Title: Don’t make genetic data disposable: Best practices for genetic and genomic data archiving

Authors: Deborah M. Leigh[1]*, Amy 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[12,13]

***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 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, Sirenia Project 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

Chloe 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) Department of Biology, Colorado State University, 1878 Campus Delivery, Fort Collins, CO, 80523-1878, USA
ORCID: 0000-0002-0043-9289

13) Université Claude Bernard Lyon 1, CNRS, ENTPE, UMR5023 LEHNA, F-69622, Villeurbanne, France
ivan.paz-vinas@univ-lyon1.fr

Don’t make genetic data disposable:
Best practices for genetic and genomic data archiving

Abstract

In ecology and evolution, genetic and genomic data are commonly collected for a vast array of scientific and applied purposes. Despite mandates for public archiving, such data are typically used only once by the data-generating authors. The repurposing of genetic and genomic datasets remains uncommon because it is often difficult, if not impossible, due to non-standard archiving practices and lack of contextual metadata. But as the new research field of macrogenetics is demonstrating, if genetic data and their metadata were more accessible, they could be reused for many additional purposes, far beyond their initial intended impact. In this review, we outline the main challenges with existing genetic and genomic data archives, factors underlying the challenges, and current best practices for archiving genetic and genomic data. Recognising that this is a longstanding issue due to an absence of formal data management training within the research field of ecology and evolution, we highlight key steps that universities, funding bodies, and scientific publishers could take to ensure timely change towards good data archiving.

Introduction

Synthesis of Open Data (publicly archived data free to reuse) is a powerful tool that is increasingly being used to test pressing big-picture questions at large scales in ecology and evolution. However, it still remains common for valuable datasets to be forgotten and mislaid after a single use (Vines et al. 2014; Roche et al. 2015; Tedersoo et al. 2021). This is a missed opportunity and hinders scientific progress. Producing and collecting scientific data is often expensive and time-consuming. Furthermore, most data have numerous potential applications beyond their original use (Piwowar et al. 2011).

Public archiving of genetic and genomic sequence data (hereafter ‘genetic data’) became standard practice in the 1980s (Cochrane et al. 2012), but notably, public archiving of associated metadata (metadata are data that describe other data, including species name, sampling coordinates, sampling year, etc.), still remains discretionary. Nevertheless, genetic data repositories were some of the earliest Open Data projects and databases (e.g. Genbank; Strasser et al. 2011) and continue to arise to meet the increasing needs of genomic data archiving (e.g. International Human Genome Mapping Consortium 2001; BOLD, Ratnasingham and Hebert 2007).

Repurposed ‘open’ population genetic data has only just begun to accumulate but has facilitated reconstruction of, for example, endangered species’ demographic histories (e.g. orangutans; Nater et al. 2015), and inference of global invasion pathways (e.g. *Trachemys scripta elegans*; Espindola et al. 2022). Multi-species genotype data are also being synthesized across large spatial and temporal ranges for macrogenetic studies

(Blanchet et al. 2017; Leigh et al. 2021), rapidly advancing molecular ecology and evolution by characterizing global biodiversity patterns, genetic diversity trends, and informing biodiversity conservation (see Leigh et al. 2021 and references therein; Schmidt et al. 2023). Sequences within the National Center for Biotechnology Information (NCBI) are frequently reused as a biodiversity reference database, facilitating new species discovery and the emergence of environmental DNA methods (Ruppert et al. 2019). Accessible raw genomic read datasets have been important for teaching bioinformatics and developing genomic analyses (e.g. Günther and Coop, 2013). Yet the future repurposing potential of genetic data extends further, as an abundance of unattempted and unknown uses remain. Vitally, repurposing of public genetic data is one way for countries to report genetic indicators required by the Convention of Biological Diversity (CBD) post-2020 Kunming-Montreal global biodiversity framework (e.g. headline indicator A.4; CBD 2022; Hoban et al. 2023; Hoban et al. 2020).

Despite the abundance of genetic data in open repositories, long-standing willingness of journals to mandate Open Data (e.g. JDAP Dryad 2011; Rieseberg et al. 2010; Moore et al. 2010; Whitlock 2011; Fairbairn 2011), and increasing popularity of the FAIR principles (Findable, Accessible, Interoperable, and Reusable; Wilkinson et al. 2016), there are still numerous issues with genetic data archiving that inhibit comprehensive repurposing. Many genetic datasets remain in private hands, often stored on private storage devices, shared only on request (see Box 1). However, such devices will quickly depreciate, leading to data loss. Furthermore, cross-disciplinary studies have shown that many authors do not share data upon request, despite committing to do so in data availability statements of their articles (Gabelica et al. 2022; Crandall et al. 2023). Notably, ecologists (including molecular ecologists) were the most likely researchers to ignore data request emails (Tedersoo et al. 2021). Data findability is also limited by the repositories available, because genetic databases are built for nucleotide sequence data and do not accept non-nucleotide data (e.g. processed Single Nucleotide Polymorphism “SNP” genotypes and/or genotypic datasets based on microsatellites). These abundant and valuable genotype data are subsequently archived, if at all, across an array of public and private data repositories (e.g. Dryad, FigShare, ScienceBase, Zenodo, and personal or institutional servers), making them hard to find (see Box 1).

Alongside issues of data accessibility and findability, genetic data that are publicly archived are often in an unsuitable format. This is predominantly because they lack key metadata. It is easy to underestimate the severity of poor metadata archiving: only ~6.5% of published nucleotide data for land-living vertebrates are georeferenced in GenBank (Gratton et al. 2017). Similarly, only 13% of biodiversity-relevant Sequence Read Archive (SRA) BioProjects have spatiotemporal metadata (Toczydlowski et al. 2021), and less than 33% of metagenomic data are archived with vital contextual environmental data (Schriml et al. 2020). Beyond the absence of critical metadata, non-sequence genotypic data are also archived in an array of formats often tailored for specific software packages, which themselves change over time. Moreover, studies frequently differ in exactly what they archive. This ranges from aligned sequences to raw data in genomic datasets, to newly identified haplotypes only vs the entire set of sequences obtained in amplicon or barcoding studies (Paz-Vinas et al. 2021).

Consequently, despite the long history of publicly archived genetic data in ecology and evolution, most remain difficult to repurpose. Poor archiving represents a significant waste of public funds and a loss of time, resources, and opportunities for scientists and practitioners. This also represents an unnecessary ethical footprint because some genetic studies require animal handling. It severely limits the development of promising Open Data reliant research avenues and biodiversity monitoring. Collectively, poor archiving lowers the impact of each dataset. There is a pressing need in ecology and evolution to improve the archiving standards of genetic data and their metadata by establishing best practices. In this review, we seek to address these challenges by offering guidance on archiving different types of genetic data and their associated metadata. We also discuss additional steps or infrastructure needed to improve the status quo. Ultimately, our goal is to prevent data loss and facilitate data reuse.

Box 1: Estimating the unknown number of 'missing' datasets

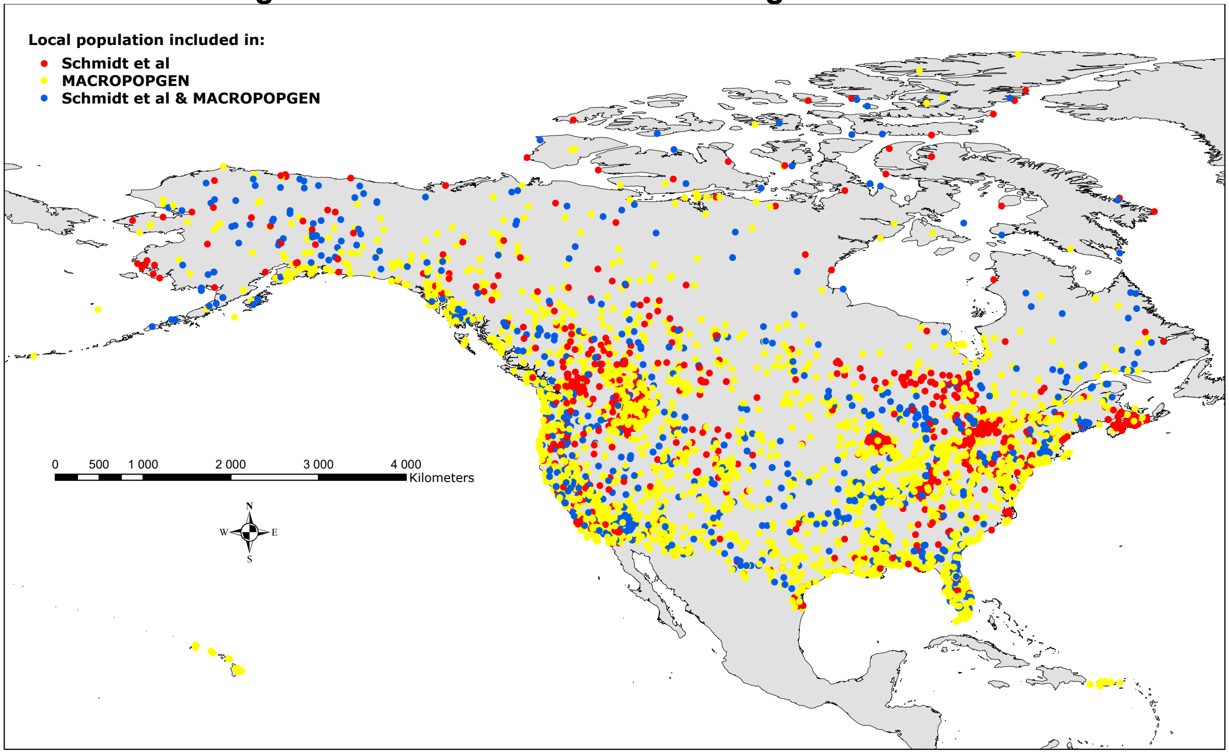

Figure 1: The plotted data from MACROPOPGEN and Schmidt et al. macrogenetic databases (described in box). Yellow dots represent data exclusively in MACROPOPGEN, red dots represent data exclusively in the Schmidt et al. database, blue dots indicate those included in

both databases.

To estimate the number of genetic datasets that have not been publicly archived, we compared the datasets synthesized in two macrogenetic databases with similar taxonomic and spatial scopes that differed in the form of included data. The first macrogenetic database was MACROPOPGEN (Lawrence et al. 2019), which is a compilation of georeferenced vertebrate microsatellite-based genetic summary statistics and related metadata (e.g. taxonomic group) across the Americas for 897 species and 9,090 populations. This database was built by extracting data (e.g. genetic summary statistics) from published articles and reports, irrespective of whether raw genotypic datasets from which the compiled data were derived are publicly available. The second macrogenetic database (hereafter, Schmidt et al. database) was compiled by Schmidt and Garroway (2021; 2022) and Schmidt et al. (2020) and comprises repurposed microsatellite genotype datasets that were archived in open repositories (mostly those archived on Dryad) for terrestrial vertebrates across USA and Canada.

If all raw genetic datasets used in the studies identified in MACROPOPGEN were publicly available, we would expect the proportion of datasets overlapping with Schmidt et al. to be high (e.g. 80-90%), and if all data were clearly linked and/or archived with metadata we would expect the overlap to increase further (e.g. 90-100%).

We extracted datasets from birds, amphibians, mammals, and reptiles located in USA and Canada from both databases, and combined them to create a pooled database containing data for 5,395 populations from 412 species, with 68.48% and 31.52% of data originating from MACROPOPGEN and the Schmidt et al. database, respectively. Data were at the population level (Figure 1). We then crosschecked different metadata fields (e.g. species name, DOI identifier of the dataset and/or of the original article, author names) among data entries from each macrogenetic database to identify populations and datasets that were included in both macrogenetic databases.

Only 21.38% of the data entries in the combined database were found in both macrogenetic databases (Figure 1, blue dots), while 59.5% were included exclusively in MACROPOPGEN (Figure 1, yellow dots). While this does not comprehensively assess the number of publications missing from public data archives, this strongly indicates a large proportion of the studies behind this 59.5% are unlikely to have public genetic or genotype datasets, given the similarity in spatial and taxonomic scopes used by the authors of both databases.

Why are we not archiving comprehensively?

Several general issues drive variation in data archiving practices that ultimately hinder data reuse (see Box 2; Tedersoo et al. 2021; Gomes et al. 2022; Huang et al. 2012; Roche et al. 2014; Hostler et al. 2023). Most commonly, poor data archiving is driven by researchers not having sufficient support or time to archive comprehensively (Hostler et al. 2023). Furthermore, authors often may not realize their archives are incomplete and challenging to reuse. A well-archived dataset should contain all of the relevant information needed to reproduce or repurpose a study without the need to consult multiple sources.

Limited data archiving can sometimes be a requirement in ecology and evolution. There are legal and ethical considerations in data archiving and reuse that can directly limit what data can be made public (see Box 3). These limitations should be accepted because it is essential that Open Data does not infringe on privacy, benefit sharing, or species protection efforts (Frank et al. 2016). Unfortunately, however, limited data archiving can sometimes be intentional to prevent repurposing; in contrast to sensitive data intentionally poor archiving needs to be rectified as it inhibits FAIR compliance (see Box 2 and 3).

Box 2: Limiting competition through intentionally poor data archiving

FAIR data archiving can fail due to a range of reasons (e.g. Tedersoo et al. 2021; Gomes et al. 2022; Huang et al. 2012; Roche et al. 2014; Hostler et al. 2023). These include a fear of scooping - when data are repurposed before the data-generating authors benefit through primary publication(s), collegial competition (Huang et al. 2012), or because researchers consider their study topic exclusive. Reluctance to archive data may also be related to a fear that reanalysis of data might reveal errors or lead to contradictory conclusions (Wicherts et al. 2011). Intentionally poor archiving can sadly inhibit reanalysis and synthesis that can lead to exciting new conclusions (e.g. Schumacher et al. 2022) and could harm biodiversity monitoring efforts.

We acknowledge that there are long standing valid concerns surrounding data archiving from long-term multi-grant studies (e.g. wild pedigreed populations) because reuse by external researchers at any point could have a disproportionately negative impact on data-generating scientists and the project (Mills et al. 2015; Whitlock et al. 2015). Similarly, data-repurposing from early career researchers (particularly matriculated students) can be harmful because they commonly have long delays prior to publication. In both such situations, to ensure FAIR compliance, data-generating researchers have an obligation to archive their data, but can use embargos to protect their planned analyses/publications (Mills et al. 2015; Whitlock et al. 2015). After embargos expire, ethical and sensitive data repurposing is vital in such situations to help maintain FAIR compliance; this includes discussion of data repurposing with generating authors and land owners/indigenous communities.

Box 3: How to archive data from sensitive species

Metadata from threatened or endangered species, as well as species that are commercially valuable or desirable, may need to be withheld or obscured to protect them (Frank et al. 2016). Withholding metadata facilitates species protection by mitigating the risk of poaching and/or habitat degradation caused by increased disturbance arising from species viewing or photography (Lindenmayer and Scheele 2017). Furthermore, for species on private land, this can protect the collaborations necessary for conservation (Lindenmayer and Scheele 2017). However, some argue that metadata must be published albeit with considerations (i.e. masking) or accessible upon request (Lowe et al. 2017).

The data of most concern for such species is location data – coordinates or specific habitat descriptions that would allow public access to these species. Best practices for generalizing sensitive species occurrence or geographic metadata have been developed (see Chapman 2020; Clarke 2016). For example, the Global Biodiversity Information Facility (GBIF) has mechanisms to incorporate location generalization and ways to document that information exists but is withheld for privacy (e.g. metadata field "informationWithheld"). Importantly the release of other data from sensitive species, such as genetic and genomic data, could facilitate conservation but their potential commercial value (e.g. for pharmaceutical or agricultural companies) should not be ignored. While we acknowledge some analyses cannot be performed without fairly accurate location data (e.g. genotype-environment association, macrogenetics), access can normally be arranged when needed. Release of metadata (e.g. number of individuals, age/reproductive status, sex, etc.), should be dependent on the potential risk to the species (i.e. providing age and size of a valuable tree, game species, or a medicinal herb may increase the likelihood harvest, but will be unimportant in other cases; Lowe et al. 2017). Chapman (2022) provides a decision tree to assist with such choices.

Examples of sensitive data archiving are the Greater Sage-grouse (*Centrocercus urophasianus*) and Gunnison Sage-grouse (*C. minimus*), both of which are species of significant conservation concern in North America. Males from both species gather in mating grounds (leks) to attract females and such places are often used for genetic sampling as well as observation by hobbyists. Yet human presence can disturb mating activities. Further, some leks occur on private land, requiring collaboration with landowners. While the genetic data from mating ground samples is publicly available, location information is either generalized (Zimmerman et al. 2019), or an averaged location for a group of mating grounds given (Row et al. 2018, Cross et al. 2018), or only on request (Oyler-McCance et al. 2022). This data masking step prevents disturbance increase and supports conservation.

In another example, the North American butternut (*Juglans cinerea L.*) is a species of conservation concern (IUCN Endangered), but is valuable for timber and traditional medicine (Pike et al. 2021). Sharing location data could lead to harmful timber harvest, thus population coordinates have sometimes been published with a random geographic offset (e.g. such as 10 km; Hoban et al. 2010). This simultaneously protects the species and allows for most genetic and geographic data repurposing, without the need for data access requests. Researchers also took care to remove location names (e.g. the name of a creek, landowner, or nearest town) to prevent location inference (Hoban et al. 2012).

Why should we improve genetic data archiving practices in ecology and evolution?

The FAIR guiding principles are the foundation for good, transparent, and reproducible science. A straightforward demonstration of this is where open data have been used to identify scientific misconduct, some of which impeded evolutionary understanding (e.g. Kozlov 2022). Datasets are often ultimately financed by taxpayers, making public releases an ethical - often even a legal - obligation to ensure the full value of data is obtained. Collectively, data cost many hundreds of millions of dollars to produce, which without archive enrichment, will have to be unnecessarily re-spent to generate the data anew (Crandall et al. 2023). Furthermore, due to the rapid pace of biodiversity loss (e.g. Ceballos et al. 2015), which include genetic diversity decline (Leigh et al. 2019), local extinctions may make regeneration of data impossible, rendering the data irreplaceable and priceless. Existing genetic data also represent an invaluable baseline against which to compare future measurements (i.e. for monitoring of genetic diversity, Jensen and Leigh 2023).

Further arguments for data archiving involve benefit sharing and the rights of local communities and local scientists to access data generated from specimens within their country or region. Thus, the CARE principles (Collective benefits, Authority to control, Responsibility, and Ethics; Carroll et al. 2020) could be considered in data generation, archiving, and repurposing of any genetic and genomic data. Emerging benefit-sharing requirements, such as those put forth in the Nagoya Protocol and being developed by the CBD (e.g. Digital Sequence Information or “DSI”; Scholz et al. 2022), are becoming a legal requirement (Marden et al. 2021). This is particularly pertinent to ecology and evolution where researchers often work internationally (e.g. Bhaumik 2023; Miller et al. 2023).

Researchers themselves can benefit professionally from publicly archiving data. Open datasets can enhance the scholarly recognition of individual research efforts, because data releases with DOI identifiers and data papers can be cited (e.g. MacroPopGen, Lawrence et al. 2019). The increasing popularity of data papers, journals publishing data releases (e.g. Chavan and Penev 2011), and meta-data papers (Raciti et al. 2018) is an early sign that accurate data archiving can benefit individual scientists and the community. Researchers could also benefit from the advancement of their field; synthesis is a powerful tool that has successfully tested pressing big-picture questions in ecology and evolution (Halpern et al. 2020).

Best practices for FAIR genetic and genomic data archiving

The most widely used and available genetic data types in molecular ecology and evolution are: 1) barcoding/gene sequences (e.g. mitochondrial cytochrome oxidase, the major histocompatibility complex), 2) microsatellite genotypes, and 3) genomic read data (i.e. unaligned high throughput sequences and SNPs). These come in a constellation of software-specific formats (Lischer and Excoffier 2011; Adamack & Gruber 2014) and, due to lack of standardization, open genetic data repositories contain most of these formats.

While there are several tools to convert between file formats (GUI-tools: PGDspider, Lischer and Excoffier 2011; Formatomatic, Manoukis 2007; command line: vcftools, plink, R packages: ‘adegenet’, Jombart et al. 2008), conversions are time-consuming and often need to be customized for each dataset. Understanding and working with each file format also requires specialist knowledge. Consequently, the lack of a standard archived format limits FAIR data 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 standardised on archives as FASTA files, and we recommend maintaining this approach. However, many gene sequences lack essential metadata to allow their reuse. It is important that authors archiving gene sequences include the minimum metadata needed to interpret their data (Box 4) otherwise, archives are impossible to reuse (e.g. non-georeferenced sequences in GenBank; Gratton et al. 2017).

For microsatellite data, based on its persistent popularity and flexibility, we recommend that it is archived in a “STRUCTURE” input file format (Pritchard et al. 2000; Lischer and Excoffier 2011). STRUCTURE input files also have the advantages that they can handle both haploid and diploid genotype data and have an intuitive and simple format that is conducive to editing in R (R Core Team, 2023) or common spreadsheet software without generating formatting errors. Files can be saved as a comma- (.csv) or tab-delimited text file (.txt) with missing alleles clearly coded as NA or “-9”. This file format can house (minimal) metadata (geographical coordinates, populations, sample name, phenotypes though it is essential these match with those used in published papers), 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 format; as both are accepted by major conversion tools like PGDSpider (Lischer and Excoffier 2011), either is suitable for archiving. However, we recommend use of the single line format to maximize similarity with VCF files.

Genomic data is often mandated to be publicly archived as raw read data on INSDC servers (“INSDC” International Nucleotide Sequence Database Collaboration) (Cochrane et al. 2016), or as aligned BAM files for model organisms (Li et al. 2009). What constitutes “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) (Mallick et al. 2023). 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. Ideally, sequencing read data should be archived as demultiplexed read files. A bioinformatic pipeline can also be challenging to reproduce because there are chronic issues surrounding open code archiving that make it hard to know exactly what parameters were applied, tool versions used, or even to have access to custom scripts (further detailed in: Gomes et al. 2022; Jenkins et al. 2023). Even if a pipeline is accessible, version changes of reference genomes or software programs quickly make reproducing a pipeline impossible. Thus, archiving FAIR genomic genotype files in addition to demultiplexed sequencing reads would greatly improve the Open Data compliance and reusability of genomic data.

Processed VCF files containing genotype calls (or genotype likelihoods) are standard for genomic analyses and could be archived in parallel with raw read files (though notably this is not possible on INSDC). Though 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 SNP or genotype likelihood calls. Specifically, unfiltered genotypes pruned only for basic errors (e.g. technical faults, known contaminated samples), with headers retained to allow for easy assessment of the bioinformatic steps applied. Notably, archiving genomic genotype files could allow non-bioinformatic wildlife managers to repurpose genomic data for analyses and enable researchers without High-Performance Computing access to work with genomic-derived data. Furthermore, this would limit the non-negligible energy, storage and ultimately emissions costs associated with reanalysing genomic data (Grealey et al. 2022).

Box 4: Archiving metadata

Metadata describing where, when, how and by whom genotype or sequence data was created are invaluable for making genetic data FAIR. There are currently two genomic metadata standards: the Darwin Core standard for biodiversity data (Wieczorek et al. 2012) and the Minimum Information about any(x) Sequence (MIxS) standard (Field et al. 2008). Both standards have cross-mapped terms that overlap (summarized below). What metadata to archive will vary by sample type, project goals, and what researchers deem important (Figure 2). At a minimum, we suggest that authors provide the required (solid lines) and recommended (dotted lines) categories represented in Figure 2 to ensure valuable context. To report metadata not covered here, we also recommend using Darwin Core or MixS standards terms to guarantee FAIRness. Note: Darwin Core contains terms that can handle geologic context of special samples, such as ancient DNA, where metadata related to sampling events generally does not reflect the conditions of the sampled individual before death. As discussed in Box 3, sensitive data should be withheld to ensure it is protected. This can be denoted with the terms “informationWithheld”, “dataGeneralizations” or “coordinateUncertaintyInMeters”.

The 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. They can thus be used to quickly join the two data types. 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 on file repositories unless it is archived with the genotype data directly. 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, as described above. To aid automated retrieval, authors should avoid using symbols, special characters, and/or colour-based cell codes.

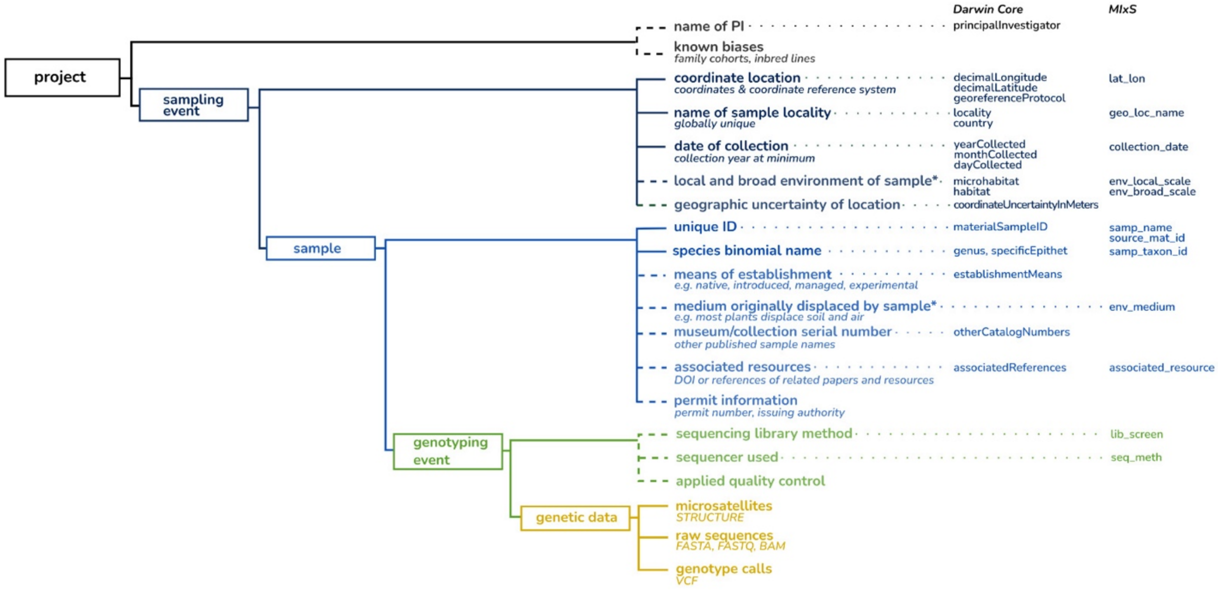

*Figure 2: Metadata can be viewed in a hierarchical manner based on how they were created. The required (solid line) and recommended (dashed line) metadata terms that would improve publicly archived genetic and genomic data reuse potential. Terms denoted with * should use controlled vocabulary from the Environment ontology ("ENVO", Buttigieg et al. 2013). Note: these 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.*

Context is key: missing metadata renders most archived data useless

Metadata are a crucial aspect of ensuring genetic data adhere to the FAIR principles (Wilkinson et al. 2016) because the context they provide vastly increases their potential reuses. Genetic metadata record the material and processes that were used for the creation of the genetic data, and can be viewed in a hierarchical manner based on how they were created: 1) sampling events, which include temporal, spatial, and methodological metadata (e.g. year, coordinates). Each sampling event can give rise to many 2) biological samples, each of which have taxonomic, biological, and methodological metadata (e.g. genus, environmental medium, sample preservative). Samples may have many 3) tissues, which might have different biological attributes (e.g. different expression of genes) and may be 4) subject to several different genotyping protocols, which have methodological metadata (e.g. library protocol, Field et al. 2008, Deck et al. 2017, Hassenrück et al. 2021, Crandall et al. 2023).

It is standard to include taxonomic metadata (species and genus) in archives, but this is often not sufficient for reuse. The minimum required and recommended metadata are shown in Box 4, without which archived data are often functionally useless and could lead to incorrect inferences. Should key metadata be unavailable to authors we suggest they provide as much information as possible to increase the chance that data will be found and re-used profitably. Currently, 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. These can take a significant amount of time to access, reformat, or convert for reuse (Crandall et al. 2023). As a result, great efforts have been made to retrospectively georeference existing genetic data to improve their reusability (e.g. Miraldo et al. 2016; Crandall et al. 2023), but this often relies on inference (e.g. inferring coordinates from place names) leaving significant room for error or lost resolution. Thus, we would encourage authors to enrich the public metadata of their data archives and ensure that the metadata included in publications is also present in the data archive.

Special considerations when working with important species or indigenous communities

CARE principles (described above, Carroll et al. 2020) are important considerations for data archiving when data are from a culturally important species or indigenous community territory. What steps researchers need to follow will be situation-specific and should be developed in conjunction with interested parties. To ensure these requirements are upheld, data-generating authors should include specific benefit-sharing statements in the publications themselves and in the data archives. This should contain contextual metadata within the statement, 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 should be stored within each sample's metadata. When reusing such data, researchers should also follow the ethical repurposing guidelines above and discuss planned analyses with interested parties. Attribution and citation of the original datasets in resulting manuscripts, and dissemination of results to the communities involved could help ensure that cultural authority and sovereignty over such data are recognized (e.g. McCartney et al. 2023), and that data are not reused inappropriately.

Which data repository should researchers use?

Centralized infrastructure already exists for genetic sequence data storage (INSDC) that makes finding and accessing data straightforward. Such databases are now impossible to replace and should continue to be used. However, these databases are designed to store only sequence data (e.g. raw reads, gene sequences, whole genomes or transcriptomes) and their metadata (e.g. BioSamples). Genotype data are not stored in sequence databases and there is limited established guidance or storage conventions for them.

Currently, genotype data are often stored in multi-purpose Open Data repositories (DRYAD, Zenodo, and increasingly FigShare). However, genetic data can quickly get lost among many of the other data types archived in multi-purpose repositories, where researchers can find everything from non-peer-reviewed ecological survey data (e.g. Shaikh 2014) to violent crime statistics (e.g. Gonzales 2010). Local rules and repository cost barriers (i.e. archiving fees) make it currently impossible to advocate for a single existing database for all genotype data. We note that there are cross-database interoperable search platforms that enable users to search multiple data repositories at once (for example DataONE). However, this functionality is not guaranteed and database linking has failed in the past (Chloé Schmidt *pers. comm.*). There is a need for a free inter-government supported public database specifically for archiving genotype data (microsatellites and SNPs).

In lieu of a dedicated repository, researchers can take a few key steps to ensure the findability of genotype data. At a minimum, researchers should ensure the database links their data to their publication. The archiving researcher should include the key metadata fields in Box 4 in the database description and/or the title to aid findability. We also encourage including marker type as a keyword (e.g. “microsatellite” or “SNP”) and key geographical descriptors (e.g. “Kruger National Park”) to make searching for data more straightforward. Researchers could also consider linking genotypic data to “metadatabases” that keep sample-level metadata in structured, searchable format, enabling users to track samples from the point of collection. These tools can also facilitate upload to the SRA, thereby making the data much more FAIR through structured queries of the metadatabases or INSDC (INSDC BioProjects and BioSamples, Barrett et al. 2021; Genomic Observatories MetaDatabase (GEOME), Deck et al. 2017; Collaborative Open Plant Omics (COPO), Shaw et al. 2020).

There are also important database features that researchers should seek out when archiving their genotype data. Researchers should look for a free (or affordable) FAIR compliant Trustworthy Digital Repository (Wilkinson et al. 2016) because they capture key accessibility criteria by definition (compliant repositories are listed on the Registry of Research Data Repositories; Pampel et al. 2013). Institution-specific databases (i.e. university or research institute level) are less desirable because they rarely produce DOIs for data citation, are not easily accessible (e.g. require a password), and might suddenly become depreciated or unsupported.

The researcher community could also request new features within existing databases that facilitate genotype data accessibility. The Web of Science’s “associated data” link is a notable advance (Web of Science, 2018), as is the soon mandatory metadata (sample location, collection date) for BioSample packages (DDBJ 2023). A desirable additional feature, which would benefit multiple scientific disciplines, is an automatic identifier for retracted data and/or data associated with retracted articles. As of writing, datasets found to be fraudulent from retracted papers remain on servers with no clear notification that the publications was retracted (e.g. Dryad, Costa-Pereira and Pruitt 2019). Similarly data found to be erroneous remains on sequence databases (e.g. Genbank, compiled by van den Burg and Vieites 2022) posing a huge challenge to researchers that automate data collection for repurposing. Researchers could benefit from an easy and anonymous way to notify data editors or database curators if they encounter incomplete non-FAIR compliant archives, who should then be responsible for formally rectifying in a harmonious manner.

The role of funding bodies and universities in increasing data archiving

University libraries, funding bodies, scientific journals, and data repositories could also take on a greater responsibility to ensure FAIR data archiving. Funding bodies can facilitate data archiving by continuing to mandate Open Data (e.g. National Institute of Health, 2020; European Commission 2016), which have undoubtedly driven an increase in accessibility. However, funding bodies need to support researchers by reviewing or assisting in data management plans (e.g. UKRI 2013), reviewing archived data accessibility and integrity, paying data archiving fees, and offering data archiving educational resources or training. We would specifically encourage funding bodies to ensure future projects budget time for data archiving in their project plan and reward researchers with an established history of Open Data in any field through positively valuing data products or dataset citations.

Universities and other science organizations could play a greater role in Open Data through hiring data “stewards” or “librarians” familiar with ecological and evolutionary genetic data. The tasks of data stewards include supporting researchers writing data management plans, identifying suitable databases for archiving, and ensuring dataset longevity through file format conversion (Peng et al. 2016). Notably, data stewards may not be able to archive data directly due to lack of resources and the specialist knowledge required.

Science organizations and funding bodies can further foster Open Data by offering data management education (e.g. short courses and training) for both students and career scientists of all disciplines (e.g. Toelch and Ostwald 2018). Few ecology, evolution, or life sciences researchers have received any formal introduction into the importance of Open Data nor in correct data archive practices. Scientific departments could also reward researchers who archive their data or whose data have been reused. Datasets and their reuse (number of views/downloads) can be credited as scientific products on a researcher’s *Curriculum Vitae or Research Record* and used during hiring, promotion and tenure decisions.

Scientific journal and reviewer roles to ensure Open Data compliance

Scientific journals can facilitate Open Data by ensuring data are archived on a FAIR compliant Trustworthy Digital Repository before final acceptance of an article (Jenkins et al. 2023, Wilkinson et al. 2016). Journals could also check that data links are activated and that authors have not added reuse clauses or unjustifiable embargos that impede the repurposing of Open Data (see Thrall et al. 2023).

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 data accessible *upon submission* is needed, particularly at the resubmission stage when papers are close to acceptance (Thrall et al. 2023). Alternatively, journals could make the final acceptance dependent on data accessibility and FAIR data compliance. Scientists concerned with data being accessed prior to publication should note that several databases offer non-public shareable links that can prevent reuse before publication acceptance.

Journals also have a role to play in improving essential metadata accessibility, which can be easily implemented by having a table of standardized terms that authors must fill out and/or ensure sample names are meaningful (see Box 4). While sample information can be included in supplementary material, versioning issues may arise if metadata are in multiple places. Thus, data editors for journals could ensure all data derived from the same sample are linked (i.e. same name) and key differences (e.g. resequencing with a new technology) highlighted. Importantly, as stated above, we recommend that journals also inform data repositories if papers have been retracted so that the dataset can be demarcated as such (though not removed).

While peer reviewers should not be tasked with ensuring data archiving, they are in a key position to help advance Open Data through a small number of tasks. Initially with novel genetic datasets, authors often need to check the integrity of the data (e.g. checking for contamination) and reviewers could consider asking for evidence of this (e.g. mapping statistics or quality, van den Burg and Vieites 2022). Reviewers could also check archived data files to ensure they are not corrupt, contain the correct number of markers or loci, and contain basic metadata. For repurposed data papers, reviewers can ensure that datasets are cited correctly (see Cousjin et al. 2018). Reviewers could also examine author contribution statements and report to the editor cases where the data-generating authors have not received equal accreditation (Nature 2022).

Rectifying past mistakes - enriching archived data

An important step many of us can take to advance Open Data, is to improve metadata archives or archiving inaccessible genetic data. For example, GEOME has successfully run remote datathons to enrich genetic metadata archives (Crandall et al. 2023). We encourage authors to similarly enrich metadata in old data archives, to archive inaccessible genetic datasets, and/or expand on what was archived (e.g. archive all mtDNA haplotypes rather than only unique haplotypes). Although old genetic marker types may be regarded as being of low value to some authors, when combined with other datasets (as in macrogenetics, Leigh et al., 2021), they can be highly informative and can even provide baselines for important biodiversity protection assessments (e.g. Figuerola-Ferrando et al. 2023; Schmidt et al. 2022).

Data enrichment initiatives could be run at the Department (similar to MoveBank, Max Planck Institute of Animal Behavior 2023), University library, or country level (e.g. GenDiB and CIEE Living Data), with support from students or technicians to upload data. Such retroactive data archives could even be collaboratively published as a "resource" paper (similar to those in Box 1). These datasets could then support mandated CBD reporting (Hoban et al. 2020), inform local conservation (e.g. Beninde et al. 2022), and identify interesting scientific opportunities (e.g. resampling populations after extreme events, Jensen & Leigh 2022).

Perspectives

Box 5: Five take-home messages to improve genetic data archives

1) Archiving genetic and genomic data in standardized file formats will facilitate reuse (i.e. microsatellites in STRUCTURE; sequences or barcodes in FASTA; SNPs in VCF; Genotype likelihoods in VCF; raw genomic data as demultiplexed FASTQ files).
2) 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*). Include additional contextual metadata when needed to interpret data correctly.
3) Carefully archive data from sensitive species and those affected by the CARE principles to ensure that metadata do not endanger the species, their habitats, or landowner relationships.
4) There is no centralized database for genotype data but this data has great value. Use keywords on FAIR compliant databases (e.g. Dryad) to improve data accessibility.
5) To help more colleagues follow the FAIR principles, request both formalized data management support and a higher value of open data from research institutes and journals.

We close on the note that genetic diversity is the most fundamental component of biodiversity (Hoban et al., 2023). Despite underlying all levels of biodiversity, the biogeographic patterns in intra-specific genetic diversity are largely understudied and poorly protected (Leigh et al. 2019; Figuerola-Ferrando et al. 2023). Consequently, perhaps the most exciting potential of improved archiving is that we can reach research scales beyond what any single research group could achieve. With data spanning such vast spatial and taxonomic scales, open genetic data is pivotal to whole new areas of research and conservation that would have previously been unimaginable due to logistic, cost, or expertise issues. Similar to data collected as part of long-term ecological monitoring programs, publicly archived genetic data is likely to only become more valuable and versatile as it accumulates. The potential of public genetic data is pertinent and timely due to the recently signed United Nations Kunming-Montreal Global Biodiversity 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 decade (Hoban et al. 2023b). Better archiving practices are likely to be central to meet these targets. Although new archiving infrastructure would undoubtedly enhance our ability to do this research, we feel the steps we propose (see Box 5) are achievable with the currently available resources and in the rapid timescale needed.

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 Box 1 is available for reviewers and will be accessible upon publication.

Conflicts of interest:

The authors declare no conflicts of interest.

References

Adamack A.T., Gruber B. (2014) POP GEN REPORT: Simplifying Basic Population Genetic Analyses in R. Edited by Stephane Dray. *Methods in Ecology and Evolution* 5(4): 384–87. https://doi.org/10.1111/2041-210X.12158.

Baack S. (2015) Datafication and Empowerment: How the Open Data Movement Re-Articulates Notions of Democracy, Participation, and Journalism. *Big Data & Society* 2(2): 205395171559463. https://doi.org/10.1177/2053951715594634.

Barrett T., Clark K., Gevorgyan R., Gorelenkov V., E. Gribov, Karsch-Mizrachi I., Kimelman M., et al. (2012) BioProject and BioSample Databases at NCBI: Facilitating Capture and Organization of Metadata. *Nucleic Acids Research* 40(1):57–63. https://doi.org/10.1093/nar/gkr1163.

Beninde J., Toffelmier E.M., Andreas A., Nishioka C., Slay M., Soto A., Bueno J.P., et al. (2022) CaliPopGen: A Genetic and Life History Database for the Fauna and Flora of California. *Scientific Data* 9(1):380. https://doi.org/10.1038/s41597-022-01479-z.

Blanchet S., Prunier J.G., De Kort H. (2017) Time to Go Bigger: Emerging Patterns in Macrogenetics. *Trends in Genetics* 33(9)579–80. https://doi.org/10.1016/j.tig.2017.06.007.

CBD (2022) Decision Adopted By The Conference Of The Parties To The Convention On Biological Diversity. https://www.cbd.int/doc/decisions/cop-15/cop-15-dec-05-en.pdf

Chapman AD (2020) Current Best Practices for Generalizing Sensitive Species Occurrence Data. Copenhagen: GBIF Secretariat. https://doi.org/10.15468/doc-5jp4-5g10.

Chavan V., Penev L. (2011) The Data Paper: A Mechanism to Incentivize Data Publishing in Biodiversity Science. *BMC Bioinformatics* 12(S15): S2. https://doi.org/10.1186/1471-2105-12-S15-S2.

Clarke K. C. A. (2016) Multiscale Masking Method for Point Geographic Data. *International Journal of Geographical Information Science* 30(2): 300–315. https://doi.org/10.1080/13658816.2015.1085540.

Cochrane G., Cook C.E., Birney E. (2012) The Future of DNA Sequence Archiving. *GigaScience* 1(1): 2047–217X–1–2. https://doi.org/10.1186/2047-217X-1-2.

Cochrane G., Karsch-Mizrachi I., Takagi T., International Nucleotide Sequence Database Collaboration. (2016) The International Nucleotide Sequence Database Collaboration, Nucleic Acids Research, 44(1) 48–D50, https://doi.org/10.1093/nar/gkv1323

Costa-Pereira R., Pruitt J. (2020) Data from: Behavior, morphology, and microhabitat use: what drives individual niche variation? Dryad, Dataset, https://doi.org/10.5061/dryad.bd26mq0

Crandall E.D., Toczydlowski R.H., Liggins L., Holmes A.E., Ghoojaei M., Gaither M.R, Wham B.E., et al. (2023) Importance of Timely Metadata Curation to the Global Surveillance of Genetic Diversity. *Conservation Biology* 37(4): e14061. https://doi.org/10.1111/cobi.14061.

Cross T.B., Schwartz M.K., Naugle D.E., Fedy B.C., Row J.R., Oyler-McCance S.J. (2018) The Genetic Network of Greater Sage-grouse: Range-wide Identification of Keystone Hubs of Connectivity. *Ecology and Evolution* 8(11): 5394–5412. https://doi.org/10.1002/ece3.4056.

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

Deck J., Gaither M.R., Ewing R., Bird C.E., Davies N., Meyer C., Riginos C., Toonen R.J., Crandall E.D. (2017) The Genomic Observatories Metadatabase (GeOMe): A New Repository for Field and Sampling Event Metadata Associated with Genetic Samples'. *PLOS Biology* 15(8): e2002925. https://doi.org/10.1371/journal.pbio.2002925.

European Commission (2016) Guidelines on FAIR data management in Horizon 2020 https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Espindola S., Vázquez-Domínguez E., Nakamura M., Osorio-Olvera L., Martínez-Meyer E., Myers E.A., Overcast I., Reid B.N., Burbrink F.T. (2022) Complex Genetic Patterns and Distribution Limits Mediated by Native Congeners of the Worldwide Invasive Red-eared Slider Turtle. *Molecular Ecology* 31(6): 1766–82. https://doi.org/10.1111/mec.16356.

Fairbairn, D.J. (2011) The Advent Of Mandatory Data Archiving. *Evolution* 65(1): 1–2. https://doi.org/10.1111/j.1558-5646.2010.01182.x.

Field D., Amaral-Zettler L., Cochrane G., Cole J.R., Dawyndt P., Garrity G.M., Gilbert J., et al. (2011) The Genomic Standards Consortium. *PLoS Biology* 9(6): e1001088. https://doi.org/10.1371/journal.pbio.1001088.

Frank R. D., Kriesberg A., Yakel E., Faniel I. M. (2016) Looting Hoards of Gold and Poaching Spotted Owls: Data Confidentiality among Archaeologists & Zoologists. *Proceedings of the*

Association for Information Science and Technology 52(1): 1–10. https://doi.org/10.1002/pra2.2015.145052010037.

Gabelica M., Bojčić R., Puljak L. (2022) Many Researchers Were Not Compliant with Their Published Data Sharing Statement: A Mixed-Methods Study. *Journal of Clinical Epidemiology* 150: 33–41. https://doi.org/10.1016/j.jclinepi.2022.05.019.

Gomes D.G.E., Pottier P., Crystal-Ornelas R., Hudgins E.J., Foroughirad V., Sánchez-Reyes L.L., Turba R., et al. (2022) Why don't we share data and code? Perceived barriers and benefits to public archiving practices. *Proceedings of the Royal Society B*. 289(1987): 2022111 https://doi.org/10.1098/rspb.2022.1113

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

Gratton P., Marta S., Bocksberger G., Winter M., Trucchi E., Kühl H. (2017) A World of Sequences: Can We Use Georeferenced Nucleotide Databases for a Robust Automated Phylogeography? *Journal of Biogeography* 44(2): 475–86. https://doi.org/10.1111/jbi.12786.

Grealey J., Lannelongue L., Saw W-Y, Marten J., Méric G., Ruiz-Carmona S., Inouye M. (2022) The Carbon Footprint of Bioinformatics. *Molecular Biology and Evolution* 39(3): msac034. https://doi.org/10.1093/molbev/msac034.

Günther T., Coop G. (2013) Robust Identification of Local Adaptation from Allele Frequencies. *Genetics* 195(1): 205–20. https://doi.org/10.1534/genetics.113.152462.

Hassenrück C., Poprick T., Helfer V., Molari M., Meyer R., Kostadinov I. (2021) FAIR Enough? A Perspective on the Status of Nucleotide Sequence Data and Metadata on Public Archives, bioRxiv 2021.09.23.461561; doi: https://doi.org/10.1101/2021.09.23.461561

Hoban S.M., Borkowski D.S., Bros S.L., McCleary T.S., Thompson L.M., McLachlan J.S., Pereira M.A., Schlarbaum S.E., Romero-Severson J. (2010) Range-Wide Distribution of Genetic Diversity in the North American Tree *Juglans cinerea*: A Product of Range Shifts, Not Ecological Marginality or Recent Population Decline. *Molecular Ecology* 19(22):4876–91. https://doi.org/10.1111/j.1365-294X.2010.04834.x.

Hoban S., McCleary T.S., Schlarbaum S.E., Anagnostakis S.L, Romero-Severson J. (2012). Human-Impacted Landscapes Facilitate Hybridization between a Native and an Introduced

Tree. *Evolutionary Applications* 5(7): 720–31. https://doi.org/10.1111/j.1752-4571.2012.00250.x.

Hoban S., Bruford M., D'Urban Jackson J., Lopes-Fernandes M., Heuertz M., Hohenlohe P. A, Paz-Vinas I., et al. (2020) Genetic Diversity Targets and Indicators in the CBD Post-2020 Global Biodiversity Framework Must Be Improved. *Biological Conservation* 248: 108654. https://doi.org/10.1016/j.biocon.2020.108654.

Hoban S., da Silva, J.M., Mastretta-Yanes A., Grueber C.E., Heuertz, M., Hunter M.E., Mergeay, J., et al. (2023). Monitoring status and trends in genetic diversity for the Convention on Biological Diversity: an ongoing assessment of genetic indicators in nine countries. *Conservation Letters*, 16, e12953. https://doi.org/10.1111/conl.12953

Hostler T.J. (2023) The Invisible Workload of Open Research. *Journal of Trial and Error* https://doi.org/10.36850/mr5.

Huang X., Hawkins B.A., Lei F., Miller G. L, Favret C., Zhang R., Qiao G. (2012) Willing or Unwilling to Share Primary Biodiversity Data: Results and Implications of an International Survey. *Conservation Letters* 5: 399–406. https://doi.org/10.1111/j.1755-263X.2012.00259.x.

Jenkins G.B., Beckerman A.P., Bellard C., Benítez-López A., Ellison A.M., Foote C.G., Hufton A.L., et al. (2023) Reproducibility in Ecology and Evolution: Minimum Standards for Data and Code. *Ecology and Evolution* 13(5): e9961. https://doi.org/10.1002/ece3.9961.

Jensen E. L., Leigh D. M. (2022) Using Temporal Genomics to Understand Contemporary Climate Change Responses in Wildlife. *Ecology and Evolution* 12:e9340 https://doi.org/10.1002/ece3.9340.

Jombart T. (2008) *Adegenet*: A R Package for the Multivariate Analysis of Genetic Markers. *Bioinformatics* 24(11): 1403–5. https://doi.org/10.1093/bioinformatics/btn129.

Kozlov M. (2022) How A Spider-Biology Scandal Upended Researchers' Lives. *Nature* 608: 658-659 doi: https://doi.org/10.1038/d41586-022-02156-2

Lawrence E. R., Benavente J. N., Matte J.-M., Marin K., Wells Z.R.R., Bernos T. A., Krasteva N., et al. (2019) Geo-Referenced Population-Specific Microsatellite Data across American Continents, the MacroPopGen Database. *Scientific Data* 6(1):14. https://doi.org/10.1038/s41597-019-0024-7.

Leigh D.M., Hendry A.P., Vázquez-Domínguez E., Friesen V.L. (2019) Estimated Six per Cent Loss of Genetic Variation in Wild Populations since the Industrial Revolution. *Evolutionary Applications* 12(8):1505–12. https://doi.org/10.1111/eva.12810.

Leigh D.M., van Rees C.B., Millette K.L., Breed M.F., Schmidt C., Bertola L.D., Hand B.K., et al. (2021) Opportunities and Challenges of Macrogenetic Studies. *Nature Reviews Genetics* 22(12): 791–807. https://doi.org/10.1038/s41576-021-00394-0.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., and 1000 Genome Project Data Processing Subgroup. (2009) The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* 25(16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

Lindenmayer D., Scheele B. (2017) Do Not Publish. *Science* 356(6340): 800–801. https://doi.org/10.1126/science.aan1362.

Lischer H.E.L., Excoffier L., (2012) PGDSpider: An Automated Data Conversion Tool for Connecting Population Genetics and Genomics Programs. *Bioinformatics* 28(2): 298–99. https://doi.org/10.1093/bioinformatics/btr642.

Lowe A.J., Smyth A.K., Atkins K., Avery R., Belbin L., Brown N., Budden A.E., et al. (2017) Publish Openly but Responsibly. *Science* 357(6347): 141–141. https://doi.org/10.1126/science.aao0054.

Mallick S., Micco A., Mah M., Ringbauer H, Lazaridis I., Olalde I., Patterson N, Reich D. (2023) The Allen Ancient DNA Resource (AADR): A Curated Compendium of Ancient Human Genomes'. biorxiv. https://doi.org/10.1101/2023.04.06.535797.

Marden E., Abbott R.J., Austerlitz F., Ortiz-Barrientos D., Baucom R.S., Bongaerts P., Bonin A., et al. (2021) Sharing and Reporting Benefits from Biodiversity Research. *Molecular Ecology* 30(5): 1103–7. https://doi.org/10.1111/mec.15702.

Max Planck Institute of Animal Behavior. 2023. Movebank Attribute Dictionary. British Oceanographic Data Centre, Natural Environment Research Council Vocabulary Server. http://vocab.nerc.ac.uk/collection/MVB/current

Mills J.A., Teplitsky C., Arroyo B., Charmantier A., Becker P. H., Birkhead T.R., Bize P., et al. (2015) Archiving Primary Data: Solutions for Long-Term Studies. *Trends in Ecology & Evolution* 30(10): 581–89. https://doi.org/10.1016/j.tree.2015.07.006.

Miraldo A., Li S., Borregaard M. K., Flórez-Rodríguez A., Gopalakrishnan S., Rizvanovic M., Wang Z. et al. (2016) An Anthropocene Map of Genetic Diversity. *Science* 353(6307): 1532–35. https://doi.org/10.1126/science.aaf4381.

Moore A.J., Mcpeek M.A., Rausher M.D., Rieseberg L., Whitlock M.C. (2010) The Need for Archiving Data in Evolutionary Biology. *Journal of Evolutionary Biology* 23(4): 659–60. https://doi.org/10.1111/j.1420-9101.2010.01937.x.

Nater A., Greminger M.P., Arora N., van Schaik C.P., Goossens B., Singleton I., Verschoor E.J., et al. (2015) Reconstructing the demographic history of orang-utans using Approximate Bayesian Computation. *Molecular Ecology* 24(2): 310-327. https://doi.org/10.1111/mec.13027

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

Nature (2022) Time to recognize authorship of open data. *Nature* 604:8 doi: https://doi.org/10.1038/d41586-022-00921-x

Oyler-McCance S.J., Cross T.B., Row J.R., Schwartz M.K., Naugle D.E., Fike J.A., Winiarski K., Fedy B.C. (2022) New Strategies for Characterizing Genetic Structure in Wide-Ranging, Continuously Distributed Species: A Greater Sage-Grouse Case Study. *PLoS ONE* 17(9): e0274189. https://doi.org/10.1371/journal.pone.0274189.

Pampel H., Vierkant P., Scholze F., Bertelmann R., Kindling M., Klump, J., Goebelbecker H.-J., et al., (2013) Making Research Data Repositories Visible: The Re3data.Org Registry *PLoS ONE* 8,(11): e78080. https://doi.org/10.1371/journal.pone.0078080.

Paz-Vinas I., Jensen E.L., Bertola L.D., Breed M.F., Hand B.K., Hunter M.E., Kershaw F., et al. (2021) Macrogenetic Studies Must Not Ignore Limitations of Genetic Markers and Scale. *Ecology Letters* 24(6): 1282–84. https://doi.org/10.1111/ele.13732.

Peng G., Ritchey N.A., Casey K.S., Kearns E.J., Prevette J.L., Saunders D., Jones P., Maycock T., Ansari S. (2016) Scientific Stewardship in the Open Data and Big Data Era Roles and Responsibilities of Stewards and Other Major Product Stakeholders. *D-Lib Magazine* 22, no. 5/6. https://doi.org/10.1045/may2016-peng.

Pike V.L., Cornwallis C.K., Griffin A.S. (2021) Why Don't All Animals Avoid Inbreeding? *Proceedings of the Royal Society B: Biological Sciences* 288(1956): 20211045. https://doi.org/10.1098/rspb.2021.1045.

Piwowar H.A., Vision T.J., Whitlock M.C. (2011) Data archiving is a good investment. *Nature.* 473(7347): 285. doi: 10.1038/473285a..

Pritchard J.K., Stephens M., Donnelly P. (2000) Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155(2): 945–59. https://doi.org/10.1093/genetics/155.2.945.

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Raciti D., Yook K., Harris T.W., Schedl T., Sternberg P.W. (2018) *Micropublication* : Incentivizing Community Curation and Placing Unpublished Data into the Public Domain. *Database* 2018: bay013 https://doi.org/10.1093/database/bay013.

Rieseberg L., Vines T., Kane N. (2010) Editorial and Retrospective 2010. *Molecular Ecology* 19(1): 1–22. https://doi.org/10.1111/j.1365-294X.2009.04450.x.

Roche D.G., Kruuk L.E.B., Lanfear R., Binning S.A. (2015) Public Data Archiving in Ecology and Evolution: How Well Are We Doing? *PLOS Biology* 13(11): e1002295. https://doi.org/10.1371/journal.pbio.1002295.

Roche D.G., Lanfear R., Binning S.A., Haff T.M., Schwanz L.E., Cain K.E., Kokko H., Jennions M.D., Kruuk L.E.B. (2014) Troubleshooting Public Data Archiving: Suggestions to Increase Participation. *PLoS Biology* 12(1): e1001779. https://doi.org/10.1371/journal.pbio.1001779.

Row J.R., Doherty K.E., Cross T.B., Schwartz M.K., Oyler-McCance S.J., Naugle D.E., Knick S.T., Fedy B.C. (2018) Quantifying Functional Connectivity: The Role of Breeding Habitat, Abundance, and Landscape Features on Range-wide Gene Flow in Sage-grouse. *Evolutionary Applications* 11(8): 1305–21. https://doi.org/10.1111/eva.12627.

Ruppert K.M., Kline R.J., Rahman S. (2019) Past, Present, and Future Perspectives of Environmental DNA (EDNA) Metabarcoding: A Systematic Review in Methods, Monitoring, and Applications of Global EDNA. *Global Ecology and Conservation* 17: e00547. https://doi.org/10.1016/j.gecco.2019.e00547.

Scholz A.H., Freitag J., Lyal C.H.C., Sara R., Cepeda M.L., Cancio I., Sett S., et al. (2022) Multilateral Benefit-Sharing from Digital Sequence Information Will Support Both Science and Biodiversity Conservation. *Nature Communications* 13(1): 1086. https://doi.org/10.1038/s41467-022-28594-0.

Schmidt C., Domaratzki M., Kinnunen R.P., Bowman J., Garroway C.J. (2022) Continent-Wide Effects of Urbanization on Bird and Mammal Genetic *Diversity. Proceedings of the Royal Society B* 287(1920): 20192497. http://dx.doi.org/10.1098/rspb.2019.2497

Schmidt C., Garroway C.J. (2022) Systemic Racism Alters Wildlife Genetic Diversity. *Proceedings of the National Academy of Sciences* 119(43): e2102860119. https://doi.org/10.1073/pnas.2102860119.

Schmidt C., Garroway C. J. (2021) The Population Genetics of Urban and Rural Amphibians in North America. *Molecular Ecology* 30(16): 3918–29. https://doi.org/10.1111/mec.16005.

Schmidt, C; Hoban, S; Jetz, W. (2023) Conservation macrogenetics: harnessing genetic data to meet conservation commitments. Trends in Genetics. In press. https://doi.org/10.1016/j.tig.2023.08.002

Schumacher, E., Brown, A., Williams, M., Romero-Severson, J., Beardmore, T. & Hoban, S. (2022). Range shifts in butternut, a rare, endangered tree, in response to past climate and modern conditions. *Journal of Biogeography*, 49(5): 866–878. https://doi.org/10.1111/jbi.14350

Schriml L.M., Chuvochina M., Davies N., Eloe-Fadrosh E.A., Finn R.D., Hugenholtz P., Hunter C. I., et al. (2020) COVID-19 Pandemic Reveals the Peril of Ignoring Metadata Standards. *Scientific Data* 7(1): 188. https://doi.org/10.1038/s41597-020-0524-5.

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

Shaw, F., Etuk A., Minotto A., Gonzalez-Beltran A., Johnson D., Rocca-Serra P., Laporte M.-A., et al. (2020) COPO: A Metadata Platform for Brokering FAIR Data in the Life Sciences. *F1000Research* 9: 495. https://doi.org/10.12688/f1000research.23889.1.

Tedersoo, L., Küngas R., Oras E., Köster K., Eenmaa H., Leijen Ä., Pedaste M., et al. (2021) Data Sharing Practices and Data Availability upon Request Differ across Scientific Disciplines. *Scientific Data* 8(1): 192. https://doi.org/10.1038/s41597-021-00981-0.

Thrall P.H., Chase J., Drake J., Espuno N., Hello S., Ezenwa V., Han B., Mori A., Muller-Landau H. (2023) From raw data to publication: Introducing data editing at Ecology Letters. *Ecology Letters* 26: 829-830. https://doi.org/10.1111/ele.14210

Toczydlowski R. H., Liggins L., Gaither M. R., Anderson T. J., Barton R. L., Berg J. T., Beskid S. G., et al. (2021) Poor Data Stewardship Will Hinder Global Genetic Diversity Surveillance. *Proceedings of the National Academy of Sciences* 118(34): e2107934118. https://doi.org/10.1073/pnas.2107934118.

Toelch U., Ostwald D. (2018) Digital open science—Teaching digital tools for reproducible and transparent research. *PLoS Biology* 16(7): e2006022. https://doi.org/10.1371/journal.

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

van den Burg M. P., Vieites D. R. (2023) Bird Genetic Databases Need Improved Curation and Error Reporting to NCBI. *Ibis* 165(2): 472–81. https://doi.org/10.1111/ibi.13143.

Vines, T.H., Albert A.Y.K., Andrew R.L., Débarre F., Bock D.G., Franklin M.T., Gilbert K.J., Moore J.-S., Renaut S., Rennison D.J. (2014) The Availability of Research Data Declines Rapidly with Article Age. *Current Biology* 24(1): 94–97. https://doi.org/10.1016/j.cub.2013.11.014.

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

Whitlock M.C. Data Archiving in Ecology and Evolution: Best Practices. *Trends in Ecology & Evolution* 26(2): 61–65. https://doi.org/10.1016/j.tree.2010.11.006.

Whitlock M.C., Bronstein J.L., Bruna E.M., Ellison A.M., Fox C.W., McPeek M.A., Moore A.J., et al. (2015) A Balanced Data Archiving Policy for Long-Term Studies. *Trends in Ecology & Evolution* 31(2): 84–85. https://doi.org/10.1016/j.tree.2015.12.001.

Wicherts J.M., Bakker M., Molenaar D. (2011) Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. *PLoS ONE* 6(11): e26828. https://doi.org/10.1371/journal.pone.0026828.

Wilkinson M.D., Dumontier M.I, Aalbersberg I.J., Appleton G., Axton M., Baak A., Blomberg N., et al. (2016) The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3(1): 160018. https://doi.org/10.1038/sdata.2016.18.

Zimmerman S.J., Aldridge C.L., Oh K.P., Cornman R.S., Oyler-McCance S.J. (2019) Signatures of Adaptive Divergence among Populations of an Avian Species of Conservation Concern. *Evolutionary Applications* 12(8): 1661–77. https://doi.org/10.1111/eva.12825.

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