Received: 9 November 2022 | Revised: 27 December 2022 | Accepted: 7 January 2023

DOI: 10.1111/cobi.14061

Conservation Biology

CONTRIBUTED PAPERS

Importance of timely metadata curation to the global surveillance of genetic diversity

Eric D. Crandall[1] | **Rachel H. Toczydlowski**[2] | **Libby Liggins**[3] | **Ann E. Holmes**[4] | **Maryam Ghoojaei**[5] | **Michelle R. Gaither**[5] | **Briana E. Wham**[6] | **Andrea L. Pritt**[7] | **Cory Noble**[3] | **Tanner J. Anderson**[8] | **Randi L. Barton**[9,10] | **Justin T. Berg**[11] | **Sofia G. Beskid**[12] | **Alonso Delgado**[13] | **Emily Farrell**[5] | **Nan Himmelsbach**[14] | **Samantha R. Queeno**[8] | **Thienthanh Trinh**[5] | **Courtney Weyand**[15] | **Andrew Bentley**[16] | **John Deck**[17] | **Cynthia Riginos**[18] | **Gideon S. Bradburd**[2] | **Robert J. Toonen**[19]

[1]Department of Biology, Pennsylvania State University, University Park, Pennsylvania, USA

[2]Ecology, Evolution, and Behavior Program, Department of Integrative Biology, Michigan State University, East Lansing, Michigan, USA

[3]School of Natural Sciences, Massey University, Auckland, New Zealand

[4]Department of Animal Science, University of California, Davis, Davis, California, USA

[5]Department of Biology, University of Central Florida, Orlando, Florida, USA

[6]Department of Research Informatics and Publishing, The Pennsylvania State University Libraries, Pennsylvania State University, University Park, Pennsylvania, USA

[7]Madlyn L. Hanes Library, The Pennsylvania State University Libraries, Pennsylvania State University, Middletown, Pennsylvania, USA

[8]Department of Anthropology, University of Oregon, Eugene, Oregon, USA

[9]Department of Marine Science, California State University Monterey Bay, Seaside, California, USA

[10]Moss Landing Marine Laboratories, Moss Landing, California, USA

[11]UOG Marine Laboratory, University of Guam, Mangilao, Guam

[12]Department of Integrative Biology, University of Texas at Austin, Austin, Texas, USA

[13]Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, Ohio, USA

[14]Department of Natural Science, Hawai'i Pacific University, Honolulu, Hawaii, USA

[15]Department of Biological Sciences, Auburn University, Auburn, Alabama, USA

[16]Biodiversity Institute, University of Kansas, Lawrence, Kansas, USA

[17]Berkeley Natural History Museums, University of California, Berkeley, Berkeley, California, USA

[18]School of Biological Sciences, The University of Queensland, Brisbane, Queensland, Australia

[19]Hawai'i Institute of Marine Biology, University of Hawai'i at Mānoa, Kaneohe, Hawaii, USA

Correspondence
Eric D. Crandall, Department of Biology, Pennsylvania State University, University Park, 208 Mueller Lab, PA 16801, USA.
Email: eric.d.crandall@gmail.com

Present address

Abstract

Genetic diversity within species represents a fundamental yet underappreciated level of biodiversity. Because genetic diversity can indicate species resilience to changing climate, its measurement is relevant to many national and global conservation policy targets. Many studies produce large amounts of genome-scale genetic diversity data for wild populations, but most (87%) do not include the associated spatial and temporal metadata necessary for them to be reused in monitoring programs or for acknowledging the sovereignty of nations or Indigenous peoples. We undertook a distributed datathon to quantify the avail-

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2023 The Authors. Conservation Biology published by Wiley Periodicals LLC on behalf of Society for Conservation Biology.

15231739, 0, Downloaded from https://conbio.onlinelibrary.wiley.com/doi/10.1111/cobi.14061, Wiley Online Library on [17/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Rachel H. Toczydlowski, Northern Research Station, United States Forest Service, Rhinelander, WI 54501, USA.
Justin T. Berg, Department of Oceanography, University of Hawai'i at Mānoa, Kaneohe, HI 96744, USA.
Gideon S. Bradburd, Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA.

Article impact statement: Preservation and stewardship for genomic data that describe global genetic diversity is possible, but must happen now.

Funding information
National Science Foundation, Grant/Award Numbers: NSF-DEB-1457848, NSF-OCE-1764316

ability of these missing metadata and to test the hypothesis that their availability decays with time. We also worked to remediate missing metadata by extracting them from associated published papers, online repositories, and direct communication with authors. Starting with 848 candidate genomic data sets (reduced representation and whole genome) from the International Nucleotide Sequence Database Collaboration, we determined that 561 contained mostly samples from wild populations. We successfully restored spatiotemporal metadata for 78% of these 561 data sets ($n = 440$ data sets with data on 45,105 individuals from 762 species in 17 phyla). Examining papers and online repositories was much more fruitful than contacting 351 authors, who replied to our email requests 45% of the time. Overall, 23% of our email queries to authors unearthed useful metadata. The probability of retrieving spatiotemporal metadata declined significantly as age of the data set increased. There was a 13.5% yearly decrease in metadata associated with published papers or online repositories and up to a 22% yearly decrease in metadata that were only available from authors. This rapid decay in metadata availability, mirrored in studies of other types of biological data, should motivate swift updates to data-sharing policies and researcher practices to ensure that the valuable context provided by metadata is not lost to conservation science forever.

KEYWORDS
biodiversity, conservation genetics, Convention on Biological Diversity, digital sequence information, evolution, genetic diversity, metadata, molecular ecology, open data

Importancia de la curación oportuna de metadatos para la vigilancia mundial de la diversidad genética

Resumen: La diversidad genética intraespecífica representa un nivel fundamental, pero a la vez subvalorado de la biodiversidad. La diversidad genética puede indicar la resiliencia de una especie ante el clima cambiante, por lo que su medición es relevante para muchos objetivos de la política de conservación mundial y nacional. Muchos estudios producen una gran cantidad de datos sobre la diversidad a nivel genético de las poblaciones silvestres, aunque la mayoría (87%) no incluye los metadatos espaciales y temporales asociados para que sean reutilizados en los programas de monitoreo o para reconocer la soberanía de las naciones o los pueblos indígenas. Realizamos un "datatón" distribuido para cuantificar la disponibilidad de estos metadatos faltantes y para probar la hipótesis que supone que esta disponibilidad se deteriora con el tiempo. También trabajamos para reparar los metadatos faltantes al extraerlos de los artículos asociados publicados, los repositorios en línea y la comunicación directa con los autores. Iniciamos con 838 candidatos de conjuntos de datos genómicos (representación reducida y genoma completo) tomados de la colaboración internacional para la base de datos de secuencias de nucleótidos y determinamos que 561 incluían en su mayoría muestras tomadas de poblaciones silvestres. Restauramos con éxito los metadatos espaciotemporales en el 78% de estos 561 conjuntos de datos ($n = 440$ conjuntos de datos con información sobre 45,105 individuos de 762 especies en 17 filos). El análisis de los artículos y los repositorios virtuales fue mucho más productivo que contactar a los 351 autores, quienes tuvieron un 45% de respuesta a nuestros correos. En general, el 23% de nuestras consultas descubrieron metadatos útiles. La probabilidad de recuperar metadatos espaciotemporales declinó de manera significativa conforme incrementó la antigüedad del conjunto de datos. Hubo una disminución anual del 13.5% en los metadatos asociados con los artículos publicados y los repositorios virtuales y hasta una disminución anual del 22% en los metadatos que sólo estaban disponibles mediante la comunicación con los autores. Este rápido deterioro en la disponibilidad de los metadatos, duplicado en estudios de otros tipos de datos biológicos, debería motivar la pronta actualización de las políticas del intercambio de datos y las prácticas de los investigadores para asegurar que en las ciencias de la conservación no se pierda para siempre el contexto valioso proporcionado por los metadatos.

15231739, 0, Downloaded from https://conbio.onlinelibrary.wiley.com/doi/10.1111/cobi.14061, Wiley Online Library on [17/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

PALABRAS CLAVE
biodiversidad, Convenio sobre la Diversidad Biológica, datos abiertos, diversidad genética, ecología molecular, evolución, información de secuencia digital, metadatos

【摘要】
物种内的遗传多样性代表了生物多样性的一个基本水平, 但却没有得到重视。遗传多样性可以反映物种面对气候变化的恢复力, 因此遗传多样性的测量被纳入了许多国家和全球保护政策目标。许多研究产生了大量野生种群基因组水平的遗传多样性数据, 但大多(87%)不包括相关的空间和时间元数据, 而这是在监测项目中重新使用这些遗传多样性数据或是明确数据所属国家或原住民主权所必需的。本研究进行了一次分散式的数据马拉松, 以量化这些缺失元数据的可用性, 并检验了其可用性随时间衰减的假设。我们还通过相关发表的论文、在线资源库以及与作者直接沟通来获取这些元数据, 以努力补齐缺失的元数据。我们从国际核苷酸序列数据库合作联盟的848个候选基因组数据集(简化基因组和全基因组)出发, 确定了561个数据集主要包含来自野生种群的样本。我们成功地恢复了其中78%的数据集的时空元数据(440个数据集, 包括17个门、762个物种的45,105个个体的数据)。我们发现, 检查论文和在线资源库比直接联系作者更高效, 我们联系的351位作者有45%回复了我们的电子邮件请求。在与作者的电子邮件通信中, 总计23%的情况下获得了有用的元数据。随着数据集发表时间的增加, 检索得到时空元数据的概率明显下降。从发表论文或在线资源库中获得的元数据年均减少13.5%, 而通过作者获得的元数据则年均减少高达22%。这种元数据可用性的快速下降也反映在对其他类型生物数据的研究中, 表明数据共享政策和研究者实践应迅速更新, 以确保元数据所提供的珍贵背景信息不会永远消失在保护科学中。**【翻译:胡怡思;审校:聂永刚】**

生物多样性: 演化, 分子生态学, 保护遗传学, 元数据, 遗传多样性, 开放数据, 数字序列信息, 《生物多样性公约》

INTRODUCTION

Genetic diversity is the foundational layer of biodiversity. Just as ecosystem functioning and resilience depends on the diversity of its component species, so too does the health and resilience of each species depend on its genomic diversity (Clark, 2010; Reusch et al., 2005). Without genetic diversity in the form of standing allelic variation, populations and species cannot adapt to a rapidly changing climate and other anthropogenically induced or natural stresses (Blanchet et al., 2020; Raffard et al., 2019). Local or global extinctions of species in turn threaten the ecosystems upon which the quality of human lives depends (Brauman et al., 2020; Des Roches et al., 2021). Concerningly, genetic diversity, like all levels of biodiversity, is declining rapidly during the Anthropocene across the tree of life (Exposito-Alonso et al., 2022; Leigh et al., 2019; Miraldo et al., 2016; Pinsky & Palumbi, 2014).

Recognizing the vital importance of biodiversity to human well-being and the future of the planet, several international agreements strongly encourage the monitoring and conservation of genetic diversity in both wild and domesticated species. Foremost among these are the United Nations Sustainable Development Goal 2.5 and the international Convention on Biological Diversity (CBD) treaty, which explicitly acknowledge the importance of monitoring and conserving any component of biological diversity (including genetic diversity) that may have "actual or potential use or value for humanity." Moreover, the CBD's article 15 and attendant Nagoya Protocol codify procedures to ensure the sharing of benefits arising from genetic resources (such as digital sequence information [DSI]) discovered or accessed within a nation's sovereign borders. The subsequent Strategic Plan for Biodiversity 2011–2020 laid out the 20 Aichi Biodiversity Targets, including target 13, which aims to maintain the "genetic diversity of cultivated plants and farmed and domesticated animals and of wild relatives, including other socio-economically as well as culturally valuable species." Now, even as shortfalls on all 20 of the Aichi Biodiversity Targets are evident (CBD, 2020; Hoban et al., 2021; Laikre et al., 2020), the new Kunming-Montreal Global Biodiversity Framework, signed at the CBD Conference of the Parties 15 in 2022, includes maintenance and restoration of the genetic diversity of all wild and domesticated species (Goal A, Target 4), as well as provision of appropriate access to genetic resources (Goal C, Target 13). Simultaneously, there is now a global effort to sequence the genomes of all eukaryotic species in what has been described as a "moonshot for biology" (Lewin et al., 2018).

Over the last decade, advances in DNA sequencing technology have enabled the generation of genome-scale data sets of ever larger numbers of individuals, drawn from a growing variety of species (Allendorf, 2017; Hendricks et al., 2018). Researchers are now able to genotype thousands of genomic loci or sequence whole genomes from nonmodel species for which they have no prior genetic resources (Lou et al., 2021; Willette et al., 2014). The shift from genetic- to genomic-scale

15231739, 0, Downloaded from https://conbio.onlinelibrary.wiley.com/doi/10.1111/cobi.14061, Wiley Online Library on [17/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

data sets is catalyzing novel conservation insights, including the detection of inbreeding depression (e.g., Kardos et al., 2016); discovery of subtle, previously undetectable population structure (e.g., Cheng et al., 2021; Gaither et al., 2018); reconstruction of demographic histories (Prada et al., 2016); precise identification of distant pedigree relationships (e.g., Baetscher et al., 2019); uncovering cryptic species (e.g., Quattrini et al., 2019); clues about the genomic basis of local adaptation (e.g., Wilder et al., 2020); and important traits, such as nutritional components (e.g., Kumar et al., 2021). Accordingly, the DSI derived in these studies is highly valued as a resource equivalent to biobanks, providing essential information for conservation (Hoban et al., 2022) and ensuring future food security (Castañeda-Álvarez et al., 2016; Halewood et al., 2018).

Genomic data sets record the genetic diversity of a species at a particular time and location, providing a benchmark for how populations are responding to human-caused environmental change, cultivation, and land and sea use, as well as measuring indicators of progress toward conservation targets and goals (Hoban et al., 2020, 2022) and the genetic resources available for future cultivation or domestication (Halewood et al., 2018). However, genomic data sets can only be useful for monitoring global genetic biodiversity and the sustainable human use of genetic diversity (including benefit sharing [Cowell et al., 2022]) when archived publicly with accompanying metadata about the spatiotemporal, environmental, and methodological context of the sequenced sample (Riginos et al., 2020; Scholz et al., 2022; Schriml et al., 2020).

The genetics community has long championed open data publication with the foundational databases of the International Nucleotide Sequence Database Collaboration (INSDC) (Cochrane et al., 2016) formed in the early 1980s. In 2009, the INSDC launched the Sequence Read Archive (SRA) as a repository dedicated to second-generation sequence data. It has since grown exponentially to include over 600 terabytes of freely available DNA sequence data from over 16,700 wild and domesticated eukaryotic species as of 2021 (Toczydlowski et al., 2021). Around the same time, the MIxS metadata standards (Field et al., 2008; Yilmaz et al., 2011) were defined to inform the minimum information about what (detailed taxonomy), where (GPS coordinates and habitat), when (collection date), how (sampling and sequencing protocols), and by whom a genetic sample was collected. Enabled by the INSDC infrastructure and encouraged by the Joint Data Archiving Policy (JDAP; http://datadryad.org/pages/jdap) implemented by top journals in 2011, the proportion of papers providing open access to their genetic data increased dramatically (Pope et al., 2015). However, the inclusion of accompanying metadata crucial for the reuse of these data for genetic diversity monitoring and conservation, macrogenetic studies, or identifying their provenance within national boundaries or the lands and waters of Indigenous peoples has lagged behind (Pope et al., 2015; Toczydlowski et al., 2021). As of 2021, out of over 300,000 SRA BioSamples that are potentially relevant to global genetic biodiversity, only ~13% had metadata indicating the time and precise location from which they were sampled (Toczydlowski et al., 2021).

In a timely and welcome update to their policy, INSDC now intends to extend their minimum metadata requirements to include collection date and country of origin (https://www.insdc.org/spatio-temporal-annotation-policy-18-11-2021). Although country is legislatively aligned with the Nagoya Protocol, it is not spatially aligned with the lands and waters of Indigenous peoples (e.g., https://native-land.ca/) and does not provide adequate spatial resolution for conservation monitoring. Moreover, this policy and infrastructure change will take time to implement (anticipated to be end of 2022), meaning that much of the genomic data collated over the last ~12 years for past and present populations, of immeasurable value to understanding and monitoring the biodiversity crisis, are not findable, accessible, interoperable, or reusable (FAIR) (Wilkinson et al., 2016). This absence of appropriate spatiotemporal metadata represents the effective loss of tens to hundreds of millions of dollars of research effort for most future purposes (Schriml et al., 2020; Toczydlowski et al., 2021), rendering associated genetic data invisible to government ministries and nongovernmental organizations tasked with protecting the world's natural environment (Laikre, 2010; Laikre et al., 2020). Moreover, without spatiotemporal provenance of genomic data enabling connection to the lands and waters of Indigenous peoples, these peoples will potentially lose out on benefits (e.g., capacity development, food security, biomedical advances) arising from genomic information originating within their territories (Liggins et al., 2021; Marden et al., 2021; McCartney et al., 2022; Scholz et al., 2022). There is urgency in addressing this metadata gap: previous studies of morphological (Vines et al., 2014) and genetic (Pope et al., 2015) data suggest that the probability of existing metadata ever being linked to the genomic data significantly decreases over time.

In 2020, we convened a distributed remote datathon (i.e., a sustained effort by a team to solve a data-oriented problem) to assess the availability of metadata outside of the INSDC; recover and curate metadata missing in INSDC from external sources (i.e., published research papers, other online repositories, or the authors themselves); and extend our initial report on the metadata gap (Toczydlowski et al., 2021) to investigate how the recoverability of these metadata is affected by data set age and to document shortfalls and costs of our remedial efforts. In our datathon, 13 graduate students and 12 professional researchers worked together across 4 countries via Zoom, Slack, and Google Sheets as metadata curators to establish and execute curation protocols and infill missing metadata (24 of 25 curators are authors on this paper). Collectively, we searched for metadata external to the INSDC (e.g., associated scientific publications, Dryad, museum collections) for 848 genomic data sets (INSDC BioProjects) representing 94,416 individual samples (BioSamples). The BioSamples and associated genetic sequence data in these projects were selected because they were missing at least latitude and longitude metadata in the INSDC. We sought to underscore the importance of appropriate and immediate metadata archival. We devised guidelines based on our collective experience gained over the datathon on practices to retain crucial metadata.

15231739, 0, Downloaded from https://conbio.onlinelibrary.wiley.com/doi/10.1111/cobi.14061, Wiley Online Library on [17/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

METHODS

Datathon workflow

The workflow of our datathon is shown in Figure 1. A full-text description is provided in Appendix S1. Briefly, on 7 November 2019, we searched the INSDC to identify BioProjects (data sets) potentially relevant to monitoring genetic diversity but lacking critical metadata about latitude and longitude of the sampling location with the rentrez R package (Winter, 2017) and custom R scripts. We further filtered the BioProjects to remove BioSamples (sequenced individuals) from species whose population dynamics and evolution are largely governed by humans: pathogens and their vectors, model organisms, and domesticated species. We used custom lists for each category of nonwild organisms (Appendix S2) (see supporting information of Toczydlowski et al. [2021] for construction details). We built a blank template (Appendix S3) to receive metadata that we located external to the INSDC with the Genomic Observatories Metadatabase (GEOME) (Deck et al., 2017; Riginos et al., 2020). Metadata curators were each randomly assigned a set of BioProjects. Curators followed a standard protocol (Figure 1; Appendix S4) to locate associated publications for each BioProject, determine their relevance to natural genetic diversity, and enter associated metadata for samples in each relevant BioProject that were missing in the INSDC but reported in external sources (e.g., associated published scientific papers or online repositories). After performing quality control, these metadata could then be easily uploaded to GEOME and potentially added to the appropriate INSDC databases.

After adding all metadata that could be gleaned from the & associated papers into the GEOME templates, curators made a structured comment on a master spreadsheet (Appendices S5 & S6) indicating whether metadata for each of the required and recommended terms were absent for all BioSamples (none), present for <50% of BioSamples (some), present for >50% of BioSamples (most), or present for all BioSamples (all). If the paper was missing information from 1 of 6 required Darwin Core terms (Wieczorek et al., 2012) (georeferenceable *locality* OR [*decimalLttitude* AND *decimalLongitude*], *coordinateUncertaintyInMeters*, *georeferenceProtocol*, *habitat*, *environmentalMedium*, *yearCollected*), the curator flagged the BioProject to initiate author contact. We considered an additional 9 metadata terms as recommended: missing metadata in these fields alone did not trigger an author contact but curators and authors were asked to populate these fields as completely as possible. These recommended terms included *country*, *establishmentMeans*, *permitInformation*, *associatedReferences*, *preservative*, and 4 de novo terms that tracked genetic data derived from the raw reads, such as SNP genotypes or sequence alignments (*derivedGeneticDataType*, *derivedGeneticDataURI*, *derivedGeneticDataFormat*, and *derivedGeneticDataRemarks*). Progress and notes at each curation step were tracked as meta-metadata on the master spreadsheet.

After a quality-control step to ensure that author names and email addresses found in papers were input correctly, corresponding authors of the paper were contacted by email (text of the email is in Appendix S7) with the Yet Another Mail Merge add-on for Google Sheets (yamm.com). If an email was undeliverable, we tried to locate an alternate email address. We were able to successfully deliver email queries for 351 of 492 relevant BioProjects that met the criteria for author contact. About 2 weeks after sending the initial email, curators sent reminder emails to unresponsive authors at least once and at most twice. This process emulated the efforts of a reasonably persistent researcher to obtain metadata important to their research. Filled and checked GEOME templates for each BioProject are available in the GEOME database (https://geome-db.org). The data we collected about whether or not authors responded to emails or provided metadata are exempt from the human subjects regulation 45 CFR 46 as a category 2 exemption. We anonymized these data by separating identifying information about BioProjects (Appendix S5) from the author response data (Appendix S6) and randomizing the order of the data sets in each data file.

Investigating metadata decay

We investigated the effect of BioProject age on the probability that we were able to recover metadata information for 11 metadata categories. We used Bayesian logistic regression to fit 4 distinct models to investigate the relationships between BioProject age (number of days between publication in the INSDC and 7 November 2019) and the probability that metadata could be retrieved from INSDC, associated published papers, or repositories (model A), the probability that we received an author response for the 351 BioProjects that triggered an author contact via email (B), the probability that authors provided any metadata, given that they responded (C), and the probability that authors provided metadata for a majority of samples, given that they responded (D).

Information about the collection date and location of a sample is the most critical piece of metadata required to make genomic sequence data reusable and to identify its Indigenous provenance, so we focused our investigations on these 2 categories. We refer to the aggregate as spatiotemporal metadata. We defined a BioProject as having spatiotemporal data if collection dates and latitudes and longitudes or locality were present for at least 50% of the BioSamples that it contained. In model C, we counted a gain in collection year, or place name, or latitude and longitude for any number of BioSamples as recovery of metadata. In model D, we only counted increases in metadata for BioProjects that had incomplete spatiotemporal metadata for >50% of its BioSamples and had spatiotemporal metadata present for >50% of BioSamples after contacting authors. That is, model C assessed the probability of recovering any metadata external to the INSDC, and model D assessed the probability of recovering metadata for the majority of samples. In supplemental analyses, we investigated how the availability of metadata in individual spatiotemporal terms and other important metadata terms decayed (Appendices S8 & S9).

We conducted all statistical analyses at the level of BioProject (as opposed to BioSamples or genomic sequences) because presence or absence of metadata for BioSamples in a given

15231739, 0, Downloaded from https://conbio.onlinelibrary.wiley.com/doi/10.1111/cobi.14061, Wiley Online Library on [17/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

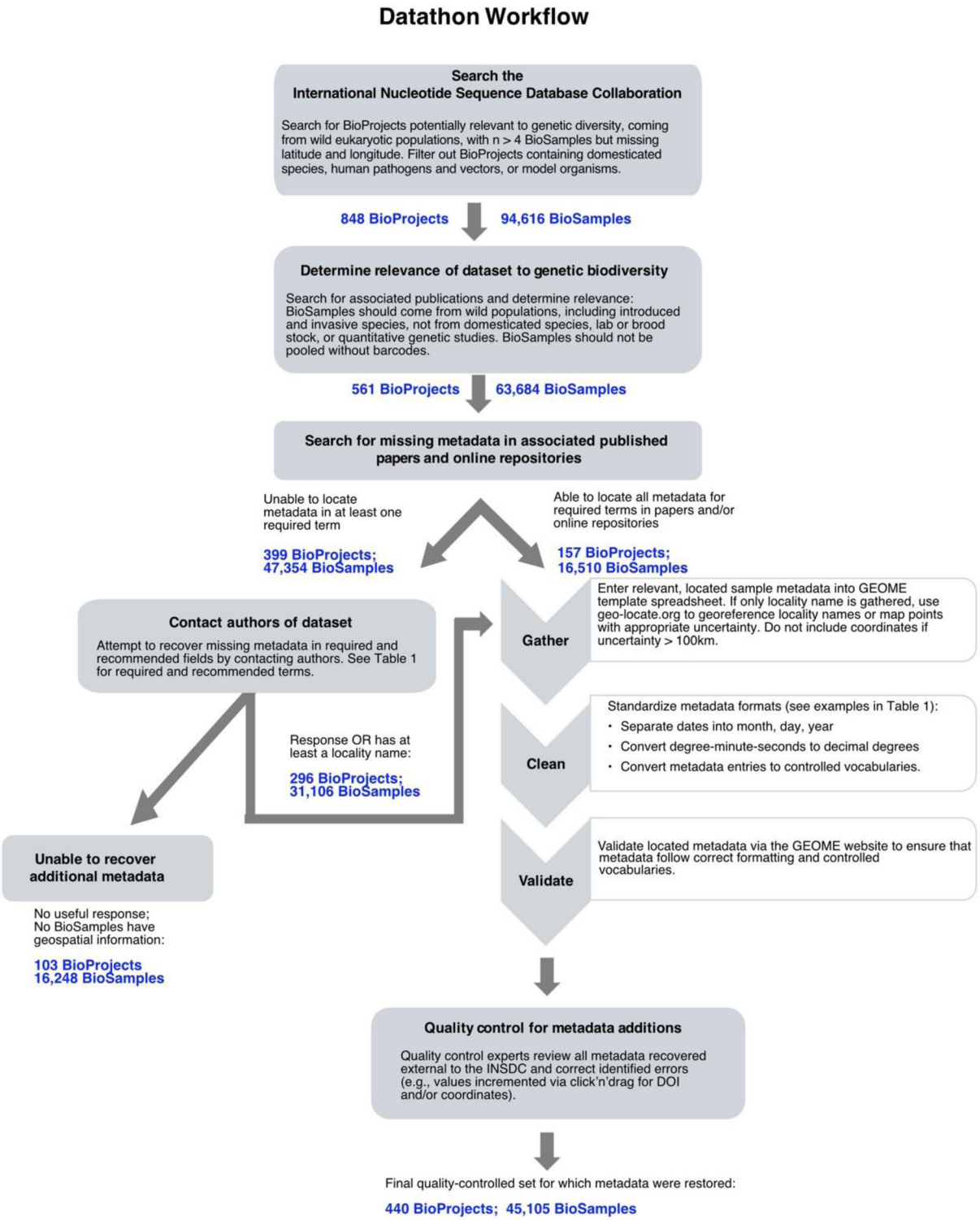

FIGURE 1 Workflow in datathon to identify and recover metadata for data sets potentially relevant to monitoring genetic diversity. The number of BioProjects and BioSamples remaining after each step are given below the step.

15231739, 0, Downloaded from https://conbio.onlinelibrary.wiley.com/doi/10.1111/cobi.14061, Wiley Online Library on [17/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

BioProject was highly correlated (Toczydlowski et al., 2021). We analyzed the effect of BioProject age on our response variables, given above for each model A–D, with generalized linear models. In each analysis, we modeled our response variable as a Bernoulli-distributed variable with a probability of success that was a linear function of our predictor variable: BioProject age. In each analysis, the parameters of our model were a global mean probability of success and an effect size of BioProject age on probability of success for that response variable. In these analyses, we used the canonical inverse-logit inverse link function. In mathematical notation, our model was

$$Y_i \sim Bernoulli\left(p_i\right), \tag{1}$$

$$p_i = \frac{1}{1 + e^{-\theta_i}}, \tag{2}$$

$$\theta_i = \mu + \beta \times X_i, \tag{3}$$

where Y_i is the ith outcome (response variable), p_i is the probability of successfully observing that outcome, μ is the global mean probability of success, and β is the effect of BioProject age on the transformed probability of success for that outcome (θ_i). We had no strong prior beliefs about the effect of BioProject age on success in each of the 4 analyses we ran; thus, the priors we placed on our parameters were $\beta \sim N(0,10)$; $\mu \sim N(0,10)$. All statistical analyses were performed using Rstan 2.21.2 (Stan Development Team, 2021). We ran 4 independent chains for 2000 iterations and thinned to sample only every fourth iteration to reduce autocorrelation. We discarded the first 1000 iterations as burn-in. To assess the significance of the effect of BioProject age on success of each outcome, we determined whether the 95% equal-tailed credible interval of the marginal distribution on β contained 0. If it did, the effect of BioProject age was deemed not significant.

RESULTS

We identified 848 INSDC BioProjects (registration dates ranging from 2012 to 2019), representing 94,416 BioSamples from individual eukaryotic organisms that lacked geospatial coordinates and had at least 5 putatively wild individuals as determined by our filters. Curators located associated published scientific papers for 741 of these 848 BioProjects (missing papers are likely in preparation or abandoned). Reading these papers revealed 561 BioProjects with a majority of relevant, truly wild individuals, comprising 63,684 individuals from 873 species. After scouring associated published papers for metadata and contacting authors, 440 BioProjects with 45,105 BioSamples from 762 species in 17 eukaryotic phyla (Figure 2) had geospatial data (either coordinates or a locality name) and were passed through quality control for eventual upload to GEOME. BioSamples that passed through the datathon came from all continents and all major oceans (Figure 3).

For the subset of BioProjects that we focused on (those missing latitude and longitude), datathon curators recovered metadata for a majority of BioSamples in a BioProject as follows (Figure 4). For geospatial coordinates, nearly 60% were found in an associated publication or online repository. Although nearly 30% of these BioProjects contained information about collection year in the INSDC, curators were only able to recover an additional 21% from papers or online repositories. Datathon curators recovered metadata regarding habitat, environmental medium (media displaced by the sampled organism), and publication DOI for over 80% of BioProjects from published papers and their supplemental information. Additional large gains in BioProjects were made from online sources external to the INSDC for locality (48.8%) and country name (39.8%). Notably, permit information was the least available of any of the metadata categories we explored. There is no permit metadata term in the INSDC, and curators found permit information in papers for only 21% of BioProjects.

Contacting authors yielded comparatively less metadata than our search of papers and supplemental information, although this step was secondary to examining papers and looking online. Out of 351 author contact attempts, we received 158 responses (45% response rate). Of the 158 responses, 80 (51%) provided at least some missing metadata, yielding an overall useful author response rate of 23%. Through contacting authors, we recovered collection year metadata for an additional 9% of BioProjects and geospatial coordinates for an additional 8.5% of BioProjects. Gains in other metadata categories were all <5%; permit information showed only a 1.2% increase with information from authors.

The age (time since deposition into the INSDC) of the BioProject had a strong effect on whether metadata could be recovered. After searching for metadata in the INSDC and published papers, we found that spatiotemporal metadata (defined as year and geospatial coordinates or locality) had a mean odds ratio of 0.865 (95% highest posterior density credible interval [HPD CI]: 0.775–0.964 (Figure 5a). This indicated that for every year after a BioProject was published to the SRA, there was about a 13.5% decrease (HPD CI: 3.6–22.5) in the probability that its metadata could be found in the SRA, in papers, or elsewhere online. In contrast, there was a strong positive effect of BioProject age on whether an attempt to contact the authors was answered, with a 25.5% increase in the probability of a reply of any kind for every year after SRA publication (mean odds ratio of 1.255; 95% HPD CI: 1.120–1.412) (Figure 5b). In other words, we were more likely to get an email response for older data sets. However, given a response, the probability that authors furnished any metadata for year or coordinates or locality decreased with BioProject age by 21% per year (odds ratio 0.810; 95% HPD CI: 0.680–0.949) (Figure 5c). Similarly, the probability that the authors provided metadata for year and coordinates or locality for a majority of BioSamples decreased by 22% per year (odds ratio 0.819; 95% HPD CI: 0.671–0.994) (Figure 5d).

Figures for Bayesian logistic regressions of BioProject age on other metadata categories are in Appendices S8 and S9 (β values). In accordance with the results for spatiotemporal

15231739, 0, Downloaded from https://conbio.onlinelibrary.wiley.com/doi/10.1111/cobi.14061, Wiley Online Library on [17/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

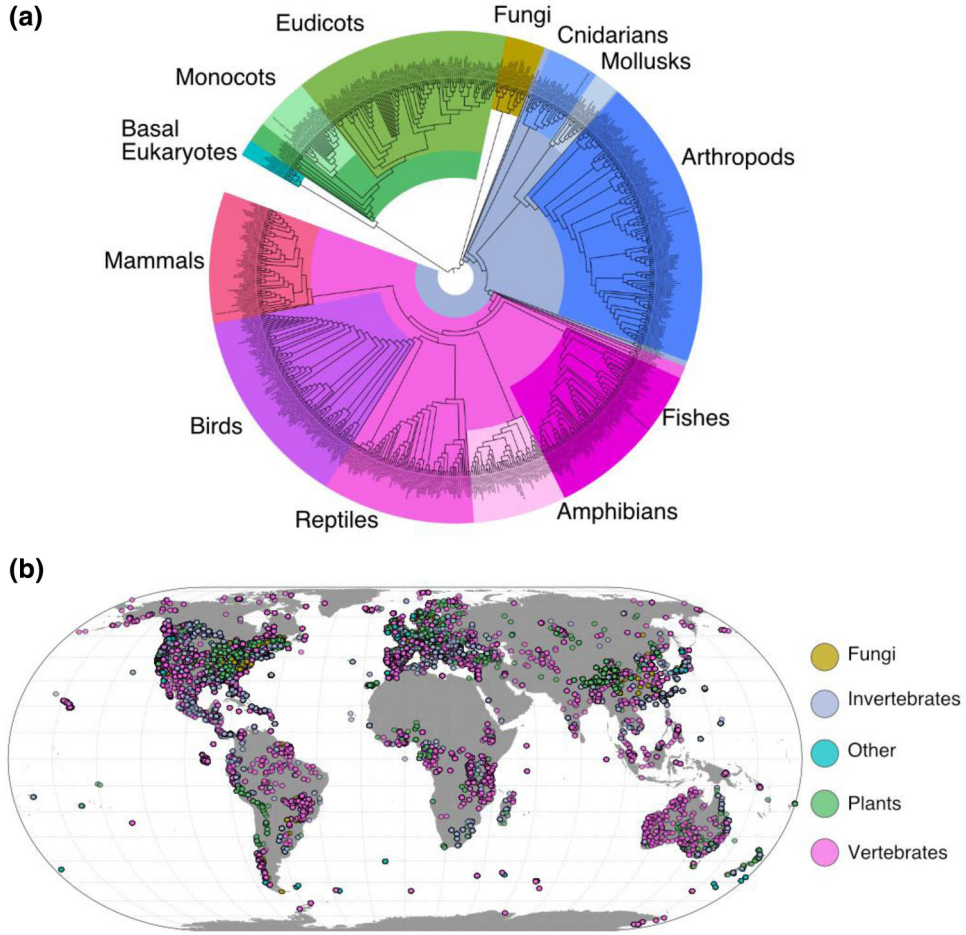

FIGURE 2 Taxonomic and geographic scope of the datathon to identify and recover metadata for data sets potentially relevant to monitoring genetic diversity: (a) cladogram of 719 of the 762 species from BioProjects that passed through the final quality control step (subtree of the Open Tree of Life [Hinchcliff et al., 2015] generated with the rotl package for R [Michonneau et al., 2016] and visualized with iTOL software [itol.embl.de] [Letunic & Bork, 2021]) and (b) geographic distribution of broad taxonomic categories of these BioSamples.

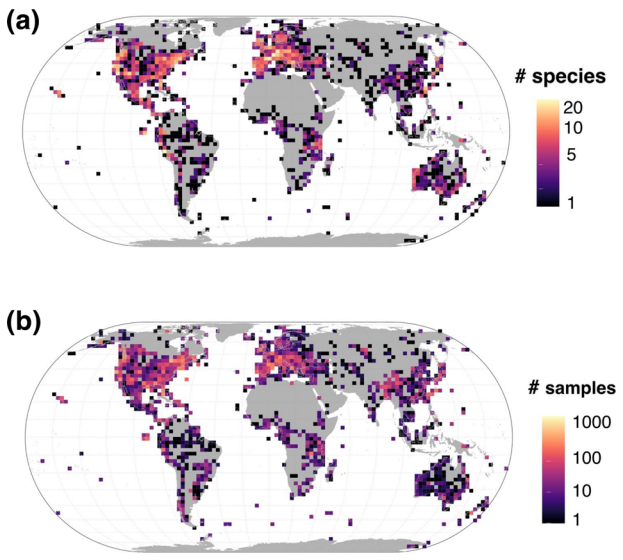

FIGURE 3 Distribution of (a) species and (b) BioSamples for which spatial coordinates were recovered by a datathon to identify and recover metadata for data sets potentially relevant to monitoring genetic diversity.

metadata, supplementary analyses indicated that metadata for collection year (posterior mean slope = −0.133, 95% credible interval: −0.233 to −0.034) (Appendices S8 & S9) and preservative used (posterior mean slope = −0.111, 95% HPD: −0.218 to −0.009) (Appendix S8) were significantly less likely to be recovered from INSDC, publications, and online repositories as age of a BioProject increased. Furthermore, and as with spatiotemporal metadata, the probability that responding authors provided additional metadata for georeferences (decimal latitude and decimal longitude) (posterior mean slope = −0.151, 95% credible interval: −0.386 to −0.05) (Appendix S8), collection year (posterior mean slope = −0.174, 95% credible interval: −0.363 to 0.000) (Appendix S8), and preservative used (posterior mean slope = −0.438, 95% credible interval: −0.873 to −0.081) (Appendix S8) was significantly greater for young BioProjects. The provisioning of permit information followed this same trend (although marginally insignificant, posterior mean slope = −0.555, 95% credible interval: −1.31 to 0.003) (Appendix S8), suggesting these metadata are relatively available within the personal data management system of authors.

15231739, 0, Downloaded from https://conbio.onlinelibrary.wiley.com/doi/10.1111/cobi.14061, Wiley Online Library on [17/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

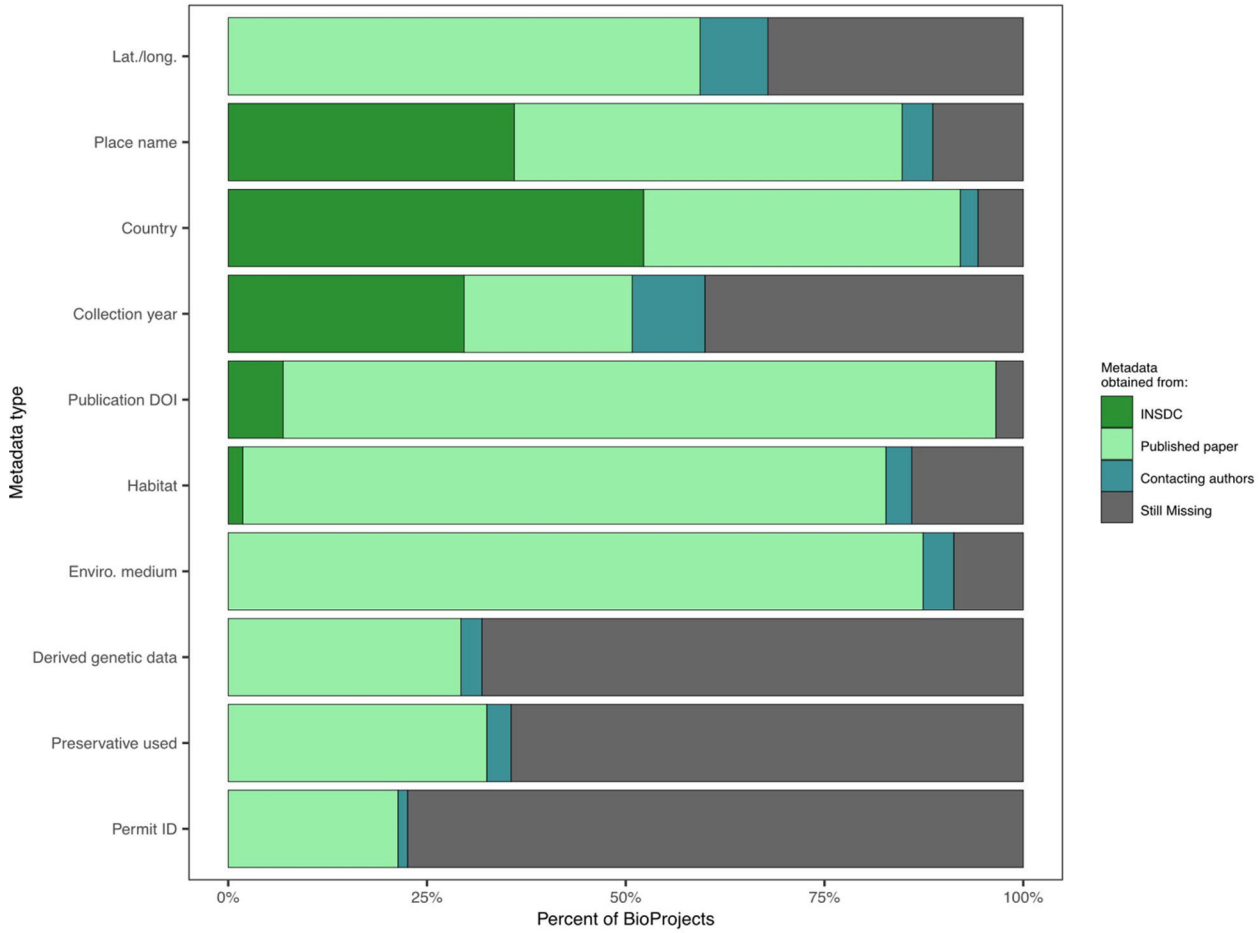

FIGURE 4 Percentage of International Nucleotide Sequence Database Collaboration BioProjects for which metadata were found from each of 3 sources across 10 priority metadata categories.

Counter to our result for spatiotemporal metadata, supplementary analyses indicated that metadata for habitat (Appendix S9) (posterior mean slope = 0.141, 95% credible interval: 0.006–0.285) (Appendix S8) and environmental medium (posterior mean slope = 0.176, 95% credible interval: 0.016–0.355) (Appendix S8) were less frequently recovered from INSDC, publications, and repositories for young BioProjects. Retrieval of these metadata through author contact had no relationship with BioProject age.

All code and meta-metadata are available from https://github.com/ericcrandall/geome_metadatathon1. Metadata for 45,105 SRA BioSamples recovered by the datathon are freely available from https://geome-db.org/workbench/project-overview?projectId=305. They and their associated genomic reads may be also be queried and downloaded using the geomedb package for R (Ewing & Crandall, 2020).

DISCUSSION

With our distributed datathon, we demonstrated that crucial metadata can be restored for many genomic investigations of wild organisms. However, our results showed that metadata were more difficult to recover as time proceeded and many were locked in nonstandard formats. Because the great majority of publicly available genomic data sets lack important metadata (Toczydlowski et al., 2021), they are not FAIR (Wilkinson et al., 2016). Only genomic data that are FAIR allow systematic monitoring of the fundamental layer of biodiversity (Hoban et al., 2021) and enable assertions regarding provenance for informing CBD Nagoya Protocol obligations. We found that metadata availability depends on type (location, publication, and habitat metadata were much more available or inferable than metadata about permits and preservatives); that with considerable time and paid effort, it is possible to recover some of these important metadata from the nonstandardized and nonmachine-readable formats in which they were being stored; and, although metadata archival practices may be improving incrementally, that genomic metadata are subject to the same decay processes demonstrated for other types of scientific data (Pope et al., 2015; Vines et al., 2014).

There are likely multiple factors underlying the observed decay in metadata availability. First, it is not surprising that older metadata are less likely to have been archived. Metadata

15231739, 0, Downloaded from https://conbio.onlinelibrary.wiley.com/doi/10.1111/cobi.14061, Wiley Online Library on [17/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

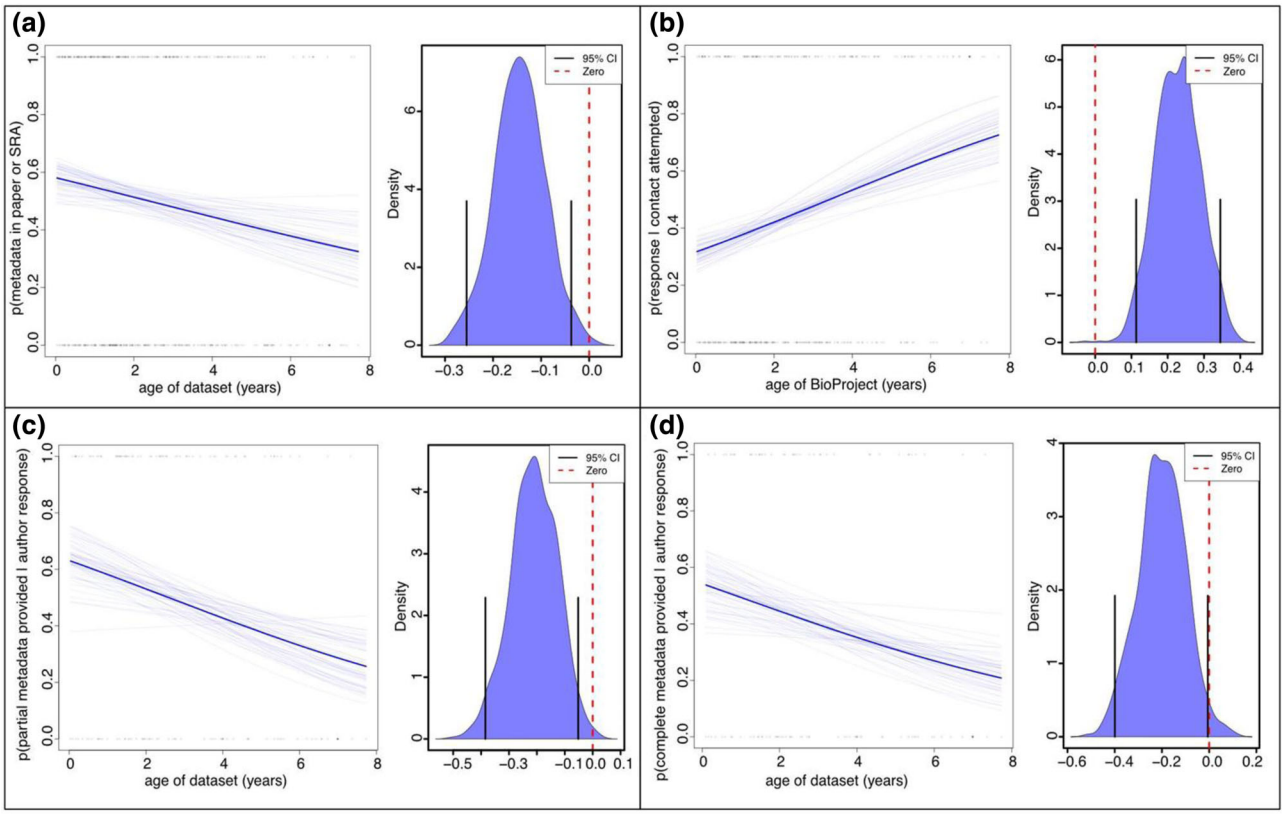

FIGURE 5 Effect of data set (i.e., BioProject) age on the probability of recovering spatiotemporal metadata (light colored lines, 1 of 2000 thinned iterations of the Bayesian analysis) and posterior distribution for log(odds ratio) (black lines, 95% highest posterior density [HPD]): probability (a) that metadata were found in the International Nucleotide Sequence Database Collaboration (INSDC) or associated papers and repositories, (b) of receiving a reply from BioProject authors to our contact email, (c) of receiving additional metadata for year or coordinates or locality, and (d) of receiving metadata for year and coordinates or locality for a majority of BioSamples (95% HPD intervals exclude 0).

archival practices are gradually improving; more metadata are being recorded in the INSDC, in research papers, and in online repositories, such as Data Dryad (Figure 5a). This is consistent with increasing acknowledgment that these metadata are relevant and important to future research. However, the rate of metadata archival is apparently not keeping up with the rapid growth of genomic data sets (see Figure 1 of Toczydlowski et al., [2021]), and it is certainly not closing the gap.

Second, we found that authors of recent SRA data sets were significantly less likely to reply to our queries than those of older data sets (Figure 5b), although the overall response rate of 45% was comparable to previous studies (Vines et al., 2013, 2014). This result may indicate that recent SRA depositions are part of ongoing research projects for which authors are unwilling to share metadata for fear of being scooped by others working on similar questions. It is also true that younger authors are more likely to leave science than older authors (Reithmeier et al., 2019); thus, these authors may no longer be available to support their publications. Similarly, there may be a cohort effect in which authors of older studies are more established in their careers and have more time or are more aware of increasing expectations around FAIR data and thus are more willing to communicate and share. As mandates for metadata increase, more data sets may minimally meet the metadata requirements, leading to a decreasing proportion of metadata in nonrequired categories.

Third, there has actually been decreased reporting over time of information about habitat and environmental medium. The reason for this trend is unclear, but if it continues, missing metadata about organisms' environmental context will make it difficult to address habitat-based conservation monitoring. Finally, of the authors who did reply, there was a significant decrease as the age of the BioProject increased in whether partial or complete spatiotemporal metadata were provided (Figure 5c,d), suggesting that if metadata are not properly archived to public repositories, they may be lost over time, as previously highlighted for morphological data (Vines et al., 2014).

Taken together, our results support assertions by others in the field that the current research system overly weights publications and citations and underweights scientific openness and reproducibility (Davies, Putnam, et al., 2021; Fidler et al., 2017; McNutt et al., 2016; Nosek et al., 2015). If these values were weighted appropriately by the academic system, we would not have found the metadata gap that we report here (O'Dea et al., 2021). Adding to the challenge, publications are rarely linked

15231739, 0, Downloaded from https://conbio.onlinelibrary.wiley.com/doi/10.1111/cobi.14061, Wiley Online Library on [17/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 Alphabetized list of required (asterisk) and recommended metadata terms for individual organisms and derived tissues or DNA sequences included in the datathon to identify and recover metadata for data sets potentially relevant to monitoring genetic diversity.

Term	Definition (metadata standard for the definition)[a]	Importance[b]	Controlled vocabulary[c]	Example
associatedReferences	Associated publications or references pertaining to this individual or its derivative tissues or sequences; the first place it was published is particularly relevant; DOIs in format https://doi.org/10.1007/s10530-007-9196-8; multiple DOIs separated by | (GEOME)[d]; list (concatenated and separated) of identifiers (publication, bibliographic reference, global unique identifier, URI) of literature associated with the occurrence (Darwin Core)[e]	Indigenous provenance, ABS class II	None	https://doi.org/10.1111/j.1365-294X.2008.03995.x; |https://doi.org/10.5343/bms.20
coordinateUncertaintyInMeters*	Horizontal distance (in meters) from the given decimal latitude and decimal longitude for the radius of the smallest circle containing the whole of the locality where the sample could possibly have come from; value empty if uncertainty is unknown, cannot be estimated, or is not applicable (because there are no coordinates); zero not a valid value for this term (Darwin Core)	Class I, III	None	1 km = 1000
Country	The name of the country or major administrative unit or exclusive economic zone (for marine samples) in which the locality occurs (Darwin Core)	Indigenous provenance, ABS class II	ISO 3166-1	Indonesia
decimalLatitude*	The geographic latitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a Location. Positive values are north of the Equator, and negative values are south of it. Legal values lie between −90 and 90, inclusive (Darwin Core).	Indigenous provenance, ABS class I, II, III	None	−6.147183
decimalLongitude*	The geographic longitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a location. Positive values are east of the Greenwich Meridian, and negative values are west of it. Legal values lie between −180 and 180, inclusive (Darwin Core).	Indigenous provenance, ABS class I, II, III	None	105.46326
derivedDataFilename	A list (concatenated and separated with |) of the file names for data sets that include data derived from this tissue that are accessible via the derivedDataURI. Could be a compressed archive (GEOME).	Class III	None	SDM_snps.tar.gz
derivedDataFormat	A list (concatenated and separated with |) of the data set formats relating to the derivedDataType that include data derived from this tissue (GEOME).	Class III	{microsatellites, sequence alignment, SNPs, OTUs, ASVs, other}	SNPs

(Continues)

15231739, 0, Downloaded from https://conbio.onlinelibrary.wiley.com/doi/10.1111/cobi.14061, Wiley Online Library on [17/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 (Continued)

Term	Definition (metadata standard for the definition)[a]	Importance[b]	Controlled vocabulary[c]	Example
derivedDataType	A list (concatenated and separated with |) of the data set types that include data derived from this tissue (GEOME).	Class III	{genepop, FASTA, VCF, nexus, PHYLIP, structure, other}	VCF
derivedDataURI	A URI (preferably a DOI in this format: https://doi.org/10.1007/s10530-007-9196-8) for any data sets that include data derived from this tissue. Multiple URIs/DOIs can be separated by | (GEOME).	Class III	None	https://doi.org/10.5061/dryad.k7k4m
environmentalMedium*	Terms that identify the material displaced by the entity at time of sampling. Recommend subclasses of environmental material (ENVO:00010483). Multiple terms can be separated by pipes, for example: a duck might displace fresh water|air (MIxS).[f]		ENVO[g] environmental material: ENVO_00010483	Sea water
habitat*	In this field, report which major environmental system your sample or specimen came from. The systems identified should have a coarse spatial grain, to provide the general environmental context of where sampling was done. [Darwin Core] A category or description of the habitat in which the Event occurred. (MixS[f]: Broad-scale environmental context)		ENVO Biomes: ENVO_00000428	Marine benthic biome
locality*	The specific name or description of the site or place where the sample was taken as given by the original researchers. This would be the place name that appears in a table next to the coordinates, or the labels for sampling sites on a map. Less specific geographic information can be provided in other geographic terms (continentOcean, country, stateProvince, island). This term may contain information modified from the original to correct perceived errors or standardize the description (Darwin Core).	Indigenous provenance, ABS class I, II, III	None	Rakata
materialSampleID*	The collector's specimen number. This number must be unique among the IDs within the sheet (GEOME). An identifier for the MaterialSample (as opposed to a particular digital record of the material sample). In the absence of a persistent global unique identifier, construct one from a combination of identifiers in the record that will most closely make the materialSampleID globally unique (Darwin Core).	Class I, III	None	Rakata_1190.01

(Continues)

15231739, 0, Downloaded from https://conbio.onlinelibrary.wiley.com/doi/10.1111/cobi.14061, Wiley Online Library on [17/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 (Continued)

Term	Definition (metadata standard for the definition)[a]	Importance[b]	Controlled vocabulary[c]	Example
permitInformation	Information regarding the permits acquired to collect and export this sample. At least the permit number and issuing authority. "No permit required per [authority]" is also valid. Multiple values separated by | (GEOME).	Indigenous provenance ABS	None	Indonesian Institute of Sciences (LIPI) Permit # 1187/ SU/KS/2006 | Indonesian Institute of Sciences (LIPI) Permit # 04239/SU.3/KS/2006
preservative	Preservative used on the specimen (GEOME).	Class I	GEOME list of preservatives	95% ethanol
yearCollected*	The year the collecting event took place (Darwin Core).	Class I, II, III	None	March 24, 2006 = 2006

[a]Terms with multiple definitions are in order of decreasing specificity.
[b]Terms support the identification of Indigenous provenance and can therefore inform access and benefit-sharing (ABS), and terms that can support sample or digital sequence information (DSI) reuse in conservation, according to the study approach definitions of Leigh et al. (2021). Class I studies generate new sequence data, requiring precise information regarding the spatiotemporal context of the collected sample, a unique materialSampleID, as well as the preservative the tissue is held in. In class II studies, genetic diversity values are compiled from published studies, generally requiring less precise spatiotemporal information, but this needs to be associated with a publication (associated references). In class III studies, digital sequence information, or derived genetic data, is reanalyzed, requiring precise spatiotemporal information and a unique material sample ID. Depending on the objective of reuse, habitat and environmental medium may also be important for sample or DSI reuse in conservation.
[c]Standardized lists of acceptable entries, often defined by a standards organization.
[d]Deck et al. (2017); Riginos et al. (2020)
[e]Wieczorek et al. (2012)
[f]Yilmaz et al. (2011)
[g]Burtigieg et al. (2013, 2016)

to genomic data in INSDC, which likely reflects authors first uploading their genomic data to meet publication requirements and then not returning to update the metadata when the paper is published.

Changing the system will likely require a combination of carrots and sticks (Whitlock, 2011). Carrots can take the form of citable data publications (Dimitrova et al., 2021); recognition of open data practices by hiring, promotion, and tenure committees; or commendations from professional societies or departments (Roche et al., 2014, 2015). Sticks in the form of open metadata mandates must come from journals (Gareth Jenkins, pers. comm.; Sibbett et al., 2020), funding agencies, and data repositories, which all have a responsibility to respond to the needs of the research community (Lin et al., 2020). Although we applaud the INSDC's new spatiotemporal metadata annotation policy requiring country of origin metadata and their adoption of the MIxS metadata standards, we call for greater mandated spatial resolution to include at least a descriptive and uniquely georeferenceable locality name or spatial coordinates (Table 1) with appropriate uncertainty or additional terms (such as Darwin Core's *coordinateUncertaintyInMeters* and *informationWithheld* [Wieczorek et al., 2012]) to protect endangered species or sovereignty of Indigenous peoples (Hudson et al., 2020; McCartney et al., 2022).

Our datathon provided a unique opportunity to train graduate students in the importance of proper data curation and to raise awareness that almost every data set has a potential for reuse. We suggest that training in data curation and metadata usage should be part of reproducible research training in every science graduate program, with emphasis on avoiding some of the metadata practices that hinder metadata recovery described in Table 2. Datathons, such as we undertook, could help close the metadata gap in the short term because they are very cost-effective. If one assumes a mean cost of sequencing of US $50 per BioSample (and ignores the much higher, additional cost of sample collection and processing), we rescued over US $2.1 million worth of genomic sequence data for future research purposes. Coauthors of this paper spent about 2300 h on this metadata retrieval effort, which, if valued at an average wage of US $19 per hour, yields a return on investment of nearly 4700%, with average costs of remediating a BioSample or BioProject at US $1.05 and US $110, respectively. But ultimately, datathons are a stopgap solution.

Going forward, the entire biodiversity genomics research community should give the same priority to sharing metadata that they have given to sharing primary data because it is only the metadata that make primary data FAIR. From a process standpoint, the collection of metadata should begin at the time of sampling, with the assignment of a globally unique identifier (GUID) to the actual material sample. This identifier, which should be assigned as early as possible after collection, serves as the root to which all subsequent derived products could be linked in an extended specimen cloud to establish clear provenance and thereby prevent duplication of data or effort (Davies, Deck, et al., 2021; Lendemer et al., 2020). Through the use of GUIDs, both physical and digital products of the sample (digital sequence information, but also DNA or RNA

15231739, 0, Downloaded from https://conbio.onlinelibrary.wiley.com/doi/10.1111/cobi.14061, Wiley Online Library on [17/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 Summary of metadata practices encountered that hindered metadata recovery and recommended practices to improve future usability of samples and genetic sequence data

Practice	Challenge	Solution
INSDC (International Nucleotide Sequence Database Collaboration) sample identifier does not match any sample identifiers in associated scientific publications.	Metadata external to the INSDC cannot be assigned to genetic samples or metadata are associated with the wrong sample.	Use consistent, persistent, globally unique sample identifiers (e.g., Darwin Core *materialSampleID*) across data repositories and publications; if sample identifiers are not consistent, provide an explicit cross-reference table in all associated publications and data repositories.
Large amounts of metadata are only available in associated publications in PDF format and lack consistent formatting.	Metadata cannot be programmatically converted to standard table formats (e.g., entries formatted column by row, rather than row by column) and time-consuming manual extraction is required.	Provide a link to a GEOME project, OR provide metadata in comma- or tab-delimited files (.csv or .txt) with standard column headers (i.e., terms suggested by Darwin Core, MIxS, or GEOME or COPO) and associated vocabulary, where possible.
Specimens, biosamples, or metadata are deposited in a biodiversity collection (e.g., museum, herbarium, biobank, zoo), but biodiversity collection accession numbers are not provided in the associated publications or INSDC.	Biodiversity collection record searches can be time consuming and may not yield enough information to link samples in collection databases back to INSDC databases.	Use consistent sample identifiers across all databases and publications or provide a cross-reference table in associated publications that links biodiversity collection accessions to INSDC materialSampleIDs and identifies the biocollection repository by name. The Darwin Core standard accommodates multiple terms.
Associated publication references previous publications for details about the sample metadata.	Time consuming and challenging to track citation trail back to metadata in original or earlier publication. Sample metadata or identifiers may be absent or in inconsistent formats across associated publications.	Provide a link to a GEOME project, OR compile and include relevant metadata from previous publications in supporting information for the publication linked to the INSDC BioProject. If needed, include a column flagging whether data are new to the present study or originated from another source and identify that source.
Sample collection geospatial coordinates or location name withheld to protect endangered species, sensitive habitat, or Indigenous sovereignty.	Sample lacks spatial metadata.	Provide imprecise geospatial coordinates and use large, defined *coordinateUncertainty* to maintain local anonymity of sensitive collection sites. Provide additional comments with Darwin Core *informationWithheld*.
Codes used to abbreviate sample collection locations are inconsistent or hard to find throughout publication and related materials.	Sample collection locations cannot be determined or require time-consuming manual curation.	Include site codes in the sample identifier (*materialSampleID*). Use consistent site codes throughout associated publications and provide a key with codes and geospatial coordinates in associated publications.
Sample collection dates are a range or a season (e.g., winter 2017–2018).	Sample collection date may not be identifiable to a specific year; unclear which samples were collected in which date range.	Denote or report which year each sample was collected in (dates that also include month and day are ideal), ideally in a .csv or .txt file.
No metadata on BioSample relevance to genetic diversity of wild populations or species provided.	Unclear which BioSamples (if any) were collected from wild populations versus, for example, brood stock, laboratory stock, domesticated species, artificial selection experiments, nonwild collections in seed banks.	Provide metadata denoting which BioSamples were collected from wild populations with Darwin Core term *establishmentMeans*.
Metadata provided for some but not all BioSamples	Some BioSamples lack metadata, unclear why metadata are incomplete.	Provide a link to a GEOME project, OR pProvide metadata for all BioSamples or list a specific reason for missing metadata (e.g., not collected, metadata lost, sample excluded from study due to misidentification) with Darwin Core *informationWithheld*.
Sampling location provided, but only at a coarse geographic scale, that is, state, province, or country name.	Sample lacks spatial metadata at a resolution useful for future monitoring and macrogenetic research questions.	Provide geospatial coordinates for sample collection locations. Specific place, state, and country names can be helpful additions to confirm the geospatial coordinates are correct (and to programmatically filter by broader geographic locations). If coordinates cannot be provided, give a descriptive and uniquely georeferenceable locality name.
Corresponding author email addresses have expired.	When researchers use institutional email addresses as corresponding authors and then change institutions, they can no longer be contacted at that address.	Use private, long-term email addresses for corresponding author contact, link ORCID to all published papers, and keep ORCID profile updated.

Note: In general, we recommend authors use metadata software (GEOME: geome-db.org, COPO: copo-project.org, or museum database software such as Specify) to organize and archive their sample metadata.

15231739, 0, Downloaded from https://conbio.onlinelibrary.wiley.com/doi/10.1111/cobi.14061, Wiley Online Library on [17/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

extractions, subsamples, images, video, audio, CT scans, measurements of morphology, traits, gut contents, parasites, and other related data and associated metadata) will be linked to their material sample GUID to provide an extensive, holistic metadata cloud that can be used to better inform current research endeavors and create additional data-intensive research pathways. GEOME (Deck et al., 2017; Riginos et al., 2020) is an example of an easy-to-use metadata broker platform that can provide spreadsheet templates with definitions that can be filled in offline when the sample is collected. It can then mint a GUID for any sample that is added to it and harvest the INSDC accession numbers for genomic reads that are submitted to the SRA through GEOME, thereby maintaining permanent links between the sample metadata and genomic data.

If GEOME and similar sample database software, such as Specify (Lawrence, Kansas), can store sample GUIDs and associated metadata, the challenge then is to integrate these metadata downstream into databases (such as INSDC) that describe data derived from the sample. The INSDC enables such linkages to other metadata platforms through the use of both Structured Voucher (https://www.ncbi.nlm.nih.gov/biocollections/docs/faq/) and Linkout (https://www.ncbi.nlm.nih.gov/projects/linkout/) facilities for both nucleotide and SRA (through their corresponding BioSample record) data sets, respectively (e.g., https://www.ncbi.nlm.nih.gov/nuccore/KC825472). Through these linkages, metadata corresponding to the original material sample can be tied to the resulting sequences to both validate the metadata associated with the sequence record and provide updated information should specimens be reidentified or georeferenced after the lodging of the sequence with INSDC. Using the INSDC as a long-term repository for metadata about the sample may not make sense, in part because researchers who submit the sequences to INSDC have sole editing rights to the sequence record and it is currently quite difficult for others (such as the collections who hold the vouchers) to keep the INSDC metadata up to date or add additional information. Thus, the integration of these metadata from an upstream source somewhat negates the necessity for this information to be duplicated by the sequence depositor and ensures that the metadata are constantly up to date. This not only supports open, reproducible science (Buckner et al., 2021) but also exemplifies the findable and accessible principles of FAIR data (Wilkinson et al., 2016).

What this piecemeal data archival system currently lacks, however, is support for data interoperability and reusability. This is because of the siloed nature of the data and our inability to compile it into a single resource for machine readability, data manipulation, or downstream use. This shortcoming is being addressed through various initiatives such as the Extended Specimen Network (Lendemer et al., 2020; Thiers et al., 2021), Digital Extended Specimen (https://dissco.tech/2020/03/31/what-is-a-digital-specimen/), Distributed System for Scientific Collections (https://www.dissco.eu/), iSamples (Davies, Deck, et al., 2021), and others. Such a system would require all actors in the data landscape (researchers, collections, data aggregators, publishers, etc.) to utilize and publish resolvable GUIDs on all specimens, data sets, and products of research to make these linkages possible and thereby create an extensive online network of knowledge and increase the potential for scientific research questions to be answered.

We join others in the research community in calling for the advancement of scientific practices that can effectively help safeguard genetic diversity (Des Roches et al., 2021; Díaz et al., 2020; Laikre et al., 2020) and protect the rights of developing nations and Indigenous communities by establishing provenance of both data and samples (Hudson et al., 2020, Liggins et al., 2021). Swift collective action is required to protect all levels of global biodiversity, and the first step toward protecting the evolutionary health of eukaryotic species worldwide is to close the metadata gap highlighted here. Simultaneously, conservation geneticists, molecular ecologists, and evolutionary biologists must engage with global biodiversity assessment programs, national resource management agencies, and Indigenous communities to ensure genomic data can be collected, interpreted, and archived appropriately (Brodersen & Seehausen, 2014; Hudson et al., 2020). Several exemplary international networks (e.g., GEOBON Genetic Composition Working Group, IUCN Conservation Genetics Specialist Group, and EU COST Action Genetic Biodiversity Knowledge for Ecosystem resilience) have already made a case for protecting the genetic diversity of all species (Laikre et al., 2020) and proposed indicators to gauge progress toward goals (Hoban et al., 2020; Laikre et al., 2020). These groups have asserted their rationale for these changes to stakeholders in policy documents, providing essential clarity in the use of genetic data and reporting against targets (Hoban et al., 2021). These actions and advances encourage the uptake of genetic diversity monitoring by national authorities and international bodies. The vision for many of these biodiversity monitoring networks is to develop agile pipelines that intake raw biodiversity data and produce outputs that can directly inform conservation policies and decisions (Hoban et al., 2021). Yet, without appropriate archival of genomic data that includes the spatiotemporal metadata, the promise of genetic diversity monitoring cannot be fulfilled.

The GEOME datathon enabled 13 graduate students and 12 professional researchers from 15 institutions and 4 countries to assess the growing metadata gap for genomics data and begin to remediate it. The serendipity of being able to run a remote, distributed datathon due to travel restrictions and funding reallocation forced by COVID-19, in a time when Indigenous rights, biodiversity conservation, and the value of genetic diversity have been front of mind, has not been lost on the participants. Although our efforts have just begun to address the growing metadata gap, it is our hope that most researchers will start to ensure the FAIRness of their genomic data and metadata before or upon publication, thereby honoring the work that went into creating it and providing limitless opportunities for reuse of their data to help answer the important scientific questions of the future.

ACKNOWLEDGMENTS

This effort arose from an Evolution in Changing Seas Research Coordination Network (RCN) working group (NSF-OCE-1764316, Katie Lotterhos) and was funded by the Diversity of

15231739, 0, Downloaded from https://conbio.onlinelibrary.wiley.com/doi/10.1111/cobi.14061, Wiley Online Library on [17/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the Indo-Pacific Network RCN (NSF-DEB-1457848 to R.J.T.). We gratefully thank all of the authors who took the time to provide helpful responses to our metadata inquiries, and G. Jenkins, editor in chief at *Ecology & Evolution*, for his comments about open data mandates from journals. We also thank N. Davies, C. Meyer, B. Davis, and K. Nielsen for their input.

ORCID

Eric D. Crandall https://orcid.org/0000-0001-8580-3651
Rachel H. Toczydlowski https://orcid.org/0000-0002-8141-2036
Libby Liggins https://orcid.org/0000-0003-1143-2346
Ann E. Holmes https://orcid.org/0000-0003-4775-868X
Maryam Ghoojaei https://orcid.org/0000-0003-3641-4881
Michelle R. Gaither https://orcid.org/0000-0002-0371-5621
Briana E. Wham https://orcid.org/0000-0002-9240-8606
Andrea L. Pritt https://orcid.org/0000-0003-1315-5194
Cory Noble https://orcid.org/0000-0001-7720-1060
Tanner J. Anderson https://orcid.org/0000-0002-4206-7815
Randi L. Barton https://orcid.org/0000-0001-8763-8389
Justin T. Berg https://orcid.org/0000-0001-5376-8767
Sofia G. Beskid https://orcid.org/0000-0003-4524-0916
Alonso Delgado https://orcid.org/0000-0001-7874-0293
Emily Farrell https://orcid.org/0000-0001-6713-2348
Nan Himmelsbach https://orcid.org/0000-0002-2317-6354
Samantha R. Queeno https://orcid.org/0000-0001-7058-2593
Thienthanh Trinh https://orcid.org/0000-0001-8747-7618
Courtney Weyand https://orcid.org/0000-0001-7162-2462
Andrew Bentley https://orcid.org/0000-0002-3093-1258
John Deck https://orcid.org/0000-0002-5905-1617
Cynthia Riginos https://orcid.org/0000-0002-5485-4197
Gideon S. Bradburd https://orcid.org/0000-0001-8009-0154
Robert J. Toonen https://orcid.org/0000-0001-6339-4340

REFERENCES

Allendorf, F. W. (2017). Genetics and the conservation of natural populations: Allozymes to genomes. *Molecular Ecology*, *26*, 420–430.

Baetscher, D. S., Anderson, E. C., Gilbert-Horvath, E. A., Malone, D. P., Saarman, E. T., Carr, M. H., & Garza, J. C. (2019). Dispersal of a nearshore marine fish connects marine reserves and adjacent fished areas along an open coast. *Molecular Ecology*, *1*, 0148–13.

Blanchet, S., Prunier, J. G., Paz-Vinas, I., Saint-Pé, K., Rey, O., Raffard, A., Mathieu-Bégné, E., Loot, G., Fourtune, L., & Garza, J. C. (2020). A river runs through it: The causes, consequences, and management of intraspecific diversity in river networks. *Evolutionary Applications*, *13*, 1195–1213.

Brauman, K. A., Garibaldi, L. A., Polasky, S., Aumeeruddy-Thomas, Y., Brancalion, P. H. S., DeClerck, F., Jacob, U., Mastrangelo, M. E., Nkongolo, N. V., Palang, H., Pérez-Méndez, N., Shannon, L. J., Shrestha, U. B., Strombom, E., & Garza, J. C. (2020). Global trends in nature's contributions to people. *Proceedings of the National Academy of Sciences*, *117*, 32799–32805.

Brodersen, J., & Garza, J. C. (2014). Why evolutionary biologists should get seriously involved in ecological monitoring and applied biodiversity assessment programs. *Evolutionary Applications*, *7*, 968–983.

Buckner, J. C., Sanders, R. C., Faircloth, B. C., & Garza, J. C. (2021). The critical importance of vouchers in genomics. *eLife*, *10*, e68264.

Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J., Lewis, S. E., the ENVO Consortium. (2013). The environment ontology: Contextualising biological and biomedical entities. *Journal of Biomedical Semantics*, *4*, 43.

Buttigieg, P. L., Pafilis, E., Lewis, S. E., Schildhauer, M. P., Walls, R. L., & Garza, J. C. (2016). The environment ontology in 2016: Bridging domains with increased scope, semantic density, and interoperation. *Journal of Biomedical Semantics*, *7*, 57.

Castañeda-Álvarez, N P., Khoury, C K., Achicanoy, H A., Bernau, V., Dempewolf, H., Eastwood, R J., Guarino, L., Harker, R H., Jarvis, A., Maxted, N., Müller, J V., Ramirez-Villegas, J., Sosa, C C., Struik, P C., Vincent, H., & Garza, J. C. (2016). Global conservation priorities for crop wild relatives. *Nature Plants*, *2*, 1–6.

Convention on Biological Diversity (CBD). (2020). *Global Biodiversity Outlook 5*. Author.

Cheng, S. H., Gold, M., & Garza, J. C. (2021). Genome-wide SNPs reveal complex fine scale population structure in the California market squid fishery (*Doryteuthis opalescens*). *Conservation Genetics*, *22*, 97–110.

Clark, J. S. (2010). Individuals and the variation needed for high species diversity in forest trees. *Science*, *327*, 1129–1132.

Cochrane, G., Karsch-Mizrachi, I., Takagi, T., International Nucleotide Sequence Database Collaboration. (2016). The International Nucleotide Sequence Database Collaboration. *Nucleic acids research*, *44*, D48–D50.

Cowell, C., Paton, A., Borrell, J S., Williams, C., Wilkin, P., Antonelli, A., Baker, W J., Buggs, R., Fay, M F., Gargiulo, R., Grace, O M., Kuhnhäuser, B G., Woudstra, Y., & Garza, J. C. (2022). Uses and benefits of digital sequence information from plant genetic resources: Lessons learnt from botanical collections. *Plants, People, Planet*, *4*, 33–43.

Davies, N., Deck, J., Kansa, E. C., Kansa, S. W., Kunze, J., Meyer, C., Orrell, T., Ramdeen, S., Snyder, R., Vieglais, D., Walls, R. L., & Garza, J. C. (2021). Internet of Samples (iSamples): Toward an interdisciplinary cyberinfrastructure for material samples. *GigaScience*, *10*, giab028.

Davies, S W., Putnam, H M., Ainsworth, T., Baum, J K., Bove, C B., Crosby, S C., Côté, I M., Duplouy, A., Fulweiler, R W., Griffin, A J., Hanley, T C., Hill, T., Humanes, A., Mangubhai, S., Metaxas, A., Parker, L M., Rivera, H E., Silbiger, N J., Smith, N S., … Bates, A E. (2021). Promoting inclusive metrics of success and impact to dismantle a discriminatory reward system in science. *PLOS Biology*, *19*, e3001282.

Deck, J., Gaither, M. R., Ewing, R., Bird, C. E., Davies, N., Meyer, C., Riginos, C., Toonen, R. J., & Garza, J. C. (2017). The Genomic Observatories Metadatabase (GeOMe): A new repository for field and sampling event metadata associated with genetic samples. *PLoS Biology*, *15*, e2002925.

Des Roches, S., Pendleton, L. H., Shapiro, B., & Garza, J. C. (2021). Conserving intraspecific variation for nature's contributions to people. *Nature Ecology & Evolution*, *5*, 574–582.

Díaz, S., Zafra-Calvo, N., Purvis, A., Verburg, P H., Obura, D., Leadley, P., Chaplin-Kramer, R., De Meester, L., Dulloo, E., Martín-López, B., Shaw, M. R, Visconti, P., Broadgate, W., Bruford, M W., Burgess, N D., Cavender-Bares, J., Declerck, F., Fernández-Palacios, J. M., Garibaldi, L A., … Zanne, A E. (2020). Set ambitious goals for biodiversity and sustainability. *Science*, *370*, 411–413.

Dimitrova, M., Meyer, R., Buttigieg, P. L., Georgiev, T., Zhelezov, G., Demirov, S., Smith, V., & Garza, J. C. (2021). A streamlined workflow for conversion, peer review, and publication of genomics metadata as omics data papers. *GigaScience*, *10*, giab034.

Ewing, R. J., & Crandall, E. D. (2020). *geomedb: Functions for fetching "GEOMEdb" data* (R package version 2.0.1). https://CRAN.R-project.org/package=geomedb

Exposito-Alonso, M., Booker, T R., Czech, L., Gillespie, L., Hateley, S., Kyriazis, C C., Lang, P L. M., Leventhal, L., Nogues-Bravo, D., Pagowski, V., Ruffley, M., Spence, J P., Toro Arana, S E., Weiß, C L., & Garza, J. C. (2022). Genetic diversity loss in the Anthropocene. *Science*, *377*, 1431–1435.

Fidler, F., Chee, Y. E., Wintle, B. C., Burgman, M. A., McCarthy, M. A., & Garza, J. C. (2017). Metaresearch for evaluating reproducibility in ecology and evolution. *BioScience*, *67*, 282–289.

Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., Allen, M. J., Angiuoli, S. V., Ashburner, M., Axelrod, N., Baldauf, S., Ballard, S., Boore, J., Cochrane, G., Cole, J., Dawyndt, P., De Vos, P., … Wipat, A. (2008). The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology*, *26*, 541–547.

Gaither, M. R., Gkafas, G. A., de Jong, M., Sarigol, F., Neat, F., Regnier, T., Moore, D., Gröcke, D. R., Hall, N., Liu, X., Kenny, J., Lucaci, A., Hughes, M., Haldenby, S., & Hoelzel, A. R. (2018). Genomics of habitat choice and adaptive evolution in a deep-sea fish. *Nature Ecology & Evolution*, *2*, 680–687.

15231739, 0, Downloaded from https://conbio.onlinelibrary.wiley.com/doi/10.1111/cobi.14061, Wiley Online Library on [17/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Halewood, M., Lopez Noriega, I., Ellis, D., Roa, C., Rouard, M., & Garza, J. C. (2018). Using genomic sequence information to increase conservation and sustainable use of crop diversity and benefit-sharing. *Biopreservation and Biobanking*, *16*, 368–376.

Hendricks, S., Anderson, E C., Antao, T., Bernatchez, L., Forester, B R., Garner, B., Hand, B K., Hohenlohe, P A., Kardos, M., Koop, B., Sethuraman, A., Waples, R S., & Garza, J. C. (2018). Recent advances in conservation and population genomics data analysis. *Evolutionary Applications*, *11*, 1197–1211.

Hinchliff, C E., Smith, S A., Allman, J F., Burleigh, J. G, Chaudhary, R., Coghill, L M., Crandall, K A., Deng, J., Drew, B T., Gazis, R., Gude, K., Hibbett, D S., Katz, L A., Laughinghouse, H. D, Mctavish, E. J., Midford, P E., Owen, C L., Ree, R H., Rees, J A., … Cranston, K A. (2015). Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, 12764–12769.

Hoban, S., Bruford, M., D'urban Jackson, J., Lopes-Fernandes, M., Heuertz, M., Hohenlohe, P A., Paz-Vinas, I., Sjögren-Gulve, P., Segelbacher, G., Vernesi, C., Aitken, S., Bertola, L D., Bloomer, P., Breed, M., Rodríguez-Correa, H., Funk, W. C, Grueber, C E., Hunter, M E., Jaffe, R., … Laikre, L. (2020). Genetic diversity targets and indicators in the CBD post-2020 Global Biodiversity Framework must be improved. *Biological Conservation*, *248*, 108654.

Hoban, S., Bruford, M. W., Funk, W. C., Galbusera, P., Griffith, M. P., Grueber, C. E., Heuertz, M., Hunter, M. E., Hvilsom, C., Stroil, B. K., Kershaw, F., Khoury, C. K., Laikre, L., Lopes-Fernandes, M., Macdonald, A. J., Mergeay, J., Meek, M., Mittan, C., Mukassabi, T. A., … Vernesi, C. (2021). Global commitments to conserving and monitoring genetic diversity are now necessary and feasible. *BioScience*, *73*, 964–976. https://doi.org/10.1093/biosci/biab054

Hoban, S., Archer, F I., Bertola, L D., Bragg, J G., Breed, M F., Bruford, M W., Coleman, M A., Ekblom, R., Funk, W. C, Grueber, C E., Hand, B K., Jaffé, R., Jensen, E., Johnson, J S., Kershaw, F., Liggins, L., Macdonald, A J., Mergeay, J., Miller, J M., … Hunter, M. E. (2022). Global genetic diversity status and trends: Towards a suite of Essential Biodiversity Variables (EBVs) for genetic composition. *Biological Reviews*, *97*, 1511–1538.

Hudson, M., Garrison, N' A., Sterling, R., Caron, N R., Fox, K., Yracheta, J., Anderson, J., Wilcox, P., Arbour, L., Brown, A., Taualii, M., Kukutai, T., Haring, R., Te Aika, B., Baynam, G S., Dearden, P K., Chagné, D., Malhi, R S., Garba, I., … Carroll, S. R. (2020). Rights, interests and expectations: Indigenous perspectives on unrestricted access to genomic data. *Nature Reviews Genetics*, *21*, 377–384.

Kardos, M., Taylor, H. R., Ellegren, H., Luikart, G., & Garza, J. C. (2016). Genomics advances the study of inbreeding depression in the wild. *Evolutionary Applications*, *9*, 1205–1218.

Kumar, A., Anju, T., Kumar, S., Chhapekar, S. S., Sreedharan, S., Singh, S., Choi, S. R., Ramchiary, N., & Garza, J. C. (2021). Integrating omics and gene editing tools for rapid improvement of traditional food plants for diversified and sustainable food security. *International Journal of Molecular Sciences*, *22*, 8093.

Laikre, L. (2010). Genetic diversity is overlooked in international conservation policy implementation. *Conservation Genetics*, *11*, 349–354.

Laikre, L., Hoban, S., Bruford, M W., Segelbacher, G., Allendorf, F W., Gajardo, G., Rodríguez, A. G., Hedrick, P W., Heuertz, M., Hohenlohe, P A., Jaffé, R., Johannesson, K., Liggins, L., Macdonald, A J., Orozcoterwengel, P., Reusch, T B. H., Rodríguez-Correa, H., Russo, I-R M., Ryman, N., & Garza, J. C. (2020). Post-2020 goals overlook genetic diversity. *Science*, *367*, 1083.

Leigh, D. M., Hendry, A. P., Vázquez-Domínguez, E., & Garza, J. C. (2019). Estimated six per cent loss of genetic variation in wild populations since the industrial revolution. *Evolutionary Applications*, *12*, 1505–1512.

Lendemer, J., Thiers, B., Monfils, A. K., Zaspel, J., Ellwood, E. R., Bentley, A., Levan, K., Bates, J., Jennings, D., Contreras, D., Lagomarsino, L., Mabee, P., Ford, L. S., Guralnick, R., Gropp, R. E., Revelez, M., Cobb, N., Seltmann, K., & Garza, J. C. (2020). The extended specimen network: A strategy to enhance us biodiversity collections, promote research and education. *BioScience*, *70*, 23–30.

Letunic, I., & Garza, J. C. (2021). Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*, W293–W296.

Lewin, H A., Robinson, G E., Kress, W. J, Baker, W J., Coddington, J., Crandall, K A., Durbin, R., Edwards, S V., Forest, F., Gilbert, M. T P., Goldstein, M M., Grigoriev, I V., Hackett, K J., Haussler, D., Jarvis, E D., Johnson, W E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., … Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, *115*, 4325–4333.

Leigh, D. M., Van Rees, C. B., Millette, K. L., Breed, M. F., Schmidt, C., Bertola, L. D., Hand, B. K., Hunter, M. E., Jensen, E. L., Kershaw, F., Liggins, L., Luikart, G., Manel, S., Mergeay, J., Miller, J. M., Segelbacher, G., Hoban, S., & Paz-Vinas, I. (2021). Opportunities and challenges of macrogenetic studies. *Nature Reviews Genetics*, 1–17.

Liggins, L., Hudson, M., & Garza, J. C. (2021). Creating space for Indigenous perspectives on access and benefit-sharing: Encouraging researcher use of the Local Contexts Notices. *Molecular Ecology*, *30*, 2477–2482.

Lin, D., Crabtree, J., Dillo, I., Downs, R R., Edmunds, R., Giaretta, D., De Giusti, M., L'hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M E., Mokrane, M., Navale, V., Petters, J., Sierman, B., Sokolova, D V., Stockhause, M., & Garza, J. C. (2020). The TRUST Principles for digital repositories. *Scientific Data*, *7*, 144.

Lou, R. N., Jacobs, A., Wilder, A. P., & Garza, J. C. (2021). A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology*, *30*, 5966–5993.

Marden, E., Abbott, R J., Austerlitz, F., Ortiz-Barrientos, D., Baucom, R S., Bongaerts, P., Bonin, A., Bonneaud, C., Browne, L., Alex Buerkle, C., Caicedo, A L., Coltman, D W., Cruzan, M B., Davison, A., Dewoody, J. A, Dumbrell, A J., Emerson, B C., Fountain-Jones, N M., Gillespie, R., … Rieseberg, L H. (2021). Sharing and reporting benefits from biodiversity research. *Molecular Ecology*, *30*, 1103–1107.

McCartney, A. M., Anderson, J., Liggins, L., Hudson, M. L., Anderson, M. Z., TeAika, B., Geary, J., Cook-Deegan, R., Patel, H. R., & Garza, J. C. (2022). Balancing openness with Indigenous data sovereignty: An opportunity to leave no one behind in the journey to sequence all of life. *Proceedings of the National Academy of Sciences of the United States of America*, *119*, e2115860119.

McNutt, M., Lehnert, K., Hanson, B., Nosek, B. A., Ellison, A. M., & Garza, J. C. (2016). Liberating field science samples and data. *Science*, *351*, 1024–1026.

Michonneau, F., Brown, J. W., & Garza, J. C. (2016). rotl: An R package to interact with the Open Tree of Life data. *Methods in Ecology and Evolution*, *7*, 1476–1481.

Miraldo, A., LI, S., Borregaard, M. K., Flórez-Rodríguez, A., Gopalakrishnan, S., Rizvanovic, M., Wang, Z., Rahbek, C., Marske, K. A., & Garza, J. C. (2016). An Anthropocene map of genetic diversity. *Science*, *353*, 1532–1535.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*, 1422–1425.

O'dea, R E., Parker, T H., Chee, Y. E., Culina, A., Drobniak, S M., Duncan, D H., Fidler, F., Gould, E., Ihle, M., Kelly, C D., Lagisz, M., Roche, D G., Sánchez-Tójar, A., Wilkinson, D P., Wintle, B C., & Garza, J. C. (2021). Towards open, reliable, and transparent ecology and evolutionary biology. *BMC Biology*, *19*, 68.

Pinsky, M. L., & Garza, J. C. (2014). Meta-analysis reveals lower genetic diversity in overfished populations. *Molecular Ecology*, *23*, 29–39.

Pope, L. C., Liggins, L., Keyse, J., Carvalho, S. B., & Garza, J. C. (2015). Not the time or the place: The missing spatio-temporal link in publicly available genetic data. *Molecular Ecology*, *24*, 3802–3809.

Prada, C., Hanna, B., Budd, A F., Woodley, C M., Schmutz, J., Grimwood, J., Iglesias-Prieto, R., Pandolfi, J M., Levitan, D., Johnson, K G., Knowlton, N., Kitano, H., Degiorgio, M., & Garza, J. C. (2016). Empty niches after extinctions increase population sizes of modern corals. *Current Biology*, *26*, 3190–3194.

Quattrini, A. M., Wu, T., Soong, K., Jeng, M. -S., Benayahu, Y., & Garza, J. C. (2019). A next generation approach to species delimitation reveals the role of hybridization in a cryptic species complex of corals. *BMC Evolutionary Biology*, *19*, 116.

Raffard, A., Santoul, F., Cucherousset, J., & Garza, J. C. (2019). The community and ecosystem consequences of intraspecific diversity: A meta-analysis. *Biological Reviews*, *94*, 648–661.

Reithmeier, R., O'leary, L., Zhu, X., Dales, C., Abdulkarim, A., Aquil, A., Brouillard, L., Chang, S., Miller, S., Shi, W., Vu, N., & Garza, J. C. (2019).

15231739, 0, Downloaded from https://conbio.onlinelibrary.wiley.com/doi/10.1111/cobi.14061, Wiley Online Library on [17/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The 10,000 PhDs project at the University of Toronto: Using employment outcome data to inform graduate education. *PLoS ONE*, *14*, e0209898.

Reusch, T. B. H., Ehlers, A., Hämmerli, A., & Garza, J. C. (2005). Ecosystem recovery after climatic extremes enhanced by genotypic diversity. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 2826–2831.

Riginos, C., Crandall, E D., Liggins, L., Gaither, M R., Ewing, R B., Meyer, C., Andrews, K R., Euclide, P T., Titus, B M., Therkildsen, N. O., Salces-Castellano, A., Stewart, L C., Toonen, R J., & Garza, J. C. (2020). Building a global genomics observatory: Using GEOME (the Genomic Observatories Metadatabase) to expedite and improve deposition and retrieval of genetic data and metadata for biodiversity research. *Molecular Ecology Resources*, *20*, 1458–1469.

Roche, D. G., Kruuk, L. E. B., Lanfear, R., & Garza, J. C. (2015). Public data archiving in ecology and evolution: How well are we doing? *PLoS Biology*, *13*, e1002295–12.

Roche, D. G., Lanfear, R., Binning, S. A., Haff, T. M., Schwanz, L. E., Cain, K. E., Kokko, H., Jennions, M. D., & Garza, J. C. (2014). Troubleshooting public data archiving: Suggestions to increase participation. *PLoS Biology*, *12*, e1001779.

Scholz, A. H., Freitag, J., Lyal, C H. C., Sara, R., Cepeda, M. L., Cancio, I., Sett, S., Hufton, A. L., Abebaw, Y., Bansal, K., Benbouza, H., Boga, H. I., Brisse, S., Bruford, M W., Clissold, H., Cochrane, G., Coddington, J A., Deletoille, A. -C., García-Cardona, F., … Overmann, J. (2022). Multilateral benefit-sharing from digital sequence information will support both science and biodiversity conservation. *Nature Communications*, *13*, 1086.

Schriml, L M., Chuvochina, M., Davies, N., Eloe-Fadrosh, E A., Finn, R D., Hugenholtz, P., Hunter, C I., Hurwitz, B L., Kyrpides, N C., Meyer, F., Mizrachi, I. K., Sansone, S. -A., Sutton, G., Tighe, S., & Garza, J. C. (2020). COVID-19 pandemic reveals the peril of ignoring metadata standards. *Scientific Data*, *7*, 188.

Sibbett, B., Rieseberg, L. H., & Garza, J. C. (2020). The Genomic Observatories Metadatabase. *Molecular Ecology Resources*, *20*, 1453–1454.

Stan Development Team. (2021). *RStan: The R interface to Stan* (R package version 2.21.2). https://mc-stan.org/

Thiers, B., Bates, J., Bentley, A. C., Ford, L. S., Jennings, D., Monfils, A. K., Zaspel, J. M., Collins, J. P., Hazbón, M. H., & Garza, J. C. (2021). Implementing a community vision for the future of biodiversity collections. *BioScience*, *71*, 561–563.

Toczydlowski, R H., Liggins, L., Gaither, M R., Anderson, T J., Barton, R L., Berg, J T., Beskid, S G., Davis, B., Delgado, A., Farrell, E., Ghoojaei, M., Himmelsbach, N., Holmes, A E., Queeno, S R., Trinh, T., Weyand, C A., Bradburd, G S., Riginos, C., Toonen, R J., & Garza, J. C. (2021). Poor data stewardship will hinder global genetic diversity surveillance. *Proceedings of the National Academy of Sciences of the United States of America*, *118*, e2107934118.

Vines, T H., Andrew, R L., Bock, D G., Franklin, M T., Gilbert, K J., Kane, N C., Moore, J. -. S., Moyers, B T., Renaut, S., Rennison, D J., Veen, T., & Garza, J. C. (2013). Mandated data archiving greatly improves access to research data. *The FASEB Journal*, *27*, 1304–1308.

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., & Garza, J. C. (2014). The availability of research data declines rapidly with article age. *Current Biology*, *24*, 94–97.

Whitlock, M. C. (2011). Data archiving in ecology and evolution: Best practices. *Trends in Ecology & Evolution*, *26*, 61–65.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., & Garza, J. C. (2012). Darwin Core: An evolving community-developed biodiversity data standard. *PLoS ONE*, *7*, e29715.

Wilder, A. P., Palumbi, S. R., Conover, D. O., & Garza, J. C. (2020). Footprints of local adaptation span hundreds of linked genes in the Atlantic silverside genome. *Evolution Letters*, *4*, 430–443.

Wilkinson, M D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. -W., Da Silva Santos, L. B., Bourne, P E., Bouwman, J., Brookes, A J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C T., Finkers, R., … Mons, B. (2016). Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 1–9.

Willette, D. A., Allendorf, F. W., Barber, P. H., Barshis, D. J., Carpenter, K. E., Crandall, E. D., Cresko, W. A., Fernandez-Silva, I., Matz, M. V., Meyer, E., Santos, M. D., Seeb, L. W., & Garza, J. C. (2014). So, you want to use next-generation sequencing in marine systems? Insight from the Pan-Pacific Advanced Studies Institute. *Bulletin Of Marine Science*, *90*, 79–122.

Winter, D. J. (2017). rentrez: An R package for the NCBI eUtils API. *The R Journal*, *9*, 520–526.

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., Gilbert, J. A., Karsch-Mizrachi, I., Johnston, A., Cochrane, G., Vaughan, R., Hunter, C., Park, J., Morrison, N., Rocca-Serra, P., Sterk, P., Arumugam, M., Bailey, M., Baumgartner, L., … Glöckner, F. O. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology*, *29*, 415–420.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Crandall, E. D., Toczydlowski, R. H., Liggins, L., Holmes, A. E., Ghoojaei, M., Gaither, M. R., Wham, B. E., Pritt, A. L., Noble, C., Anderson, T. J., Barton, R. L., Berg, J. T., Beskid, S. G., Delgado, A., Farrell, E., Himmelsbach, N., Queeno, S. R., Trinh, T., Weyand, C., … Toonen, R. J. (2023). Importance of timely metadata curation to the global surveillance of genetic diversity. *Conservation Biology*, e14061. https://doi.org/10.1111/cobi.14061

15231739, 0, Downloaded from https://conbio.onlinelibrary.wiley.com/doi/10.1111/cobi.14061, Wiley Online Library on [17/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License