Title: Best practices for genetic and genomic data archiving

Authors: Deborah M. Leigh[1]*, Amy G. Vandergast[2,] Margaret E. Hunter[3], Eric Crandall[4], W. Chris Funk[5], Colin J. Garroway[6], Sean Hoban[7], Sara J. Oyler-McCance[8], Christian Rellstab[1], Gernot Segelbacher[9], Chloe Schmidt[10], Ella Vázquez-Domínguez[11], Ivan Paz-Vinas[5,12]

***corresponding author: deborah.leigh@wsl.ch**

Deborah M. Leigh

1) Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

ORCID: 0000-0003-3902-2568

deborah.leigh@wsl.ch

Amy G. Vandergast

2) U.S. Geological Survey, Western Ecological Research Center, 4165 Spruance Road, Suite 200, San Diego CA, 92101, USA

ORCID: 0000-0002-7835-6571

avandergast@usgs.gov

Margaret E. Hunter

3) U.S. Geological Survey, Wetland & Aquatic Research Center, 7920 NW 71st Street, Gainesville, Florida 32653, USA

ORCID: 0000-0002-4760-9302

mhunter@usgs.gov

Eric D. Crandall

4) Department of Biology, Pennsylvania State University, 208 Mueller Laboratory, University Park, PA 16802, USA

ORCID: 0000-0001-8580-3651

ecrandall@psu.edu

W. Chris Funk

5) Department of Biology, Graduate Degree Program in Ecology, Colorado State University, 1878 Campus Delivery, Fort Collins, CO 80523-1878, USA

ORCID: 0000-0002-6466-3618

Chris.Funk@colostate.edu

Colin J. Garroway

6) Department of Biological Sciences, University of Manitoba, Winnipeg, Manitoba, Canada

ORCID: 0000-0002-0955-0688

colin.garroway@umanitoba.ca

Sean Hoban
7) Center for Tree Science, The Morton Arboretum, Lisle, IL, 60532, USA *and*
Committee on Evolutionary Biology, University of Chicago, Chicago, IL, 60637, USA
ORCID: 0000-0002-0348-8449
shoban@mortonarb.org

Sara J. Oyler-McCance
8) U.S. Geological Survey, Fort Collins Science Center, 2150 Centre Avenue, Building C, Fort Collins, CO, 80526, USA
ORCID: 0000-0003-1599-8769
soyler@usgs.gov

Christian Rellstab
1) Swiss Federal Research Institute WSL, Birmensdorf, Switzerland
ORCID: 0000-0002-0221-5975
christian.rellstab@wsl.ch

Gernot Segelbacher
9) Wildlife Ecology and Management, University Freiburg, Tennenbacher Str. 4, 79106 Freiburg, Germany
ORCID: 0000-0002-8024-7008
gernot.segelbacher@wildlife.uni-freiburg.de

Chloé Schmidt
10) German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Puschstrasse 4, Leipzig, 04103, Germany
ORCID 0000-0003-2572-4200
chloe.schmidt@idiv.de

Ella Vázquez-Domínguez
11) Departamento de Ecología de la Biodiversidad, Instituto de Ecología, Universidad Nacional Autónoma de México, Coyoacán, Ciudad de México, 04510, México;
ORCID: 0000-0001-6131-2014
evazquez@ecologia.unam.mx

Ivan Paz-Vinas
12) Universite Claude Bernard Lyon 1, LEHNA UMR 5023, CNRS, ENTPE, F-69622, Villeurbanne, France
ORCID: 0000-0002-0043-9289
ivan.paz-vinas@univ-lyon1.fr

Best practices for genetic and genomic data archiving

Abstract: Genetic and genomic data are collected for a vast array of scientific and applied purposes. Despite mandates for public archiving, data are typically used only by the generating authors. The reuse of genetic and genomic datasets remains uncommon because it is difficult, if not impossible, due to non-standard archiving practices and lack of contextual metadata. But as the new field of macrogenetics is demonstrating, if genetic data and their metadata were more accessible and FAIR compliant, they could be reused for many additional purposes. We discuss the main challenges with existing genetic and genomic data archives, and suggest best practices for archiving genetic and genomic data. Recognising that this is a longstanding issue due to little formal data management training within the fields of ecology and evolution, we highlight steps that research institutions and publishers could take to improve data archiving.

Main:

A brief overview on the history and value of genetic data in ecology and evolution

Synthesis of Open Data (publicly archived data, free to reuse) is a powerful tool that is increasingly being used to test pressing questions in ecology and evolution. However, it remains common for valuable datasets to be forgotten after a single use[1–3]. This is a missed opportunity and hinders scientific progress. Producing scientific data is often expensive and time-consuming. Furthermore, most data have numerous potential applications beyond their original use[4].

Public archiving of genetic and genomic sequence data (hereafter 'genetic data') became standard practice in the 1980s[5] but notably, archiving associated *metadata* (data that describe the sampling event, sample, and other derived data), still remains discretionary. Nevertheless, genetic data repositories were some of the earliest Open Data projects (e.g. National Center for Biotechnology Information 'NCBI' GenBank[6]) and continue to arise in response to the increasing needs and volume of genomic data archiving (e.g.[7,8]).

Open population genetic data have now accumulated to the point where data can be synthesized across broad scales (e.g., in macrogenetics[9,10]), rapidly advancing the fields of ecology and evolution by enabling characterization of global biodiversity patterns and genetic diversity trends[10,11]. Sequences within NCBI are now frequently reused for taxonomic assignments, facilitating species discovery and environmental DNA method development[12]. Accessible raw genomic read datasets have also become central to bioinformatic teaching and analysis development (e.g.[13]). Yet the future reuse potential of genetic data extends further, as an abundance of unattempted and unknown uses remain. Vitally, data reuse is one way for countries to help prevent genetic diversity loss through reporting the required genetic indicators of the Convention of Biological Diversity

(CBD) Kunming-Montreal Global Biodiversity Framework (e.g. headline indicator A.4[14–16]).

Despite the long history and growing abundance of Open Genetic Data, journal Open Data policies[21–25], and increasing awareness of the FAIR principles (Findable, Accessible, Interoperable, and Reusable[26]), there are still numerous issues that inhibit comprehensive reuse. These range from issues general to ecology and evolution such as inaccessible private data (comprehensively addressed elsewhere[2,3,27–29]), to field-specific issues we outline below. In this perspective, we suggest guidance on archiving different types of genetic data and their associated metadata. We discuss additional steps or infrastructure needed to improve the status quo. Ultimately, our goal is to prevent data loss and facilitate data reuse.

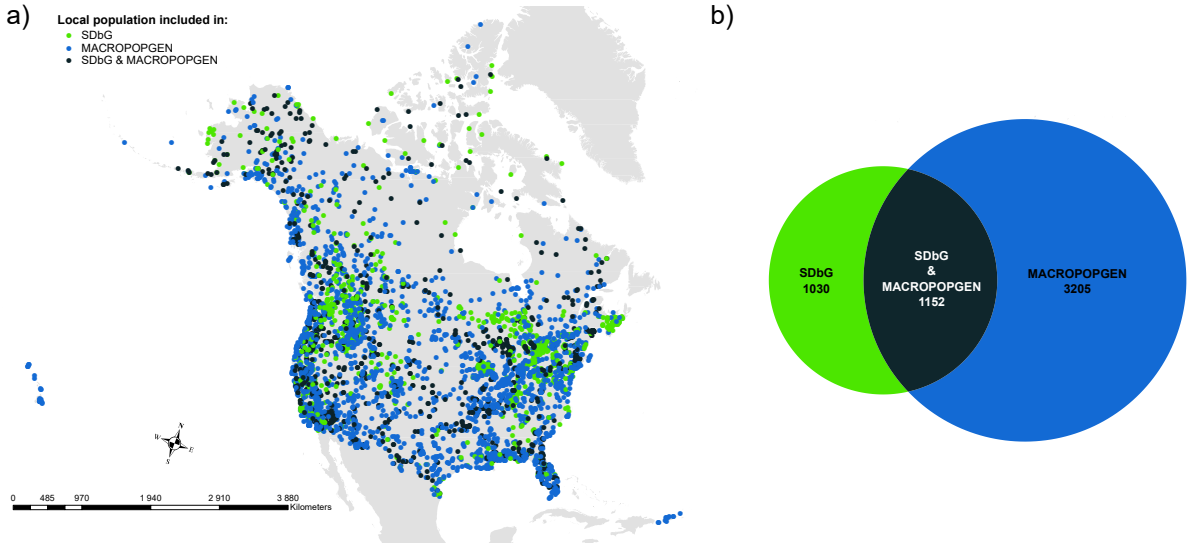


Figure 1: Estimating the unknown number of 'missing' datasets in open repositories. Spatial distribution (a) and proportion (b) of overlapping datasets available in two recently published macrogenetic databases for mammals, birds, reptiles, and amphibians from the USA and Canada: 1) MACROPOPGEN[17], consists of georeferenced microsatellite-derived summary statistics extracted from published articles; 2) SDbG[18–20] consists of raw microsatellite genotype datasets extracted directly from open repositories. After cross checking, only 21.38% of the data entries were found in both databases (black dots), while 59.5% were found exclusively in MACROPOPGEN (blue dots). Low overlap suggests a large proportion of genetic studies included in MACROPOPGEN did not have findable publicly archived data and/or sufficient metadata, and thus were not usable in the SDbG.

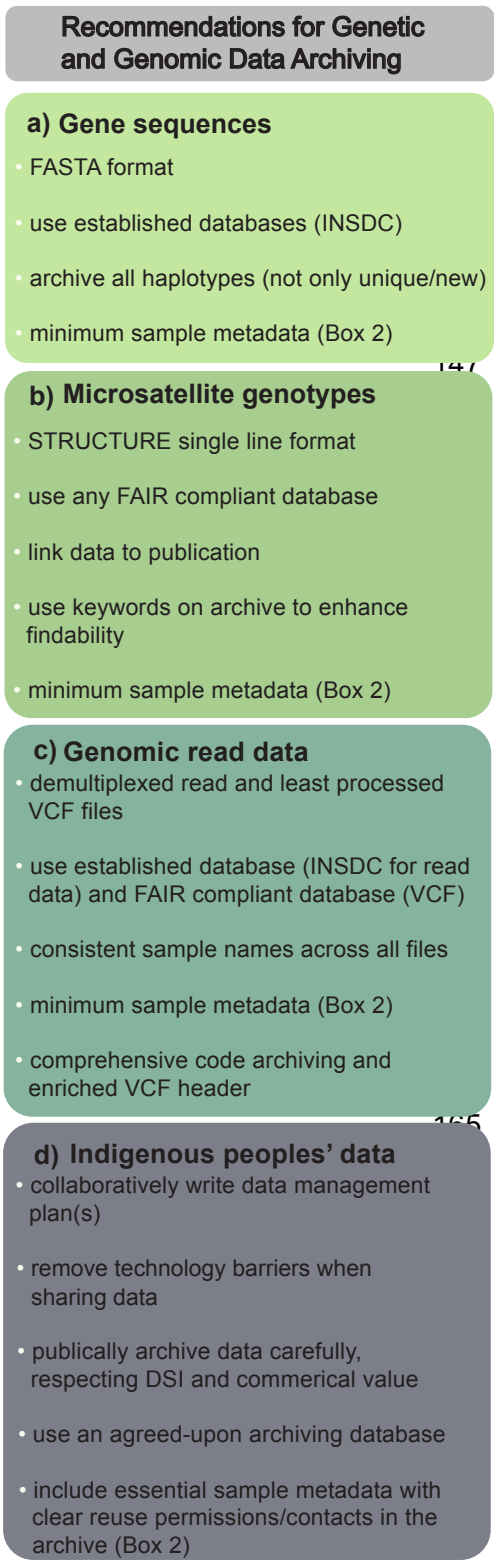


Figure 2: Summary of Best Practice recommendations for each type of genetic data.

The need to improve genetic data archiving practices in ecology and evolution

The FAIR guiding principles are the foundation for good, transparent, and reproducible science. The clearest evidence for this is that Open Data have been used to identify scientific misconduct (e.g.[30]). Datasets are often financed by taxpayers, making public releases of any data an ethical - often mandated[31] - obligation to ensure the full value is obtained. Genomic data production has been particularly expensive, costing hundreds of millions of dollars that would have to be re-spent to regenerate data without archive enrichment[32]. Furthermore, due to the rapid pace of biodiversity loss[33], which includes allelic loss and population extirpation[34], recollecting data may be impossible, rendering existing data irreplaceable. Existing genetic data consequently represent an irreplaceable baseline against which to compare future measurements (i.e. for monitoring[35]). Poor archiving represents a significant loss of time, resources, and opportunities (Figure 1). It is also an unnecessary ethical footprint when data re-generation requires animal handling. Further, increasing genetic data volumes and associated storage energy costs add urgency to the need to improve the archiving standards of genetic data and their metadata by establishing *best practices*.

Researchers can benefit in many ways from publicly archiving genetic data. Open datasets can enhance the scholarly recognition of individual research efforts, because data releases with DOIs and data papers can be cited (e.g. MacroPopGen[17]). As in many disciplines, data papers (e.g. Darwin Tree of Life's Genome Notes) are becoming increasingly popular. Synthesis can also test previously unanswerable big-picture questions in genetics, benefiting researchers through advancement of their field[10].

Best practices for FAIR genetic data archiving

The most widely available genetic data types in molecular ecology and evolution are: a) barcoding/gene sequences (e.g. mitochondrial cytochrome oxidase, the major

histocompatibility complex), b) microsatellite genotypes, and c) genomic read data (i.e. raw high throughput sequences and Single Nucleotide Polymorphisms "SNPs"; Figure 2). The latter two come in a constellation of software-specific formats[36,37] and, due to lack of standardization, repositories contain most of these formats. While format conversion tools exist (PGDspider[36]; Formatomatic[38]; vcftools[39]; plink[40]; adegenet[41]), conversions are time-consuming and often need customization. Mastering each file format also requires specialist knowledge. Consequently, the lack of a standard archived format limits interoperability and reusability. Due to fundamental differences in data types, file sizes and formats used, a single genetic data file format is unrealistic. However, a single file type for each data type is possible and would be a significant advancement.

Unlike other genetic data types, gene sequences are somewhat standardized on archives as FASTA files, and we recommend maintaining this approach (Figure 2a). However, many gene sequences lack essential metadata to allow their reuse. It is important that authors include the minimum metadata needed to interpret their archived data (briefly summarised in Box 1), otherwise archives are challenging to reuse (e.g. non-georeferenced sequences in GenBank[42]).

For microsatellite data (Figure 2b), we suggest archiving in the popular and flexible STRUCTURE format[43]. STRUCTURE input files can handle genotype data of varying ploidy and have a simple format that is conducive to editing in R, spreadsheet software or on the command line, without generating formatting errors. This file format can house metadata (Box 1), as well as marker information (i.e. presence of recessive alleles, inter-marker distances, phase information). We note that there are also variations within the STRUCTURE line format, notably the 1 vs 2 lines *per* individual. Either is suitable for archiving as both are accepted by conversion tools like PGDSpider[36]. However, we recommend use of the single line format to maximize similarity with VCF ("Variant Call Format") files.

Genomic data are often mandated to be publicly archived as raw read data on INSDC servers ("INSDC" International Nucleotide Sequence Database Collaboration)[44], or as aligned BAM files for model organisms[45]. Raw read data can be highly variable, ranging from completely unprocessed files containing several individuals, demultiplexed read files, cleaned files (i.e. with low-quality reads or individuals removed), to error-corrected files (e.g. in ancient DNA)[46]. In contrast to microsatellite data, the variable archiving of genomic data means basic error removal, sample delimitation, and genotype calls are not expected to be present in archived data. We recommend sequencing read data are archived as demultiplexed read files to ensure that separation from key barcode metadata will not render the read data unusable (Figure 2c). This also facilitates archiving of individual sample metadata which has higher reuse potential than study-level metadata. A bioinformatic pipeline can also be challenging to reproduce because there are chronic issues surrounding open code archiving that make it hard to identify what parameters were applied, tool versions used, or even to have access to custom scripts

(further detailed in[27,47]). Even if a pipeline is accessible, version changes of reference genomes or software programs quickly make reproducing it impossible. Thus, we recommend archiving genotype files (detailed below) in addition to demultiplexed sequencing reads to improve the Open Data compliance and reusability of genomic data.

Processed VCF files containing genotype calls (or genotype likelihoods/probabilities) are standard for genomic analyses and we recommend archiving them in parallel with raw read files (notably, this is not possible on INSDC). Although such processed files are not currently widely archived, the practice is becoming more common. Standardization of exactly which variant file is archived also needs consideration. Maximum reusability would be achieved if the archived file represents the least processed genotype (i.e. unfiltered and pruned only for basic errors like technical faults or contamination, with file headers retained to indicate the bioinformatic steps applied and versions used). Notably, archiving VCF files could allow reuse by researchers, managers, or benefit-holders without High-Performance Computing capabilities (e.g. for conservation). Furthermore, VCF archiving would reduce the non-negligible energy, storage and ultimately emissions costs associated with reanalysing raw genomic data[48].

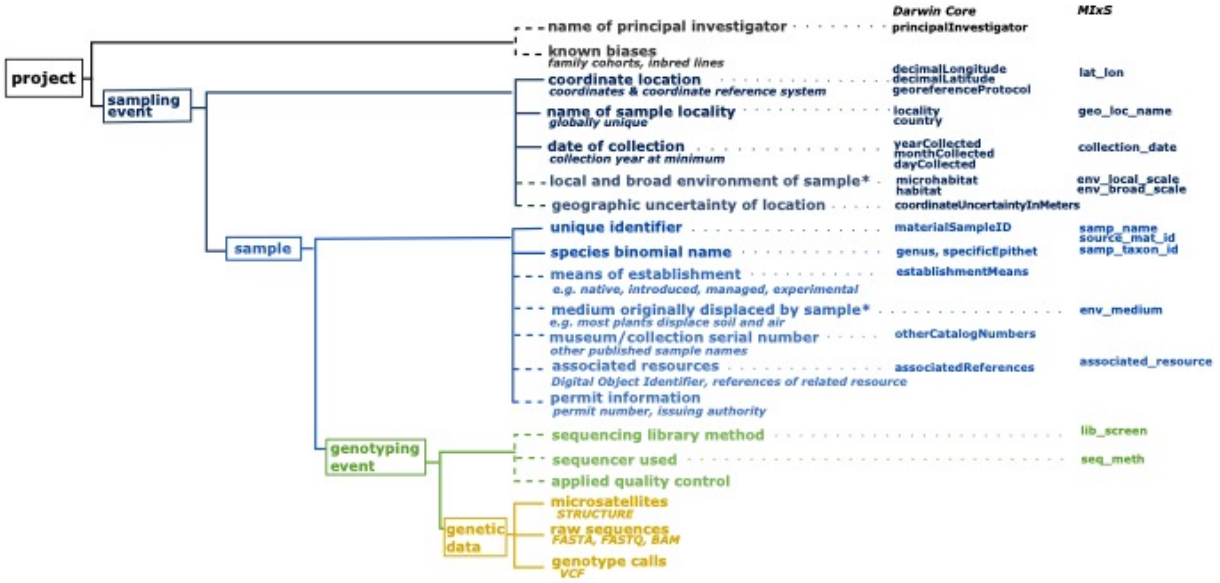
Box 1: Archiving metadata for genetic studies

In genetic studies metadata describing where, when, how and by whom genotype or sequence data were created are invaluable for making data FAIR. There are currently two genomic metadata standards: the Darwin Core standard for biodiversity data[49] and the Minimum Information about any(x) Sequence (MIxS) standard[50]. Both have cross-mapped terms that overlap[51]. The box Figure summarizes term overlap.

Metadata can be viewed in a hierarchical manner based on how they were created starting from the sampling events moving down to the genotyping event. While metadata will vary by sample type and project goals, at a minimum, we suggest that authors provide the required (solid lines) and recommended (dotted lines) categories in the Figure below to improve publicly archived genetic data reuse potential. Terms denoted with * could use controlled vocabulary from the Environment ontology ("ENVO" [52]). To report metadata not covered here, we also recommend using Darwin Core or MixS standards terms to guarantee FAIRness. Sensitive data should be withheld to ensure it is protected. This can be denoted with the terms "informationWithheld", "dataGeneralizations" or "coordinateUncertaintyInMeters". Note, metadata fields might not be adapted for ancient DNA, for which metadata related to sampling events generally does not reflect the age and the environmental conditions of the sampled individual before death. Geological context names may be needed.

An often-overlooked key to FAIR metadata lies in the sample identifier (materialSampleID or samp_name in Darwin Core and MIxS respectively). These identifiers should be unique within the project, and identical between the genetic data and the metadata. To protect genetic data being separated from metadata and help spot errors during complex uploads to databases, we recommend introducing metadata-enriched unique sample names enriched with core metadata like species name, coordinates and/or sampling year (i.e. Capra.ibex_46.97.8.25 or

Capra.ibex.pilatus.2014). Samples that need to be linked across files or studies must be named consistently. We also discourage archiving metadata in separate file repositories from genotype or sequencing data. If unavoidable, we recommend that metadata are stored in a simple table (CSV or text format) with clearly labelled columns (e.g. using MIxS or Darwin Core terms), and consistent sample identifiers. To aid automated retrieval, authors can ensure metadata are machine actionable by avoiding the use of symbols, special characters, and/or colour-based cell codes.



Missing metadata renders most archived data useless

Metadata are a crucial aspect of ensuring genetic data adhere to the FAIR principles [26] because data context vastly increases potential reuses. It was historically standard to only include taxonomic metadata (species and genus) in genetic data archives. Recent mandates have expanded to include the country of collection and collection dates[53,54]. However, neither is sufficient for comprehensive reuse. The minimum required and recommended metadata are shown in Box 1, without which archived data are often functionally useless and could foster incorrect inferences.

We note that metadata from at-risk species, and species that are commercially valuable or desirable, may need to be withheld or obscured to protect them[55]. The most concerning data for such species is location data – coordinates or specific habitat descriptions that would allow public access to these specimens or locations. This is particularly pertinent in genetic studies as individual-level high accuracy coordinates are often collected. Recommended best practices for generalizing sensitive species occurrence or geographic metadata involve masking, controlled access, or not reporting such metadata[56,57]. These limitations should be accepted because it is essential that Open Data do not infringe on privacy, benefit sharing, or species protection efforts[55].

Special considerations when working internationally, with sensitive species, or Indigenous communities

Genetic data and metadata archiving are also key to benefit sharing, including the rights of local communities and local scientists to access data generated from specimens within their country or region. Emerging benefit-sharing requirements, such as those put forth in the Nagoya Protocol and being developed by the CBD (e.g. Digital Sequence Information 'DSI'[58]), are becoming a legal requirement[59]. This is particularly pertinent to ecology and evolution where researchers often work transnationally(e.g.[60,61]) and steps are needed to overcome parachute science[62]. Importantly, while the release of genetic data from sensitive or commercially relevant species could facilitate conservation or evolutionary understanding, protecting their potential commercial value (e.g. pharmaceutical or agricultural) should be carefully considered during archiving.

For work involving Indigenous communities, the CARE principles (Collective benefits, Authority to control, Responsibility, and Ethics[63,64]) could be considered in archiving and reuse of genetic data (Figure 2d). What steps researchers should follow will be situation-specific and developed in conjunction with the benefit holders[65]. Data-generating authors can include specific benefit-sharing statements in publications and in data archives to ensure data sovereignty is upheld. The statement should contain contextual metadata, for instance provenance information, community names, and also clearly outline community-granted permissions for reuse and circulation. Links to biocultural notices created by researchers and endorsement labels issued by Indigenous peoples can also be stored as sample metadata. Authors should also be aware that respecting data sovereignty can influence the data repository used (e.g. Aotearoa Genomic Data Repository, which allows for access only once permission is granted[66]), data storage location[67], and archiving formats (e.g. archiving VCF files is key to limiting technology barriers which otherwise may inhibit reuse by Indigenous researchers or communities). When reusing Indigenous owned genetic data, researchers should also discuss or co-design planned reanalyses with Indigenous communities. Attribution and citation of the original datasets in resulting manuscripts and dissemination of results to the communities involved could further help ensure that cultural authority and sovereignty over reused data remain recognized (e.g.[68]), and that data are not reused inappropriately. Overall, best practices involve improving respect and compliance with the rights Indigenous peoples have for agency over their data.

Key features of data repositories for FAIR data

Currently, genotype data are often stored in generalist Open Data repositories (DRYAD, Zenodo, and increasingly FigShare). However, genetic data can quickly get lost among many other data types archived, where researchers can find everything from non-peer-reviewed ecological survey data (e.g.[69]) to violent crime statistics (e.g.[70]). Local rules and repository fees make it impossible to advocate for a single database for all genotype

data. While there are interoperable search platforms that facilitate simultaneous cross-repository search (e.g. DataONE), their functionality is not guaranteed and database linking has failed in the past (Chloé Schmidt *pers. comm.*). Thus, there is a need for a free[71] inter-government supported public database specifically for archiving genotype data (e.g. microsatellites calls, SNPs, etc). Note that for data involving Indigenous communities, particularly that with restricted reuse, special repositories may be needed to prevent automated retrieval and improper reuse[66].

In lieu of a dedicated repository, researchers can take a few key steps to ensure genotypic data findability. At a minimum, the repositories used must clearly link data to publications and provide citable DOIs. Key metadata fields (Box 1) should be included in the database description to aid findability. Marker type (e.g. “microsatellite” or “SNP”) and key geographical descriptors (e.g. “Kruger National Park”) can also be used as keywords to aid search functions. Researchers could also link genotypic data to “metadatabases” that track samples through metadata and can facilitate upload to the SRA (INSDC BioProjects and BioSamples[72]; Genomic Observatories MetaDatabase ‘GEOME’[73]; Collaborative OPen Omics ‘COPO’[74]).

Researchers can also request new features within existing databases to facilitate data accessibility. The Web of Science’s “associated data” link is a notable advance[75], as is the increased mandatory metadata (sample location, collection date) for BioSample packages and European Nucleotide Archive archives[54,76]. An additional feature, which would benefit multiple disciplines, is the implementation of an automatic identifier for data associated with retracted articles. While datasets from retracted papers should remain online as important records of technical errors, or even fraud, as of writing fraudulent data remain on data repositories with no notice of retraction (e.g.[77]). Similarly, data found to be erroneous remains on sequence databases (e.g. GenBank[78]). Collectively, this poses a huge challenge for studies based on automated data reuse. Archives could flag data with “concerns raised”, “under evaluation”, “technical errors present”, or “retracted” for clarity. Researchers could also benefit from an easy and anonymous way to notify database curators if they encounter incomplete non-FAIR compliant archives to improve database integrity.

The role of scientific institutions and journals in improving data archiving

Funding bodies could take on a greater responsibility to ensure cross-discipline FAIR data archiving (Figure 3). While mandating Open Data has undoubtedly increased data accessibility (e.g.[79,80]), funding bodies could also support researchers with data management plans and confirm implementation (e.g.[81]), check data accessibility, pay archiving fees, and offer general data archiving educational resources or training. For genetic projects, funding bodies can ensure sufficient time is budgeted for archiving because it can take several days.

University libraries or research organizations could support Open Data by hiring data "stewards" or "librarians" familiar with ecological and evolutionary genetic data. Data stewards can help write data management plans, identify suitable databases for genotype file archiving, and ensure dataset longevity through best practice compliance[82]. Few ecology and evolution researchers receive formal training in Open Data or data archiving. Thus they could also offer data management education (e.g. short courses and training) for both students and career scientists (e.g.[83]).

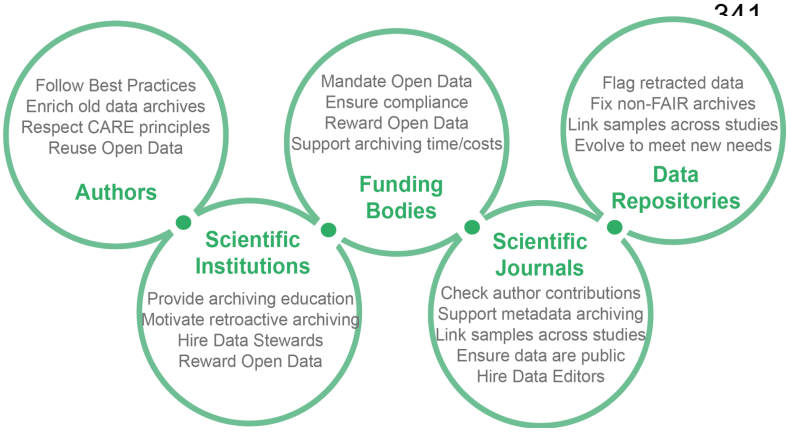


Figure 3: A brief summary of the roles that would improve Open Genetic Data

All scientific institutions could reward researchers with an established history of Open Data through positively valuing Open Data products or dataset citations on *Curriculum Vitae* and in grant proposals.

Scientific journals can also help improve genetic data archiving by ensuring genotype and read data accessibility on FAIR Trustworthy Digital Repositories before final article acceptance[26,47]. Data are often made *accessible upon publication* with links activated when papers are *in press*. However, this makes it impossible for journals to assess data presence and support archiving. A shift to making data accessible *upon submission* to journals is thus needed, particularly at the resubmission stage when papers are close to acceptance[84]. Journal data editors could also check data archives to ensure files are not corrupt, contain the reported number of markers or loci, and contain basic metadata. Scientists concerned with data being accessed prior to publication should note that several databases offer non-public shareable links. Alternatively, journals could make the final acceptance dependent on evidence of FAIR data compliance[84].

Journals could also improve genetic data reuse potential by establishing a mandatory table of standardized metadata terms (see Box 1). However, we note that versioning issues may arise if metadata are in multiple places (e.g. supplementary materials, INSDC, Dryad). Journal data editors could help prevent this by ensuring data from the same sample are linked (i.e. same name) and key differences highlighted (e.g. resequencing with a new technology). As noted above, journals (or monitoring efforts like Retraction Watch) can also inform data repositories if papers have been retracted to advance data reuse. For papers reusing data, journal editors can ensure that datasets are cited correctly (see[85]) and that generating authors receive equal accreditation for their work[86]. We note that peer reviewers should not be tasked with these jobs, because this may be out of their realm of expertise and would increase their already high burden.

Rectifying past mistakes by enriching archived data

An important step many can take to advance Open Data is to improve metadata for existing archives (e.g. GEOME datathons to enrich genetic metadata archives[32]). Publicly accessible metadata are often housed in non-standardized file formats, archived with non-standard terms, or present only in published manuscripts and supplementary files, consequently these can take a significant amount of time to convert for reuse[32]. Retrospective georeferencing is often needed in such situations (e.g.[32, 53,87]), but this often relies on inference (e.g. coordinates derived from place names) leaving significant room for error or lost resolution. We therefore encourage authors to enrich metadata in their old data archives. We would also encourage public archiving of currently inaccessible genetic datasets, and expansion of what was archived (e.g. archive all mtDNA haplotypes rather than only unique haplotypes). Although older datasets may be regarded as being of low value to some authors, when combined with other datasets they can be highly informative and can even provide baselines for important biodiversity protection assessments (e.g.[11,88]).

Data enrichment initiatives could be run at the Department (similar to MoveBank[89]), University library, or country level (e.g. GenDiB and CIEE Living Data). Such retroactive data archives could even be collaboratively published as a "resource" paper (similar to those in Figure 1). These datasets could then support mandated CBD reporting on genetic indicators[16], inform local conservation (e.g.[90]), and identify interesting scientific opportunities (e.g. resampling populations after extreme events[35]).

Box 2: Five take-home messages to improve genetic data archives

1) Archive genetic data in standardized file formats to facilitate reuse (i.e. sequences or barcodes in FASTA; microsatellites in STRUCTURE; SNPs or genomic genotypes in VCF; Genotype likelihoods in VCF; raw genomic data as demultiplexed FASTQ files).
2) Although no centralized database for genotype data exists, these data have great value and should be retroactively archived on FAIR compliant databases to facilitate data rescue.
3) Publicly archive key metadata with the genetic or genomic data, and use enriched sample names (including *a study identifier, species name, coordinates, and sampling year*).
4) To help more colleagues follow the FAIR principles, request both formalized data management support and a higher value of Open Data from research institutions, journals, and funding bodies.
5) For work involving Indigenous communities, carefully archive data affected by the CARE principles so data sovereignty is maintained.

We close on the note that genetic diversity is the most fundamental component of biodiversity[16]. Despite underlying all levels of biodiversity, the biogeographic patterns in

intra-specific genetic diversity are largely understudied and poorly protected[34,88]. Improved archiving (summarised in Box 2) would expand genetic research scales far beyond what any single study or research group could achieve due to logistic, cost, or expertise issues. With data spanning such vast spatial and taxonomic scales, open genetic data will be pivotal to new previously unimaginable areas of research and conservation. Similar to data collected as part of long-term ecological monitoring programs, publicly archived genetic data are likely to only become more valuable and versatile when used in aggregation. This potential is pertinent and timely, due to the recently signed CBD Framework which includes commitments by 192 countries to conserve and restore genetic diversity within and among species' populations, and to monitor and report on progress towards that commitment within the next decades[91]. Better archiving practices will be central to meeting these targets.

Corresponding Author

deborah.leigh@wsl.ch

Acknowledgements

D.M.L. was funded by the BiodivERsA project "ACORN" granted by the Swiss National Science Foundation (SNSF Project 31BD30_193900). I.P-V. was supported by the U.S. Geological Survey John Wesley Powell Center for Analysis and Synthesis. Thanks to Torsten Günther and Bastiaan Star for their comments on ancient DNA archiving practices and considerations. Thanks also to Jennifer Gibson for her helpful discussions about FAIR databases. Thanks to Felix Gugerli and Corine Buser-Schoebel for their helpful feedback on the manuscript. This work was conducted as a part of the Standardizing, Aggregating, Analyzing and Disseminating Global Wildlife Genetic and Genomic Data for Improved Management and Advancement of Community Best Practices Working Group supported by the John Wesley Powell Center for Analysis and Synthesis, funded by the U.S. Geological Survey. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Data accessibility statement:

The data underpinning Figure 1 are available for reviewers at this private link https://docs.google.com/spreadsheets/d/14RyPOCwmZL8LdqH4t-xPIwy2u9QDoWiK/edit?usp=sharing&ouid=110189835630079798646&rtpof=true&sd=true and will be uploaded to a Dryad when a journal link is supplied.

Conflicts of interest:

The authors declare no conflicts of interest.

Author contribution statement

All authors contributed to the inception and writing of this work. DML supervised this work and conducted the editing, with support from IPV, AGV, and MEH. IPV conducted the database included analysis.

References

1. Vines, T. H. *et al.* The Availability of Research Data Declines Rapidly with Article Age. *Curr. Biol.* **24**, 94–97 (2014).
2. Roche, D. G., Kruuk, L. E. B., Lanfear, R. & Binning, S. A. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? *PLOS Biol.* **13**, e1002295 (2015).
3. Tedersoo, L. *et al.* Data sharing practices and data availability upon request differ across scientific disciplines. *Sci. Data* **8**, 192 (2021).
4. Piwowar, H. A., Vision, T. & Whitlock, M. C. Data archiving is a good investment. *Nature* **473**, (2011).
5. Cochrane, G., Cook, C. E. & Birney, E. The future of DNA sequence archiving. *GigaScience* **1**, 2 (2012).
6. Strasser, B. J. The Experimenter's Museum: GenBank, Natural History, and the Moral Economies of Biomedicine. *Isis* **102**, 60–96 (2011).
7. International Human Genome Mapping Consortium *et al.* A physical map of the human genome. *Nature* **409**, 934–941 (2001).
8. Ratnasingham, S., Hebert P,D. bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Mol Ecol Notes.* 7, 355-364 (2007)
9. Blanchet, S., Prunier, J. G. & De Kort, H. Time to Go Bigger: Emerging Patterns in Macrogenetics: Trends in Genetics. *Trends Genet.* **33**, 579–580 (2017).
10. Leigh, D. M. *et al.* Opportunities and challenges of macrogenetic studies. *Nat. Rev. Genet.* **22**, 791–807 (2021).

11. Schmidt, C., Hoban, S. & Jetz, W. Conservation macrogenetics: harnessing genetic data to meet conservation commitments. *Trends Genet.* **39**, 816–829 (2023).
12. Ruppert, K. M., Kline, R. J. & Rahman, M. S. Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Glob. Ecol. Conserv.* **17**, e00547 (2019).
13. Günther, T. & Coop, G. Robust Identification of Local Adaptation from Allele Frequencies. *Genetics* **195**, 205–220 (2013).
14. CBD. Decision Adopted By The Conference Of The Parties To The Convention On Biological Diversity. (2022) doi:https://www.cbd.int/doc/decisions/cop-15/cop-15-dec-05-en.pdf.
15. Hoban, S. *et al.* Genetic diversity targets and indicators in the CBD post-2020 Global Biodiversity Framework must be improved. *Biol. Conserv.* **248**, 108654 (2020).
16. Hoban, S. *et al.* Monitoring status and trends in genetic diversity for the Convention on Biological Diversity: An ongoing assessment of genetic indicators in nine countries. *Conserv. Lett.* **16**, e12953 (2023).
17. Lawrence, E. R. *et al.* Geo-referenced population-specific microsatellite data across American continents, the MacroPopGen Database. *Sci. Data* **6**, 14 (2019).
18. Schmidt, C., Domaratzki, M., Kinnunen, R. P., Bowman, J. & Garroway, C. J. Continent-wide effects of urbanization on bird and mammal genetic diversity. *Proc Roc B* **287**, (2020).
19. Schmidt, C. & Garroway, C. J. Systemic racism alters wildlife genetic diversity. *Proc. Natl. Acad. Sci.* **119**, e2102860119 (2022).
20. Schmidt, C. & Garroway, C. J. The population genetics of urban and rural amphibians in North America. *Mol. Ecol.* **30**, 3918–3929 (2021).
21. Rieseberg, L., Vines, T. & Kane, N. Editorial and retrospective 2010. *Mol. Ecol.* **19**, 1–22 (2010).
22. Moore, A. J., Mcpeek, M. A., Rausher, M. D., Rieseberg, L. & Whitlock, M. C. The need for archiving data in evolutionary biology. *J. Evol. Biol.* **23**, 659–660 (2010).

23. Whitlock, M. C. Data archiving in ecology and evolution: best practices. *Trends Ecol. Evol.* **26**, 61–65 (2011).
24. Fairbairn, D. J. The advent of mandatory data archiving. *Evolution* **65**, 1–2 (2011).
25. Berberi, I. & Roche, D. G. No evidence that mandatory open data policies increase error correction. *Nat. Ecol. Evol.* **6**, 1630–1633 (2022).
26. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
27. Gomes, D. G. E. *et al.* Why don't we share data and code? Perceived barriers and benefits to public archiving practices. *Proc Roc B* **289**, 2022111 (2022).
28. Huang, X. *et al.* Willing or unwilling to share primary biodiversity data: results and implications of an international survey: Biodiversity data sharing and archiving. *Conserv. Lett.* **5**, 399–406 (2012).
29. Hostler, T. J. The Invisible Workload of Open Research. *J. Trial Error* (2023) doi:10.36850/mr5.
30. Kozlov, M. How A Spider-Biology Scandal Upended Researchers' Lives. *Nature* **608**, (2022).
31. European Commission. H2020 Programme: AGA – Annotated Model Grant Agreement. (2019).
32. Crandall, E. D. *et al.* Importance of timely metadata curation to the global surveillance of genetic diversity. *Conserv. Biol.* **37**, e14061 (2023).
33. Ceballos, G. *et al.* Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Sci. Adv.* **1**, e1400253 (2015).
34. Leigh, D. M., Hendry, A. P., Vázquez-Domínguez, E. & Friesen, V. L. Estimated six per cent loss of genetic variation in wild populations since the industrial revolution. *Evol. Appl.* **12**, 1505–1512 (2019).
35. Jensen, E. L. & Leigh, D. M. Using temporal genomics to understand contemporary climate change responses in wildlife. *Ecol. Evol.* **12**, (2022).

523 36.Lischer, H. E. L. & Excoffier, L. PGDSpider: an automated data conversion tool for
524 connecting population genetics and genomics programs. *Bioinformatics* **28**, 298–299 (2012).
525 37.Adamack, A. T. & Gruber, B. P OP G EN R EPORT : simplifying basic population genetic
526 analyses in R. *Methods Ecol. Evol.* **5**, 384–387 (2014).
527 38.Manoukis, N. C. formatomatic: a program for converting diploid allelic data between common
528 formats for population genetic analysis. *Mol. Ecol. Notes* **7**, 592–593 (2007).
529 39.Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158
530 (2011).
531 40.Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based
532 Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
533 41.Jombart, T. *adegenet* : a R package for the multivariate analysis of genetic markers.
534 *Bioinformatics* **24**, 1403–1405 (2008).
535 42.Gratton, P. *et al.* A world of sequences: can we use georeferenced nucleotide databases for
536 a robust automated phylogeography? *J. Biogeogr.* **44**, 475–486 (2017).
537 43.Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of Population Structure Using
538 Multilocus Genotype Data. *Genetics* **155**, 945–959 (2000).
539 44.Cochrane, G., Cook, C. E. & Birney, E. The future of DNA sequence archiving. *GigaScience*
540 **1**, 2 (2012).
541 45.Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–
542 2079 (2009).
543 46.Mallick, S., *et al.* The Allen Ancient DNA Resource (AADR) a curated compendium of ancient
544 human genomes. *Sci Data* **11**, 182 (2024).
545 47.Jenkins, G. B. *et al.* Reproducibility in ecology and evolution: Minimum standards for data
546 and code. *Ecol. Evol.* **13**, e9961 (2023).
547 48.Grealey, J. *et al.* The Carbon Footprint of Bioinformatics. *Mol. Biol. Evol.* **39**, msac034
548 (2022).

549 49.Wieczorek, J. *et al.* Darwin Core: An Evolving Community-Developed Biodiversity Data
550 Standard. *PLOS ONE* **7**, e29715 (2012).
551 50.Field, D. *et al.* The Genomic Standards Consortium. *PLoS Biol.* **9**, e1001088 (2011).
552 51.Meyer, R. *et al.* Aligning Standards Communities for Omics Biodiversity Data: Sustainable
553 Darwin Core-MIxS Interoperability. *Biodivers. Data J.* **11**, e112420 (2023).
554 52.Buttigieg, P. *et al.* The environment ontology: contextualising biological and biomedical
555 entities. *J. Biomed. Semant.* **4**, 43 (2013).
556 53.Böhne, A. *et al.* Contextualising samples: Supporting reference genomes of European
557 biodiversity through sample and associated metadata collection. biorxiv.
558 https://doi.org/10.1101/2023.06.28.546652 (2024).
559 54.Stroe, O. ENA to introduce mandatory spatiotemporal annotations.
560 https://www.ebi.ac.uk/about/news/updates-from-data-resources/ena-spatiotemporal-
561 metadata/ (2023).
562 55.Frank, R. D., Kriesberg, A., Yakel, E. & Faniel, I. M. Looting hoards of gold and poaching
563 spotted owls: Data confidentiality among archaeologists & zoologists. *Proc. Assoc. Inf. Sci.*
564 *Technol.* **52**, 1–10 (2015).
565 56.Chapman, A. D. Current Best Practices for Generalizing Sensitive Species Occurrence Data.
566 57.Clarke, K. C. A multiscale masking method for point geographic data. *Int. J. Geogr. Inf. Sci.*
567 **30**, 300–315 (2016).
568 58.Scholz, A. H. *et al.* Multilateral benefit-sharing from digital sequence information will support
569 both science and biodiversity conservation. *Nat. Commun.* **13**, 1086 (2022).
570 59.Marden, E. *et al.* Sharing and reporting benefits from biodiversity research. *Mol. Ecol.* **30**,
571 1103–1107 (2021).
572 60.Bhaumik, V. Global inequities in local science. *Nat. Ecol. Evol.* **7**, 793–793 (2023).
573 61.Miller, J., White, T. B. & Christie, A. P. Parachute conservation: Investigating trends in
574 international research. *Conserv. Lett.* **16**, e12947 (2023).

62.de Vos, A. & Schwartz, M. W. Confronting parachute science in conservation. *Conserv. Sci. Pract.* **4**, e12681 (2022).

63.Carroll, S. R. The CARE Principles for Indigenous Data Governance - Data Science Journal. *Data Sci. J.* **19**, 1–12.

64.Carroll, S. R., Herczog, E., Hudson, M., Russell, K. & Stall, S. Operationalizing the CARE and FAIR Principles for Indigenous data futures. *Sci. Data* **8**, 108 (2021).

65.Kukutai, T. Indigenous data sovereignty—A new take on an old theme. *Science* **382**, eadl4664 (2023).

66.Te Aika, B. *et al.* Aotearoa genomic data repository: An āhuru mōwai for taonga species sequencing data. *Mol. Ecol. Resour.* 1-14 (2023)

67.Hudson, M. *et al.* Indigenous Peoples' Rights in Data: a contribution toward Indigenous Research Sovereignty. *Front. Res. Metr. Anal.* **8**, (2023).

68.Mc Cartney, A. M. *et al.* Indigenous peoples and local communities as partners in the sequencing of global eukaryotic biodiversity. *Npj Biodivers.* **2**, 1–12 (2023).

69. Shaikh A. (2014). Ecology week 4: field sample with animals. figshare. Dataset. https://doi.org/10.6084/m9.figshare.1194651.v1

70. Gonzalez L. (2010). Sexual crime in Colombia 2010-2022. figshare. Dataset. https://doi.org/10.6084/m9.figshare.21937154.v1

71.Roche, D. G., Jennions, M. D. & Binning, S. A. Fees could damage public data archives. *Nature* **502**, 171–171 (2013).

72.Barrett, T. *et al.* BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.* **40**, D57–D63 (2012).

73.Deck, J. *et al.* The Genomic Observatories Metadatabase (GeOMe): A new repository for field and sampling event metadata associated with genetic samples. *PLOS Biol.* **15**, e2002925 (2017).

74.Shaw, F. *et al.* COPO: a metadata platform for brokering FAIR data in the life sciences. *F1000Research* **9**, 495 (2020).

75.Web of Science. Associated Data. (2018). https://images.webofknowledge.com/images/help/WOK/hp_associated_data.html

76. DDBJ Including Sample Location and Collection Date and Time for BioSample submissions Including Sample Location. (2023) https://www.ddbj.nig.ac.jp/news/en/2023-05-02-e.html

77.Costa-Pereira, R. & Pruitt, J. Retraction: Behaviour, morphology and microhabitat use: what drives individual niche variation? *Biol. Lett.* **16**, 20200588 (2020).

78.van den Burg, M. P. & Vieites, D. R. Bird genetic databases need improved curation and error reporting to NCBI. *Ibis* **165**, 472–481 (2023).

79. National Institute for Health. Final NIH Policy for Data Management and Sharing: NOT-OD-21-013 https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.htm (2020)

80.European Commission. Guidelines on FAIR data management in Horizon 2020. (2016).

81.UKRI. Data Management Plan: Guidance for Peer Reviewers. (2013) https://www.ukri.org/wp-content/uploads/2021/07/ESRC-200721-DataManagementPlan-GuidanceforPeerReviewers.pdf .

82.Peng, G. *et al.* Scientific Stewardship in the Open Data and Big Data Era Roles and Responsibilities of Stewards and Other Major Product Stakeholders. *D-Lib Mag.* **22**, (2016).

83.Toelch, U. & Ostwald, D. Digital open science—Teaching digital tools for reproducible and transparent research. *PLOS Biol.* **16**, e2006022 (2018).

84.Thrall, P. H. *et al.* From raw data to publication: Introducing data editing at Ecology Letters. *Ecol. Lett.* **26**, 829–830 (2023).

85.Cousijn, H. *et al.* A data citation roadmap for scientific publishers. *Sci. Data* **5**, 180259 (2018).

86.Nature. Time to recognize authorship of open data. *Nature* **604**, 8–8 (2022).

87.Miraldo, A. *et al.* An Anthropocene map of genetic diversity. *Science* **353**, 1532–1535 (2016).

88. Figuerola-Ferrando, L. *et al.* Global patterns and drivers of genetic diversity among marine habitat-forming species. *Glob. Ecol. Biogeogr.* **32**, 1218–1229 (2023).
89. Kays, R. *et al.* The Movebank system for studying global animal movement and demography. *Methods Ecol. Evol.* **13**, 419–431 (2022).
90. Beninde, J. *et al.* CaliPopGen: A genetic and life history database for the fauna and flora of California. *Sci. Data* **9**, 380 (2022).
91. Hoban, S. *et al.* Genetic diversity goals and targets have improved, but remain insufficient for clear implementation of the post-2020 global biodiversity framework. *Conserv. Genet.* **24**, 181–191 (2023).

******** END*******