guillén,R. (2010) In: *Història natural dels Països Catalans Catalana*. Edició: Barcelona: Enciclopèdia Catalana, 1985–2010.
20. Böhne,A., Fernández,R., Leonard,J.A., McCartney,A.M., McTaggart,S., Melo-Ferreira,J., Monteiro,R., Oomen,R.A., Pettersson,O.V. and Struck,T.H. (2023) Contextualising samples: supporting reference genomes for European biodiversity through sample and associated metadata collection. bioRxiv doi: https://doi.org/10.1101/2023.06.28.546652, 30 June 2023, preprint: not peer reviewed.
21. Lawniczak,M.K.N., Durbin,R., Flicek,P., Lindblad-Toh,K., Wei,X., Archibald,J.M., Baker,W.J., Belov,K., Blaxter,M.L., Bonet,T.M., *et al.* (2022) Standards recommendations for the Earth BioGenome Project. *Proc. Natl. Acad. Sci. U.S.A.*, **119**, e2115639118.
22. Santesmasses,D., Mariotti,M. and Guigó,R. (2018) Selenoprofiles: a computational pipeline for annotation of selenoproteins. In: *Methods in Molecular Biology*. Vol. 1661.
23. Vlasova,A., Pulido,T.H., Camara,F., Ponomarenko,J. and Guigó,R. (2021) FA-nf: a functional annotation pipeline for proteins from non-model organisms implemented in nextflow. *Genes (Basel)*, **12**, 1645.
24. Fuentes,D.., Molina,M., Chorostecki,U., Capella-Gutiérrez,S., Marcet-Houben,M. and Gabaldón,T. (2022) PhylomeDB V5: an expanding repository for genome-wide catalogues of annotated gene phylogenies. *Nucleic Acids Res.*, **50**, D1062–D1068.
25. DI Tommaso,P., Chatzou,M., Floden,E.W., Barja,P.P., Palumbo,E. and Notredame,C. (2017) Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, **35**, 316–319.
26. Mölder,F., Jablonski,K.P., Letcher,B., Hall,M.B., Tomkins-Tinch,C.H., Sochat,V., Forster,J., Lee,S., Twardziok,S.O., Kanitz,A., *et al.* (2021) Sustainable data analysis with Snakemake [version 2; peer review: 2 approved]. *F1000Res*, **10**, 33.
27. Ledoux,J.B., Cruz,F., Gómez-Garrido,J., Antoni,R., Blanc,J., Gómez-Gras,D., Kipson,S., López-Sendino,P., Antunes,A., Linares,C., *et al.* (2020) The genome sequence of the octocoral Paramuricea clavata - a key resource to study the impact of climate change in the mediterranean. *G3: Genes, Genomes, Genetics*, **10**, 2941–2952.

28. Cruz,F., Gómez-Garrido,J., Gut,M., Alioto,T.S., Pons,J., Alós,J. and Barcelo-Serra,M. (2023) Chromosome-level assembly and annotation of the Xyrichtys novacula (Linnaeus, 1758) genome. *DNA Res.*, **30**, dsad021.
29. Gomez-Garrido,J., Cruz,F., Alioto,T.S., Feiner,N., Uller,T., Gut,M., Sanchez Escudero,I., Tavecchia,G., Rotger,A., Otalora Acevedo,K.E., *et al.* (2023) Chromosome-level genome assembly of Lilford's wall lizard, Podarcis lilfordi (Günther, 1874) from the Balearic Islands (Spain). *DNA Res.*, **30**, dsad008.
30. Marlétaz,F., Firbas,P.N., Maeso,I., Tena,J.J., Bogdanovic,O., Perry,M., Wyatt,C.D.R., de la Calle-Mustienes,E., Bertrand,S., Burguera,D., *et al.* (2018) Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature*, **564**, 64–70.
31. Torruella,G., Galindo,L.J., Moreira,D., Ciobanu,M., Heiss,A.A., Yubuki,N., Kim,E. and López-García,P. (2023) Expanding the molecular and morphological diversity of Apusomonadida, a deep-branching group of gliding bacterivorous protists. *J. Eukaryot. Microbiol.*, **70**, e12956.
32. Righi,E. and Guigó,R. (2023) The BioGenome Portal: a web-based platform for biodiversity genomics data management. bioRxiv doi: https://doi.org/10.1101/2023.12.20.572408, 22 December 2023, preprint: not peer reviewed.
33. Genovart,M., Arcos,J.M., Álvarez,D., McMinn,M., Meier,R., Wynn,B., R.,G. and Oro,D. (2016) Demography of the critically endangered Balearic shearwater: the impact of fisheries and time to extinction. *J. Appl. Ecol.*, **53**, 1158–1168.
34. Arcos,J.M. (compiler) (2011) International species action plan for the Balearic shearwater, Puffinus mauretanicus. In: *SEO/BirdLife & BirdLife International*.
35. Cortés,V. and González-Solís,J. (2018) Seabird bycatch mitigation trials in artisanal demersal longliners of the Western Mediterranean. *PLoS One*, **13**, e0196731.
36. Rodríguez,A., Arcos,J.M., Bretagnolle,V., Dias,M.P., Holmes,N.D., Louzao,M., Provencher,J., Raine,A.F., Ramírez,F., Rodríguez,B., *et al.* (2019) Future directions in conservation research on petrels and shearwaters. *Front. Mar. Sci.*, **6**, 00094.
37. Louzao,M., Delord,K., García,D., Boué,A. and Weimerskirch,H. (2012) Protecting persistent dynamic oceanographic features: transboundary conservation efforts are needed for the critically endangered balearic shearwater. *PLoS One*, **7**, e35728.
38. Frankham,R. (2005) Genetics and extinction. *Biol. Conserv.*, **126**, 131–140.
39. Charlesworth,D. and Willis,J.H. (2009) The genetics of inbreeding depression. *Nat. Rev. Genet.*, **10**, 783–796.
40. Maffi,L. (2005) Linguistic, cultural and biological diversity. *Annu. Rev. Anthropol.*, **34**, 599–617.
41. Gorenflo,L.J., Romaine,S., Mittermeier,R.A. and Walker-Painemilla,K. (2012) Co-occurrence of linguistic and biological diversity in biodiversity hotspots and high biodiversity wilderness areas. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 8032–8037.
42. Nettle,D. and Romaine, S (2000) In: *Vanishing voices: the extinction of the world's languages*. Oxford University Press.
43. Cámara-Leret,R. and Bascompte,J. (2021) Language extinction triggers the loss of unique medicinal knowledge. *Proc. Natl. Acad. Sci. U.S.A.*, **118**, e2103683118.
44. Ortega i Gonzàlez,E. and (2017) In: *Diccionari etimològic dels noms científics dels ocells dels Països Catalans*. 1st edn. Institut d'Estudis Catalans, Barcelona.
45. Aguiló,C. and Mestre,A. (2017) *Atles ornitonímic de les Illes Balears*.Barcelona/Palma,Institut d'Estudis Catalans / Institut d'Estudis Baleàrics.
46. Vila,R., Stefanescu,C. and Sesma,J.M. (2018) *Guia de les papallones diürnes de Catalunya*. Lynx Edicions, Bellaterra.
47. Lleonart,J. (2012) *Els mamífers marins i els seus noms*. *Terminàlia*, **5**, 7–25.
48. Alegre i Urgell,M. (1992) In: *Espècies pesqueres d'interès comercial: Nomenclatura oficial catalana*. Generalitat de Catalunya. Departament de Cultura, Barcelona.
49. Mercader,L., Lloris,D. and Rucabado,J. (2001) In: *Tots els peixos del Mar Català: Diagnosis i claus d'identificació*. Institut d'Estudis Catalans, Barcelona,.
50. Niell,M. and Girbal,J.M. (2006) *Els noms populars dels bolets a Andorra*. *Rev. Catalana Micol.*, **28**, 209–216.
51. Cuello,J. (2007) In: *Els noms dels bolets*. Lynx Edicions, Bellaterra.
52. Gras,A., Garnatje,T., Marín,J., Parada,M., Sala,E., Talavera,M. and Vallès,J. (2020) The power of wild plants in feeding humanity: a meta-analytic ethnobotanical approach in the Catalan linguistic area. *Foods*, **10**, 61.
53. Vallès,J., Veny,J., Vigo,J., Bonet,M.À., Julià,M.A. and Villalonga,J.C. (2014) In: *Noms de plantes. Corpus de fitonímia catalana. Termcat - Centre de Terminologia & Universitat de Barcelona*. Barcelona.
54. Blythe,J. and Brown,R.M. (2003) Maintaining the links : Language, identity and the land. In: *Proceedings of the 7th FEL Conference, Broom, Western Australia, 22–24 September 2003*. Foundation for Endangered Languages, Bath, UK.
55. Guigó,R. (2023) Genome annotation: from human genetics to biodiversity genomics. *Cell Genomics*, **3**, 100375.
56. Marques,J.P., Alves,P.C., Amorim,I.R., Lopes,R.J., Moura,M., Myers,E.W., Sim-Sim,M., Sousa-Santos,C., Judite Alves,M.V., Borges,P.A., *et al.* (2023) Building a Portuguese Coalition for Biodiversity Genomics. *EcoEvoRxiv, ver 2*.
57. Ebenezer,T.G.E., Muigai,A.W.T., Nouala,S., Badaoui,B., Blaxter,M., Buddie,A.G., Jarvis,E.D., Korlach,J., Kuja,J.O., Lewin,H.A., *et al.* (2022) Africa: sequence 100,000 species to safeguard biodiversity. *Nature*, **603**, 388–392.
58. Shaffer,H.B., Toffelmier,E., Corbett-Detig,R.B., Escalona,M., Erickson,B., Fiedler,P., Gold,M., Harrigan,R.J., Hodges,S., Luckau,T.K., *et al.* (2022) Landscape genomics to enable conservation actions: the California Conservation Genomics Project. *J. Hered.*, **113**, 577–588.
59. Crisci,J.V., Katinas,L., Apodaca,M.J. and Hoch,P.C. (2020) The end of botany. *Trends Plant Sci.*, **25**, 1173–1176.

Appendix

Additional members of the CBP who participated in the project and are not mentioned in the author list: Laura Aguilera, Francisco Cámara, Jèssica Gómez-Garrido, Fernando Cruz, Ignacio Sánchez-Escudero (Centre Nacional d'Anàlisi Genòmica, Universitat de Barcelona, 08028 Barcelona, Spain); Cristian R. Altaba (Conselleria d'Agricultura, Pesca i Medi Natural, Govern de les Illes Balears; Research Group on Human Evolution and Cognition, Universitat de les Illes Balears, 07122 Palma, Illes Balears, Spain); Rui Alves (Departament de Ciències Mèdiques Bàsiques, Universitat de Lleida, 25002 Lleida, Catalonia, Spain); Enrique Arboleda, Maria Capa, Laura Triginer (Centre Balear de Biodiversitat, Univesitat de les Illes Balears, 07122 Palma, Illes Balears, Spain); Vicent Arbona (Departament de Biologia, Bioquímica i Ciències Naturals. Universitat Jaume I. 12071 Castelló, País Valencià, Spain); Conxita Avila, Laura Baldo, Alba Enguidanos, Andrea Prófumo, Owen S. Wangesteen (Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals, Facultat de Biologia; Institut de Recerca de la Biodiversitat, Universitat de Barcelona, 08028 Barcelona, Catalonia, Spain); Margarida Barceló-Serra (Institut Mediterrani d'Estudis Avançats, CSIC-UIB, 07190 Esporles, Illes Balears, Spain); Laura Botigué, Amparo Monfort (Centre for Research in Agricultural Genomics, CSIC-IRTA-UAB-UB, Campus UAB, 08193 Bellaterra, Catalonia, Spain); Bernat Burriel-Carranza; Salvador Carranza, Javier del Campo, Gabriel Mochales-Riaño, Marc Palmada-Flores (Institute of Evolutionary Biology, IBE, UPF-CSIC, PRBB, 08003 Barcelona, Spain); Héctor Candela (In-

stituto de Bioingeniería, Universitat Miguel Hernández d'Elx, 03202, Elx, Spain); Cristian Cañestro, Carlos Carreras, Marc Domènech, Carles Galià-Camps, Sara Guirao-Rico, Jesús Lozano-Fernández, Marta Pascual, Cinta Pegueroles, Alejandro Sánchez-Gracia (Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Institut de Recerca de la Biodiversitat, Universitat de Barcelona, 08028 Barcelona, Catalonia, Spain); Miguel A. Carretero (CIBIO Research Centre in Biodiversity and Genetic Resources, InBIO. Universidade do Porto. Campus de Vairão. 4485-661 Vairão, Portugal); Oriol Grau (Observatori de Recerca del Parc Natural de l'Alt Pirineu, CSIC, Global Ecology Unit CREAF-CSIC-UAB. 08913 Bellaterra. Catalonia, Spain); Romina Garrido, Xavier Grau-Bové, Toni Hermoso, Iana Kim (Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, 08003 Barcelona, Catalonia, Spain); José A. Jurado-Rivera (Departament de Biologia, Genètica. Universitat de les Illes Balears. 07122 Palma de Mallorca, Illes Balears, Spain); Carles Lalueza-Fox (Natural Sciences Museum of Barcelona, 08019, Barcelona, Spain); Jean-Baptiste Ledoux (CIIMAR/CIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, 4099-002 Porto, Portugal); Ramiro Logares, Ramon Massana, Francesc Piferrer (Institut de Ciències del Mar, CSIC, 08003, Barcelona, Catalonia, Spain); Juli Mauri (Fundació Barcelona Zoo, 08013, Barcelona, Catalonia, Spain); Francesc Mesquita-Joanes, Juan S. Monrós (Institut Cavanilles de Biodiversitat i Biologia Evolutiva, ICBIBE, Universitat de València, 46980 Paterna, València, Spain); Jaume Pellicer (Institut Botànic de Barcelona, CSIC-CMCNB, 08038 Barcelona, Catalonia, Spain; Royal Botanic Gardens, Kew, Richmond TW9 3DS, United Kingdom); Josep Peñuelas (Global Ecology Unit CREAF-CSIC-UAB, Campus UAB, 08913 Bellaterra, Catalonia, Spain); Cristina Roquet (Departament de Biologia Animal, Biologia Vegetal i Ecologia, Facultat de Biociències, Universitat Autònoma de Barcelona, 08193 Bellaterra, Catalonia, Spain); Aurora Ruiz-Herrera (Departament de Biologia Cel·lular, Fisiologia i Immunologia, Institut de Biotecnologia i Biomedicina, Universitat Autònoma de Barcelona, 08913 Bellaterra, Spain); Guifré Torruella (Barcelona Supercomputing Centre (BSC-CNS), 08034 Barcelona, Spain); Xavier Turón, Marc Ventura (Centre d'Estudis Avançats de Blanes, CEAB-CSIC, 17300 Blanes, Girona, Catalonia, Spain).

Received: December 29, 2023. **Revised:** May 10, 2024. **Editorial Decision:** June 17, 2024. **Accepted:** June 19, 2024