Received: 1 March 2020 | Revised: 22 July 2020 | Accepted: 9 September 2020

DOI: 10.1111/1755-0998.13269

FROM THE COVER

MOLECULAR ECOLOGY RESOURCES WILEY

# 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

Cynthia Riginos[1] | Eric D. Crandall[2,3] | Libby Liggins[4] | Michelle R. Gaither[5] | Rodney B. Ewing[6] | Christopher Meyer[7] | Kimberly R. Andrews[8] | Peter T. Euclide[9] | Benjamin M. Titus[10] | Nina Overgaard Therkildsen[11] | Antonia Salces-Castellano[12] | Lucy C. Stewart[13] | Robert J. Toonen[14] | John Deck[15]

[1]School of Biological Sciences, The University of Queensland, St Lucia, Qld, Australia

[2]Department of Biology and Chemistry, California State University, Seaside, CA, USA

[3]Department of Biology, Pennsylvania State University, University Park, PA, USA

[4]School of Natural and Computational Sciences, Massey University, Auckland, New Zealand

[5]Department of Biology, Genomics and Bioinformatics Cluster, The University of Central Florida, Orlando, FL, USA

[6]Biocode LLC, Junction City, OR, USA

[7]Smithsonian Institution, National Museum of Natural History, Washington, DC, USA

[8]Institute for Bioinformatics and Evolutionary Studies (IBEST), University of Idaho, Moscow, ID, USA

[9]Wisconsin Cooperative Fishery Research Unit, College of Natural Resources, University of Wisconsin–Stevens Point, Stevens Point, WI, USA

[10]Division of Invertebrate Zoology, American Museum of Natural History, New York, NY, USA

[11]Department of Natural Resources, Cornell University, Ithaca, NY, USA

[12]Island Ecology and Evolution Research Group, Instituto de Productos Naturales y Agrobiología (IPNA-CSIC), Santa Cruz de Tenerife, Spain

[13]GNS Science, Lower Hutt, New Zealand

[14]Hawai'i Institute of Marine Biology, University of Hawai'i at Mānoa, Kāne'ohe, HI, USA

[15]University of California at Berkeley, Berkeley, CA, USA

**Correspondence**
Cynthia Riginos, School of Biological Sciences, The University of Queensland, St Lucia, Qld., Australia.
Email: c.riginos@uq.edu.au

**Funding information**
Division of Environmental Biology, Grant/Award Number: DEB-1457848 and OISE-1243541

## Abstract

Genetic data represent a relatively new frontier for our understanding of global biodiversity. Ideally, such data should include both organismal DNA-based genotypes and the ecological context where the organisms were sampled. Yet most tools and standards for data deposition focus exclusively either on genetic or ecological attributes. The Genomic Observatories Metadatabase (GEOME: geome-db.org) provides an intuitive solution for maintaining links between genetic data sets stored by the International Nucleotide Sequence Database Collaboration (INSDC) and their associated ecological metadata. GEOME facilitates the deposition of raw genetic data to INSDCs sequence read archive (SRA) while maintaining persistent links to standards-compliant ecological metadata held in the GEOME database. This approach facilitates findable, accessible, interoperable and reusable data archival practices. Moreover,

Cynthia Riginos, Eric D. Crandall, Libby Liggins and Michelle R. Gaither co-first authors.

  

GEOME enables data management solutions for large collaborative groups and expedites batch retrieval of genetic data from the SRA. The article that follows describes how GEOME can enable genuinely open data workflows for researchers in the field of molecular ecology.

**KEYWORDS**
bioinformatics, ecoinformatics, FAIR principles, genomic, open data, reproducible research

## 1 | INTRODUCTION

Genetic data represent the foundations of global biodiversity (Davies et al., 2014) and their value is internationally recognized by the Convention on Biological Diversity (2007). Since the founding of Molecular Ecology in 1992, records linked to DNA sequences have been created at a prodigious and exponentially increasing rate (Cochrane et al., 2016). Motivations for studies and their associated data structures vary widely, with examples including whole genomes of single species, surveys of intraspecific genomic diversity, and characterization of ecological communities. Concurrently, the culture of science is shifting toward an "open" model (Hampton et al., 2015; Powers & Hampton, 2018), prioritizing transparency and reproducibility of research, and enabling data reuse. Indeed, most major governmental funding bodies (e.g., the National Science Foundation, the National Institutes of Health, the European Research Council, the Australian Research Council) and many journals, including major journals in evolution and ecology, require open data deposition (Joint Data Archiving Policy: wiki.datadryad.org/Joint_Data_Archiving_Policy_[JDAP]).

Yet, genetic biodiversity data straddle the traditions of genetics and ecology, whereby community tools and standards for data deposition emphasize either genetic attributes (gene annotations, translated protein, sequencing instrument) or ecological attributes (location, time, environment) but not both. Databases in the genetics tradition (e.g., INSDC, which includes National Center for Biotechnology Information, European Molecular Biology Laboratory, and DNA Data Bank of Japan) are not inherently configured to record the ecological context from where the organism(s) were sampled (i.e.. the sample metadata; but see the Barcode of Life Database [BOLD]). Databases in the ecological tradition (e.g.. Global Biodiversity Information Facility [GBIF] and Ocean Biogeographic Information System), conversely, are poorly configured to accommodate genetic information especially multilocus or genomic data. As a consequence, ecological metadata for genetic biodiversity records are frequently ad hoc, do not match ecological data standards, and are incomplete (Gilbert et al., 2012; Pope et al., 2015).

Ideally, deposition of genetic biodiversity data should follow the FAIR guiding principles (Wilkinson et al., 2019). These principles were developed for data-intensive fields such as genomics with future data reuse as an overarching aim. FAIR guidelines insist that data are findable, accessible, interoperable, and reusable. Thus, they emphasize the importance of unique and persistent identifiers, legibility for both humans and machines, and controlled vocabularies for metadata. Several large initiatives in biology are embracing FAIR, including the US National Institutes of Health's Big Data to Knowledge (commonfund.nih.gov/bd2k).

The Genomic Observatories Metadatabase (GEOME: geome-db.org; Deck et al., 2017) augments long-standing genetic repositories (Leinonen et al., 2010). by providing a straightforward and FAIR-compliant solution for maintaining links between raw genomic data and associated ecological and geographic metadata. When such metadata are uploaded to GEOME, GEOME will provide either a preformatted package to facilitate upload of the genomic data via the SRA portal, or a new and easy-to-use portal for direct upload of genomic data to the SRA. GEOME will then automatically harvest the SRA accession identities, thereby creating permanent links between the genetic data and their metadata. Metadata in GEOME and their linked genomic data in the SRA are thus "findable" via a human user-friendly web interface that supports searching by taxonomy, geographic extents, type of genetic data or programmatically via the R package geomedb (CRAN.R-project.org/package=geomedb). Data are accessible, as the permanent unique identifiers provide a means for persistent access and the application programming interface (API) allows for batch retrieval of metadata and genomic sequences through R functions wrapping SRA toolkit command-line programs. Interoperability is central to GEOME, as metadata follow controlled vocabularies consistent with DarwinCore and MIxS standards (Wieczorek et al., 2012; Yilmaz et al., 2011) and new records on GEOME are pushed onto GBIF. Finally, "reusability" is supported by metadata field choices and a posted data usage policy. Synthetic studies based on reused GEOME data are enabling novel biodiversity and evolutionary insights (Crandall, et al., 2019a; Crandall et al., 2019b; Matias & Riginos, 2018) and complementary research programmes have been motivated through the visualization of data gaps and opportunities for previously underrepresented geographic regions (e.g., Liggins & Arranz, 2018).

GEOME was developed as a collaboration between biologists and data scientists and thus strives to solve challenges common to molecular ecologists while maintaining the aforementioned FAIR criteria. An especially notable feature is the incorporation of maps for visual verification of coordinates before data are uploaded or downloaded (Figure 1). The new portal for direct upload of genomic read data to the SRA should be a timesaver for molecular ecologists by bypassing potentially tedious and complex interim data manipulation steps such as using the Sequin interface

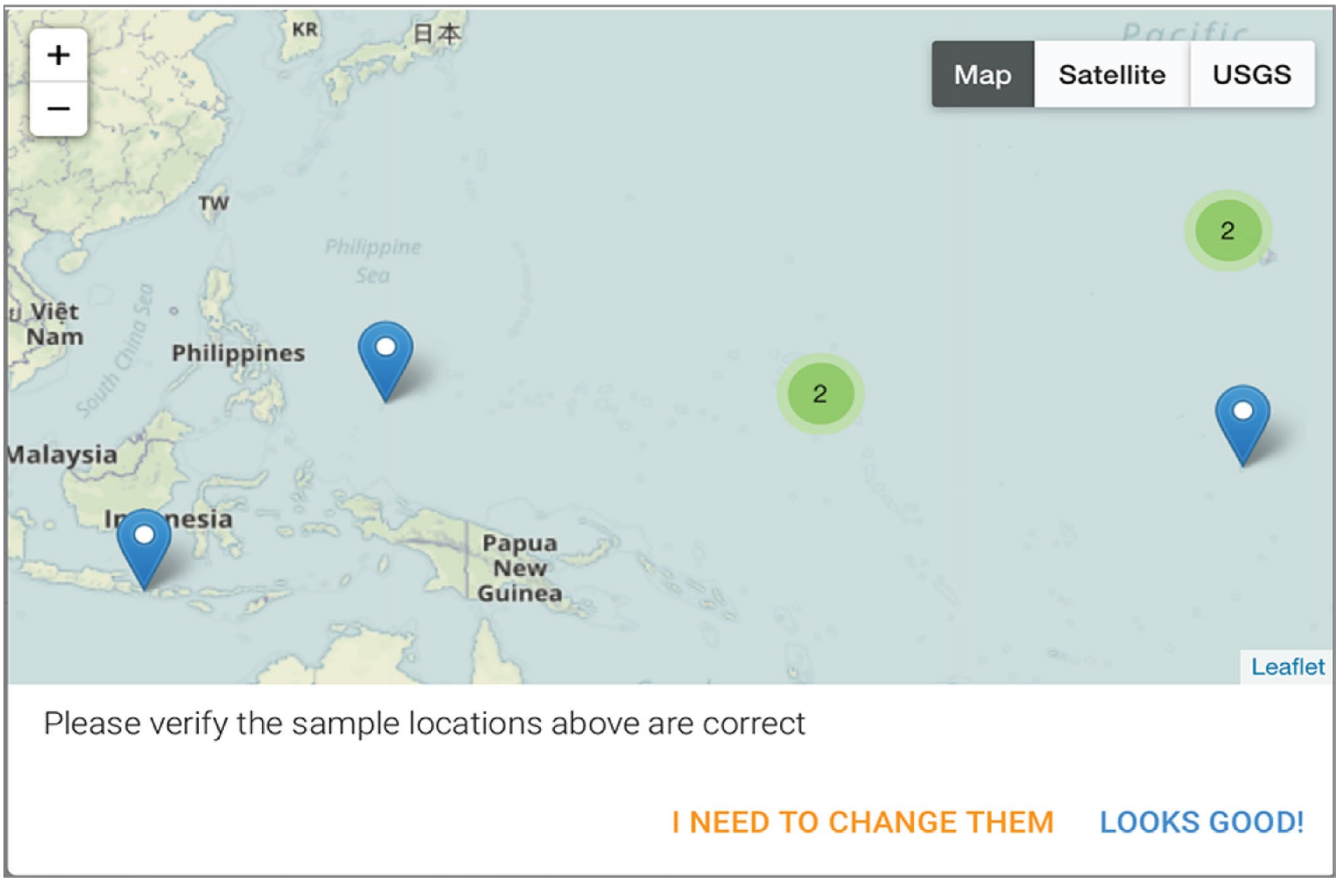


**FIGURE 1** Data validation step includes visualization of sampling locations on a map so that georeferencing errors can be quickly recognized and fixed [Colour figure can be viewed at wileyonlinelibrary.com]

or the newer tbl2asn command-line program. GEOME also aims to balance flexibility and consistency of data and project characteristics. It can now support all DNA sequence-type data including whole genome sequences, reduced representation sequences (such as restriction-site associated DNA), single marker sequences (such as DNA barcoding), and community level sequencing (i.e., metagenomics and metabarcoding). Although metadata fields are maintained as controlled vocabularies, the choices available within GEOME are extensive and include open comment fields; a number of suggested templates are provided based on the nature of the data to aid novice or casual users but can be modified to suit a project's needs. Project submissions to GEOME can be one-off or can now be managed as "expeditions" within a project "team" (Figure 2). Finally, GEOME now allows flexible embargoes that are under the control of the data-submitter. Thus, since the launch of GEOME (Deck et al., 2017) the metadatabase features have been significantly extended to provide increased flexibility and to accommodate a much greater breadth of molecular ecological research needs while simultaneously ensuring that projects follow the FAIR principles.

In this article, we demonstrate the newly enhanced utility of GEOME to molecular ecologists (see Box 1). Our aim is to familiarize potential users with the functionality and benefits in using GEOME across diverse molecular ecological studies and throughout project stages. We provide examples of team and individual research projects (Box 2) that have employed GEOME and highlight the ease and flexibility in its use. We hope that GEOME will help to enable FAIR best practices within the molecular ecology community, so that the molecular ecology community can keep pace with evolving expectations in research conduct and the stewardship of data resources.

## 2 | TEAM PROJECTS LEVERAGE THE FULL FUNCTIONALITY OF GEOME

A core aim of GEOME is to enable international collaboration among large teams by helping individual research groups to contribute data towards a larger team project (Figure 2). Teams define their own sets of rules, attributes, and controlled vocabularies. Thus, each contribution ("expedition") will be submitted and validated in alignment with the team's self-defined criteria and new data automatically added to the team's repository. Indeed, GEOMEs genesis arose from the data-management needs of two distinct projects, the Mo'orea Biocode Project (biocode.swala.org/) and the Diversity of the Indo-Pacific Network (diversityindopacific.net/). We describe these two projects here to highlight how GEOME supports structured collaboration and data-management.

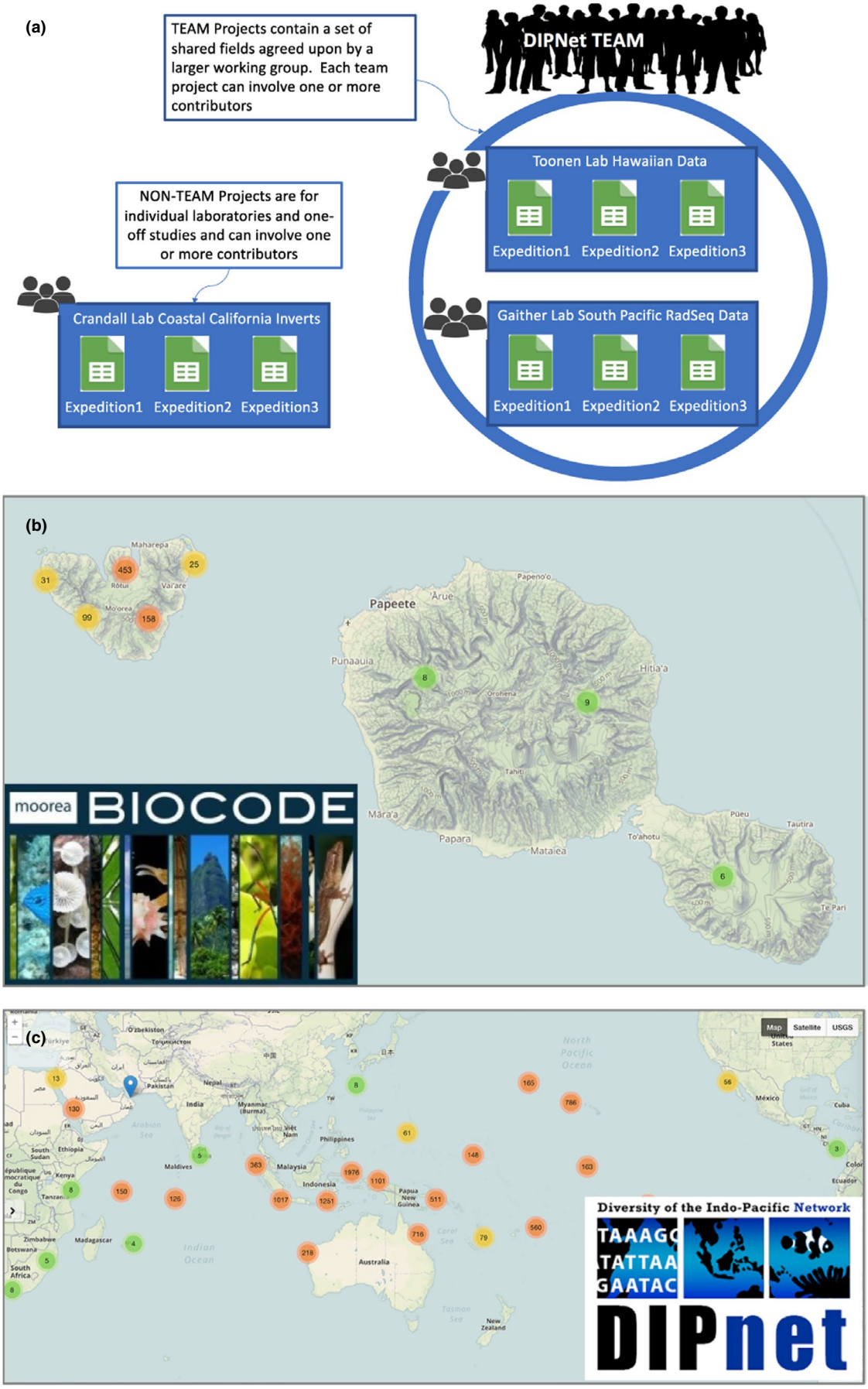
(a)
TEAM Projects contain a set of shared fields agreed upon by a larger working group. Each team project can involve one or more contributors
DIPNet TEAM
NON-TEAM Projects are for individual laboratories and one-off studies and can involve one or more contributors
Toonen Lab Hawaiian Data
Expedition1 Expedition2 Expedition3
Crandall Lab Coastal California Inverts
Expedition1 Expedition2 Expedition3
Gaither Lab South Pacific RadSeq Data
Expedition1 Expedition2 Expedition3
(b)
Maharepa
Rōtui
Vaʻarae
Moʻorea
Papeete
ʻArue
Papenoʻo
Punaauia
Hitiaʻa
Orohena
Tahiti
Māraʻa
Papara
Mataiea
Pueu
Tautira
Toʻahotu
Teʻ Pari
moorea BIOCODE
(c)
Map
Satellite
USGS
Diversity of the Indo-Pacific Network
TAAAGG
TATTAA
GAATAC
DIPnet

**FIGURE 2** Team projects. (a) The Team project configuration allows collaborating research groups to standardize metadata fields to suit that project. (b) Mo'orea Biocode Project biosamples contain mixed assemblages from marine, fresh-water, and terrestrial locations and barcode sequences. (c) The DIPnet project contains mitochondrial and nuclear DNA sequences (primarily RADseq) identified to individual marine organisms

**Box 1 Eight reasons why GEOME could be useful to you**

1. **Built for purpose and aligned with standards:** GEOME guides the generation of appropriate metadata templates using standard and quality assured fields with flexibility to meet the needs of a specific study, the expectations of colleagues, and requirements of publishers and funders. Moreover, GEOME provides a ready-made data-management plan, with predefined yet customizable standards for genetic data and its associated metadata. For the conscientious scientist, GEOME provides the infrastructure and assurance of adhering to community best practices.
2. **Facilitates data management across research groups:** For research groups or multi-institution consortia, GEOME now provides a coordinated team-wide platform to maintain and curate collections of biological samples, genomic DNA, and genetic data that is standardized for all team members. Teams may customize their data standards, controlled vocabularies, and validation rules. Projects built as part of a team may be kept private and then made public upon publication.
3. **Facilitates data management throughout project stages:** Once a project or team decides upon their study's data standards, a GEOME metadata template can be exported (as a spreadsheet or text file). This template can be completed from remote locations and does not require an Internet connection. Metadata pertaining to these biological collections can be uploaded immediately, and updated as the project advances, for example, enabling information exchange regarding available sample tissues (see reason number 5). Finally, GEOME facilitates the upload of sequences to the SRA through a dedicated portal, assuring the use of metadata fields required by the SRA.
4. **Enhances data accessibility:** GEOME now has in excess of 100 metadata fields, with 90% approved by a standards organization (e.g., Genomic Standards Consortium or Biodiversity Information Standards) that can be used in structured queries of the entire metadatabase through the web interface or custom R package. Query results can be fed to *geomedb* R functions that wrap SRA toolkit programs and allow rapid download of relevant genetic data. The opportunity to extensively query existing data increases the potential for data submitted to GEOME to be re-used with appropriate attribution.
5. **Builds community and identifies new collaboration opportunities:** When data are made public, GEOME now provides peers rapid visualization of research intentions and progress before publishing (i.e., available genetic data and tissues collected) through an interactive, queryable map, table presentation, and dynamic dashboard. In the same open-science ethos as GitHub and arXiv, GEOME enables viewing of research collections and progress in real time, avoiding the publication lag. Thus, there is the opportunity to join teams and discover colleagues interested in the same taxonomic groups, geographic regions, or types of studies. For genetic nonexperts, GEOME provides an intuitive means to visualize and interface with studies and collections in progress, allowing managers, communities, and students to contact data creators in real time.
6. **Protects privacy:** For those who work with sensitive genetic data, GEOME can point to the embargoed or secured genetic data. The "coordinateUncertaintyInMeters" field allows users to obscure precise locations of endangered or culturally sensitive species. This ensures that such data are still "findable", and if/when released, will ensure the data are interoperable and the study reproducible. Thus, for internal and external audits of sample collections for reasons of biosecurity, permitting, and safekeeping of endangered species, or other sensitive collections, GEOME assures collection integrity, longevity, and visibility (if desired).
7. **Respects rights:** For scientists that work with indigenous communities, GEOME now provides metadata fields (e.g., "traditionalKnowledgeNotice") that can be used to indicate that there may be cultural sensitivities related to the sample and/or genetic data that would need to be considered prior to any access or re-use of the data. GEOME's Data Usage Policy specifically addresses concerns raised by the Convention on Biological Diversity and its attendant Bonn Guidelines and Nagoya Protocol.
8. **Facilitates data attribution and citations:** GEOME has three levels of data attribution. At the most atomic level, persistent and resolvable archival resource key identifiers are registered for every material sample, collecting event, tissue, and diagnostic observation registered in the system. Expeditions (which comprises collections of samples and events) are also assigned archival resource key identifiers, containing a project title, description, and username. Finally, GEOME enables users to couple DOIs minted elsewhere (such as DOI's for associated publications), at the project level.

## Box 2 Individual projects and new user experiences

**Reduced representation sequencing**

*RADseq of bighorn sheep and Mexican grey wolves:* The data we uploaded to GEOME were from a project focused on developing a bioinformatic pipeline for parentage assignment with RADseq data (Andrews et al., 2018). The pipeline included options for analysis with or without a reference genome, as well as options for identifying smaller subsets of markers for developing SNP genotyping assays with high power and accuracy for parentage assignment. We tested this pipeline using blood and tissue samples from Mexican grey wolves of known parentage and bighorn sheep of unknown parentage, and compared results generated using RADseq and microsatellites. RADseq outperformed microsatellites and had strong power and accuracy, even with subsets of fewer than 300 SNPs. The GEOME platform provided a straightforward method for organizing and uploading metadata, and a user-friendly format for searching and viewing uploaded metadata. When compared to uploading to NCBI, I found GEOME to be more intuitive and it provided more options for uploading ecological information. I found the video tutorials to be particularly helpful. - *Kimberly Andrews*.

*GT-seq of walleye fish:* We used GEOME to store metadata and produce NCBI upload files for the 288 tissue samples testing a new GT-seq panel (genotyping-in-thousands: Campbell et al., 2014) for walleye (*Sander vitreus*). The panel was designed to include markers that maximize parentage estimates and the ability to discriminate between local populations to inform management and improve stocking practices in Wisconsin and Minnesota, USA. This GT-seq panel is the first of its kind in the region and is already being used throughout Wisconsin to evaluate genetic variation of walleye in new systems. Therefore, ensuring that the sequence data for the baseline samples used to test the panel were catalogued and publicly available was important for future evaluation and use of this new genetic resource. Like any new database, GEOME has a learning curve, but unlike NCBI, the GEOME website was designed with the purpose of being user friendly. The database building process guides the user with intuitive step-by-step descriptions that allowed me to go from creating a new account to downloading a csv database template in a short amount of time. When I have uploaded sequence data to the NCBI SRA in the past, I struggled to get all of the necessary data formatted correctly for the NCBI system. However, because GEOME constructs the SRA files for you, it was easier to pass my data through GEOME's interface to construct SRA files than it was to go directly to try to make them myself. - *Peter Euclide*.

*UCEs of sea anemones:* Giant tropical sea anemones that serve as hosts to mutualistic clownfishes (or anemonefishes) have been historically understudied: the 10 nominally described host taxa have only been described morphologically. Recently, I (Titus et al., 2019) recovered preliminary molecular evidence that some host taxa may be cryptic species complexes. Now, using fine-scale sampling across coral atoll habitats in the Maldives (outer fore reef, reef flat, inner lagoon) I am testing whether the magnificent anemone, *Heteractis magnifica*, is a cryptic species complex that has undergone ecological speciation. I am using recently developed bait-capture probes targeting Ultra Conserved Elements (UCEs) and exon loci for Class Anthozoa (Quattrini et al., 2018) to test for cryptic speciation in this group. Uploading my georeferenced bait-capture data to the GEOME database was straight-forward and far simpler than uploading data directly to NCBI's SRA. GEOME's best feature is that it automatically generates the NCBI SRA submission files needed to archive the raw genomic data, which is a huge time saver. Additionally, because the geographic localities of the data are required for GEOME database submissions, it will make finding and incorporating publicly available data into future studies much easier. - *Benjamin Titus*.

**Whole genome sequences**

*WGS of Atlantic cod:* I first used GEOME for a data set consisting of low-coverage whole genome sequencing of 306 Atlantic cod collected across most active spawning sites in the Gulf of Maine (USA) and surrounding waters (a total of 20 locations). The goal of this study was to examine fine-scale population structure and search for evidence of adaptive divergence between spawning groups. The whole-genome analysis provided unprecedented resolution for characterizing geographic differentiation and revealed strong signatures of divergent selection that had not been detectable in prior studies using less dense genome sampling like microsatellites, RAD-seq, and SNP arrays (Clucas et al., 2019). Our raw sequence data amounted to almost 600 Gb and GEOME provided a really useful framework for organizing and archiving the metadata behind our 462 fastq files prior to SRA submission. Multiple different collaborators had collected our samples over several years, and GEOME's built-in validation rules and helpful error messages made it much easier to detect inconsistencies in data recording conventions across different merged metadata files (I especially love the mapping of sampling sites to quickly sanity-check geographic coordinates). GEOME's list of controlled vocabulary sample attributes will also be really useful for generating templates to ensure a more comprehensive and consistent metadata collection in future projects. - *Nina Overgaard Therkildsen*.

**DNA barcoding and metagenomics**

*Barcoding of codistributed beetles from Tenerife:* Our data consist of mitochondrial DNA (COI) sequences amplified from beetles. We are using a community-level approach to investigate the consequences of climate across limited spatial scales (a few kilometres) with

**Box 2 (Continued)**

respect to geographic isolation and incipient speciation. In this study (Salces-Castellano et al., 2020), we demonstrate congruent diversification with gene flow across different species, mediated by Quaternary climate oscillations that have facilitated a dynamic of isolation and secondary contact. Uploading data to GEOME was a simple and fast process, in which a large database (mtDNA sequences from 1,787 individuals sampled from 10 sampling sites) with associated information was submitted using csv and FASTA files. I have not yet had the opportunity to use NCBI, so I cannot compare, but I would use GEOME again because of its easy to use interface and useful data handling features. It is quite intuitive, and for any questions the "help" section is very useful. I also found the feedback and help from the developers to be very effective. - *Antonia Salces Castellano.*

*Microbial eDNA metagenomic sequencing:* I was first introduced to GEOME by the Ira Moana Project, which is using GEOME to facilitate sharing of metadata associated with marine genetic and genomic samples in Aotearoa New Zealand. I used GEOME's metadata template and upload interface for two research projects, which both included 16S rRNA barcodes and eDNA metagenome datasets generated from environmental samples taken across terrestrial and marine sites. The first project was a survey of microbial diversity in hydrothermal environments along the Kermadec Arc subduction zone, for which we generated nine 16S rRNA libraries from four expeditions (Stewart, et al., 2018; Stewart et al., 2018). The second project investigated the potential impact of gas hydrate extraction on methane-oxidising microbial communities and included twenty-six 16S rRNA libraries and three metagenomes that were generated from one sampling expedition (Stewart et al., 2020). As an early career researcher moving between institutions, it is important to me that my data is archived in a contextual and persisting framework. GEOME makes it easy to connect data sets that are generated across different expeditions and published as separate research articles but are part of an overarching survey or larger research programme. It also provides the possibility that my data will be complemented and put into a wider geographic context by other researchers – an opportunity not easily realized or possible through use of archives such as the SRA alone. I found the interface very easy to use because of its compatibility with the MIxS metadata standard, and I intend to use the template generator when planning data collection for future projects. - *Lucy C. Stewart* (Figure 3).

## 2.1 | Mo'orea Biocode Project

Funded primarily by the Gordon and Betty Moore Foundation, the Mo'orea Biocode Project (MBP) has endeavored to genetically characterize the entire macrobiota of a tropical ecosystem: algae, fungi, plants, and animals; marine, freshwater, and terrestrial; from undersea to the tops of the mountains by creating a voucher-based DNA barcode library for all species encountered. Led by an international team of researchers from the University of California at Berkeley, the Smithsonian Institution, and the French National Center for Scientific Research, through both French and American field stations on the French Polynesian island of Mo'orea, the project engaged over 200 persons from more than 30 countries over the course of seven years. This collaborative project created an informatics challenge pertaining to tracking and updating both local and distributed data systems. A field management information system was developed based on a set of core standards (Deck et al., 2012) that linked through a tissue identifier to a laboratory information management system (Geneious laboratory information management system; Parker et al., 2012) that enabled near real-time capture of an agreed upon set of minimum metadata fields, augmented by voucher photos. These data could be updated at any time and provided a clearinghouse for downstream holding institutions and repositories. A key strategy was to enforce this core set of standards for all MBP participants, focused first on the event data, rather than the voucher, in order to lock down critical metadata early in the data maturation chain. This first information management system served a sufficiently broad set of users to cover the majority of workflows encountered in standard biodiversity expeditions and became the prototype of the current GEOME toolkit described herein.

## 2.2 | Diversity of the Indo-Pacific Network

The Diversity of the Indo-Pacific Network (DIPnet) is a National Science Foundation funded Research Coordination Network founded in 2012 in an effort to create greater synchronization and collaboration among marine researchers working in the Indo-Pacific Ocean. It started as a National Evolutionary Synthesis Center working group of collaborating scientists who aimed to compile decades of existing molecular data into a single database for asking evolutionary questions at large spatial and taxonomic scales. As a first step in the larger goal of aggregating all Indo-Pacific molecular data, the working group amassed and curated mitochondrial sequence data and agreed upon 66 relevant metadata descriptors, with a final dataset of nearly 39,000 samples across > 250 species (Crandall, et al., 2019a). Initially these data were stored in ASCII file formats (text tables for metadata and FASTA files for associated sequences) that were then converted into the first GEOME project. DIPnet has grown to include over 60 members from more than two dozen countries and has continued to support new GEOME functions such as the upload portal and dashboard. DIPnet has now entered a second phase of data deposition as member laboratories contribute high throughput sequencing data such as restriction-site associated DNA

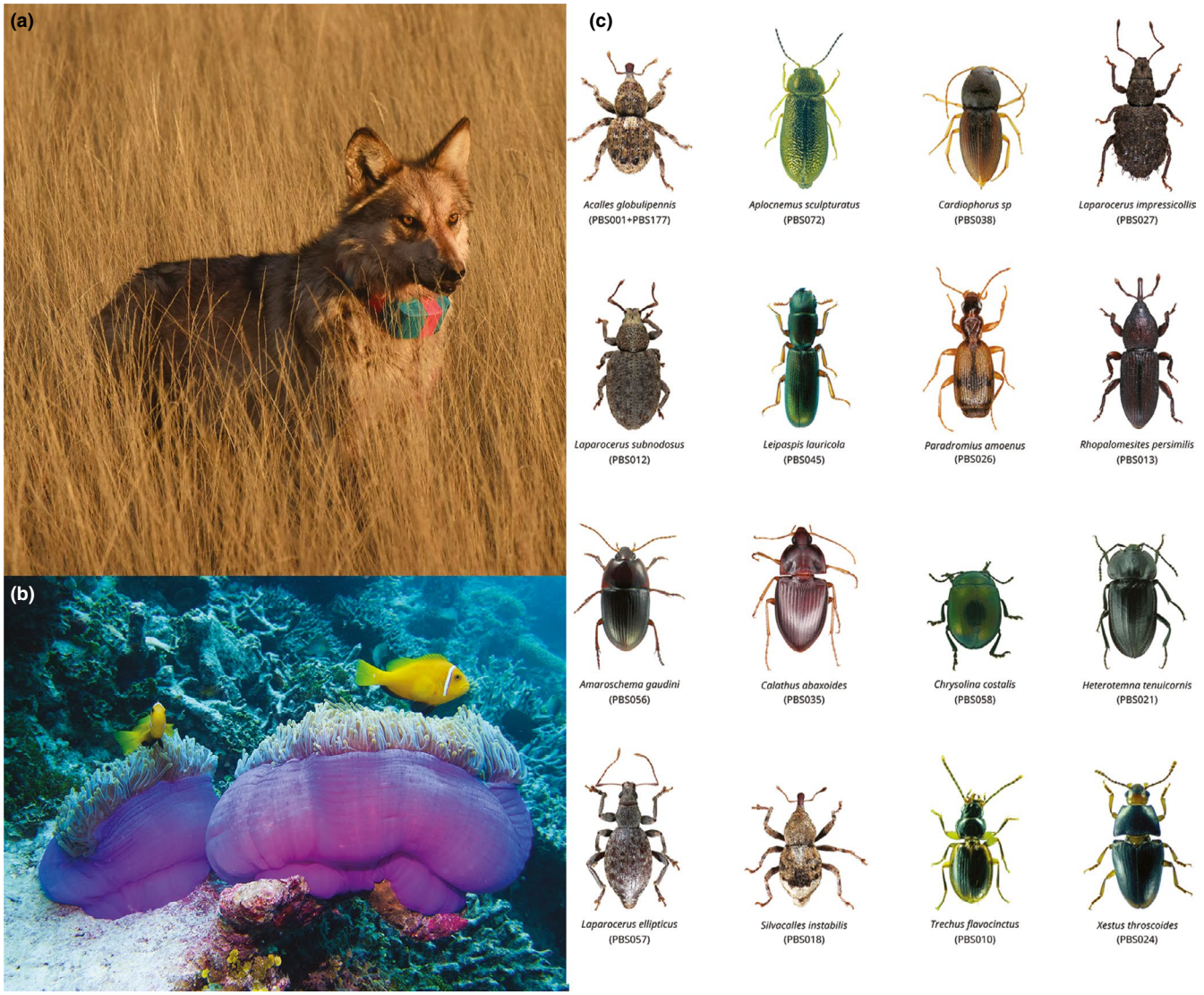


**FIGURE 3** Examples of some of the organisms studied by new GEOME users. (a) Mexican wolf. (b) Magnificent sea anemone. (c) Beetles from Tenerife. Photo credits: Mexican Wolf Interagency Field Team, B. Titus & A. Salces-Castellano

sequencing data sets for population genomic studies. The DIPnet project now also includes the metadata for 800 SRA biosamples and is growing.

## 3 | DOWNSTREAM USES: RETRIEVING LINKED METADATA AND GENOTYPES VIA GEOME

Properly archived scientific data that meet FAIR criteria have far more value than just fulfilling their original purpose (Vision, 2010). Such data are essential for verifying and reproducing results, contributing to meta-analysis, developing new questions and new proposals, and providing case studies for instruction (Whitlock, 2011). Whereas there are numerous "sticks" to compel data archiving in the form of mandates from journals and funding bodies, there are unfortunately few "carrots" that reward researchers for this seemingly altruistic behavior (Roche et al., 2014). As described above, we have shown how GEOME lowers barriers for submitting genetic data and attendant metadata. But, GEOME also facilitates downstream data usage as well as responsible attribution and recognition of data contributors.

Because GEOME endeavors to store ecologically and environmentally relevant metadata that are permanently linked to genetic data archived at the SRA, it makes no attempt to "reinvent the wheel" of genetic data archives created by INSDC (Cochrane et al., 2016; Leinonen et al., 2010), but instead augments INSDC. Thus, while all genetic data submitted through GEOME remain BLAST-able, they are now also searchable by locality, bounding-box coordinates, sampling year, principal investigator, and by a large selection of DarwinCore metadata fields. These queries can be conducted on the GEOME website, with easily browsable results presented in map or table form, or they can be conducted programmatically through the geomedb R package, or the API. The R package also provides

functions that wrap the fastq-dump, fasterq-dump, and prefetch functions in the SRA toolkit (Leinonen et al., 2009) and can seamlessly download all SRA data associated with any GEOME query. This enhanced search and retrieval adds great value to genetic data accessioned to the SRA through GEOME.

GEOME also facilitates proper attribution and credit for biological samples and genetic data. Every sampling expedition, event, sample, and tissue on GEOME has a unique landing page, with a globally unique archival resource key that will be persistently resolvable, for example, by the Names to Things resolver (n2t.net). Moreover, every GEOME project has a linked unique, citable, digital object identifier (DOI), while the "associatedReferences" field can store links to attendant publications from the data. The updated GEOME data dashboard efficiently summarizes all of this information for publicly available projects such that it could be consulted as a metric of research progress similar to the way Google Scholar is used for scholarly output.

Finally, all GEOME metadata are subject to an updated data-usage policy, developed in consultation with an expert in international law, which explicitly addresses the Convention on Biological Diversity (1992). The agreement ensures that uploaded data describe samples that were lawfully collected under appropriate research permits and that downloaded data remain in the public domain and are appropriately cited to their source.

## 4 | CONCLUSIONS AND FUTURE DIRECTIONS

In the burgeoning ecosystem of tools supporting open science, GEOME fills a unique niche. GEOME offers a bridge between long-established standards for DNA sequence archiving via INSDC and for biodiversity records via GBIF. Simultaneously, GEOME complements other useful tools for biodiversity genetics. For example, BOLD is aimed towards species identification but does not support extensive metadata fields and project customization. IMapGene (macroecology.ku.dk/resources/imapgenes/) provides data access and visualization for terrestrial mtDNA sequences compiled and described by Miraldo et al. (2016) but is not a platform for new data deposition. MacroPopGen (Lawrence et al., 2019) from vertebrate population genetic studies and IntraBioDiv for alpine plant species of Europe (www.wsl.ch/en/projects/intrabiodiv.html) are both static spreadsheets containing summary statistics. Sample sharing is facilitated by Otlet (otlet.io) but this portal does not link to derived data. Whereas these resources support specific data needs, none overlap in substance with GEOME.

GEOME has been designed for maximal project flexibility within the constraints of maintaining controlled vocabularies. By focusing on DNA sequence-based data, data across projects can be easily combined by end users (for this reason, called genotypes that are difficult to confirm across laboratories such as microsatellites and SNPs are not directly supported on GEOME, although the functionality to point to derived datasets is in development). At present, most GEOME projects consist of samples based on single individuals associated with one or more sequences (e.g., one-to-one for barcoding and Sanger sequences and one-to-many for short reads in genomic projects). Other configurations such as many-to-many for pools of individuals (poolSeq, eDNA, metabarcoding) and hierarchies such as individual hosts and their parasites or hosts and their epiphytic communities can also be accommodated using linkages through the materialSampleID field (and/or unique tissue identifiers corresponding to that materialSampleID). Whereas GEOME currently facilitates new submissions to NCBI's SRA via a direct portal, submission of barcoding data is not yet so streamlined (but see Box 2 for user case studies) and future development would allow direct submission of GEOME data to either BOLD or NCBI's nucleotide database. Right now, Sanger data (i.e., single nucleotide sequences) can be stored as text directly in the GEOME database. Using the Laboratory Information Management System plug-in for Biomatter's Geneious platform can facilitate preparation of Sanger data for submission to the NCBI Nucleotide database. Alternatively, it is recommended that researchers submit their metadata to GEOME first, obtain a unique ID, and embed that ID in a subsequent NCBI or BOLD submission.

Another area of important future development for GEOME will be to continue to innovate tools and metrics for data provenance. Ensuring that data creators receive credit is an important aspect of community buy-in for open data standards (Kaye et al., 2009). In the short term, we are using the "associatedReference" field to provide links to a unique DOI for the GEOME project and DOIs for publications describing the data in line with Data Citation Principles (Data Citation Synthesis Group, 2014). In the longer term, we are exploring using an evolving block chain that would support complete analytics for contributed data, including citations for data usage at the individual sample level.

Complementing the adherence to FAIR principles, GEOME is actively developing its infrastructure to further acknowledge indigenous rights by considering the CARE principles for indigenous data governance (Collective benefit; Authority to control; Responsibility; Ethics: www.gida-global.org/care#). Sample provenance, indigenous names, and value of the samples to indigenous communities can already be accommodated using existing metadata fields within GEOME. For example, in using existing and standard metadata fields to incorporate indigenous provenance and value, the New Zealand based Ira Moana Project (www.massey.ac.nz/iramoana, hosted by GEOME) is ensuring that the indigenous Māori origin of samples is not erased from interoperable database structures. Furthermore, in collaboration with Local Contexts (localcontexts.org/) and Te Mana Rauranga (the Māori Data Sovereignty Network, www.temanararaunga.maori.nz/) GEOME is now beta-testing the capacity for researchers to add Traditional Knowledge, Biocultural Notices, and Labels as metadata for DNA sequence data (www.enrich-hub.org/bc-labels). Researchers apply the Notices to signal that there are accompanying indigenous rights that need further attention and Labels (applied by indigenous communities) provide provenance information and community expectations for future data use. The

development of GEOME's infrastructure to host Notices and Labels is a first for a biological resource and for genetic data, establishing new ethical standards in this research community aligned with international expectations for fair and equitable sharing of benefits arising from genetic resources (i.e., Nagoya Protocol).

Open data and data reuse are principles long embraced by the genetics community and made possible by the INSDC repositories. Likewise, repositories of species occurrence data such as GBIF have been transformational for ecological studies. With the exponential increase in genomic data from free-living populations and species, molecular ecologists can now pursue expansive research programmes to yield deeper insights into how and why alleles, species, communities, and ecosystems are arrayed in space and time (Crandall, et al., 2019a; Crandall, et al., 2019b; Gratton et al., 2017; Manel et al., 2020; Millette et al., 2020; Miraldo et al., 2016; Salces-Castellano et al., 2020; Smith et al., 2014) and toward the conservation of biodiversity (e.g., supporting the Group on Earth Observations Biodiversity Observation Networks; geobon.org/). These ambitious enterprises, however, necessitate that the metadata providing ecological context remain linked to DNA sequences. Moreover, given pervasive ongoing global environmental change, the genomic data collected now will undoubtedly provide a contrast against future measures of biodiversity. GEOME attempts to maximize the value of this genomic observatory approach to DNA data (Davies et al., 2014) by creating permanent links to the contextual metadata. Thus, GEOME increases reusability and reproducibility, enables collaboration across space and time, and incorporates best and emerging principles for genetic data deposition.

## ACKNOWLEDGEMENTS

Development of GEOME has been supported by NESCent, the National Evolutionary Synthesis Center (2012 Catalysis Meeting to EDC and CR; 2013-2014 Working Group to CR, EDC and RJT); the National Science Foundation (OISE-1243541 to CM, DEB-0956426 to CM and JD, and DEB-1457848 to EDC and RJT); the Gordon and Betty Moore Foundation (CM and JD), the Berkeley Natural History Museums, and contracts from the Smithsonian National Museum of Natural History as components of the Global ARMS Programme, Global Genome Initiative and the SI Barcoding Network. The implementation of some metadata fields in GEOME were jointly initiated by the Ira Moana – Genes of the Sea – Project (www.massey.ac.nz/iramoana) and the project 'Te Tuākiri o te Tāonga: Recognizing Indigenous Interests in Genetic Resources' both supported by Catalyst Seeding funds provided by the New Zealand Ministry of Business, Innovation and Employment and administered by the Royal Society Te Apārangi (17-MAU-309-CSG to LL, CR, EDC, MRG, JD, RJT and seven others; and 19-UOW-008-CSG to Maui Hudson and Jane Anderson, respectively), as well as a Massey University Research Fund to L.L. We thank Dr Karen Chambers and Dr Benjamin Sibbett from Wiley for providing encouragement and suggestions, alongside reviewers including Dr Shawn Narum and Dr Nick Fountain-Jones. NESCent, we miss you.

## AUTHOR CONTRIBUTIONS

Core concepts for GEOME emerged from previous projects where E.D.C., C.R., C.M., M.R.G., R.J.T., and J.D. were central participants in planning discussions. J.D., and R.E. wrote the main code with web-based interface with early beta testers including E.D.C., M.R.G., C.M., and C.R. R.J.E., J.D., and E.D.C. wrote the geomedb R package for querying GEOME and extracting SRA data. C.R., E.D.C., L.L., and M.R.G. conceived the present manuscript and took the lead in writing, organising, and editing with written contributions by C.M., and J.D., K.R.A., P.T.E., B.M.T., N.O.T., A.S.C., and L.C.S. uploaded data to GEOME as naïve users, described their experiences in Box 2, and provided useful feedback that improved the user interface. All authors read and approved the paper.

## DATA AVAILABILITY STATEMENT

The GEOME portal can be accessed at geome-db.org. Instructions for using GEOME and general additional information including FAQs: geome-db.org/about. Data usage policy for GEOME based on FAIR principles: geome-db.org/about#dataPolicy. Ira Moana project, incorporating CARE principles with genetic data: www.massey.ac.nz/iramoana/ and Ira Moana Project FAQs.

## ORCID

*Cynthia Riginos* https://orcid.org/0000-0002-5485-4197
*Libby Liggins* https://orcid.org/0000-0003-1143-2346
*Michelle R. Gaither* https://orcid.org/0000-0002-0371-5621
*Kimberly R. Andrews* https://orcid.org/0000-0003-4721-1924
*Peter T. Euclide* https://orcid.org/0000-0002-1212-0435
*Benjamin M. Titus* https://orcid.org/0000-0002-0401-1570
*Nina Overgaard Therkildsen* https://orcid.org/0000-0002-6591-591X
*Lucy C. Stewart* https://orcid.org/0000-0001-7352-3329
*John Deck* https://orcid.org/0000-0002-5905-1617

## REFERENCES

Andrews, K. R., Adams, J. R., Cassirer, E. F., Plowright, R. K., Gardner, C., & Dwire, M. (2018). A bioinformatic pipeline for identifying informative SNP panels for parentage assignment from RAD seq data. *Molecular Ecology Resources*, *18*(6), 1263–1281. https://doi.org/10.1111/1755-0998.12910.

Campbell, N. R., Harmon, S. A., & Narum, S. R. (2014). Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. *Molecular Ecology Resources*, *15*(4), 855–867. https://doi.org/10.1111/1755-0998.12357

Clucas, G. V., Lou, R. N., Therkildsen, N. O., & Kovach, A. I. (2019). Novel signals of adaptive genetic variation in northwestern Atlantic cod revealed by whole-genome sequencing. *Evolutionary Applications*, *12*(10), 1971–1987. https://doi.org/10.1111/eva.12861

Cochrane, G., Karsch-Mizrachi, I., Takagi, T., & International Nucleotide Sequence Database Collaboration (2016). The international nucleotide sequence database collaboration. *Nucleic Acids Research*, *44*(D1), D48–D50. https://doi.org/10.1093/nar/gkv1323

Convention on Biological Diversity of 5 June (1992). *(1760 U.N.T.S. 69)*.

Crandall, E. D., Riginos, C., Bird, C. E., Liggins, L., Treml, E., Beger, M., Barber, P. H., Connolly, S. R., Cowman, P. F., DiBattista, J. D., Eble, J. A., Magnuson, S. F., Horne, J. B., Kochzius, M., Lessios, H. A., Liu,

S. Y. V., Ludt, W. B., Madduppa, H., Pandolfi, J. M., … Gaither, M. R. (2019a). The molecular biogeography of the Indo-Pacific: Testing hypotheses with multispecies genetic patterns. *Global Ecology and Biogeography*, *58*(5), 403–418. https://doi.org/10.1111/geb.12905

Crandall, E. D., Toonen, R. J., Laboratory, T. B., & Selkoe, K. A. (2019b). A coalescent sampler successfully detects biologically meaningful population structure overlooked by F-statistics. *Evolutionary Applications*, *12*(2), 255–265. https://doi.org/10.1111/eva.12712

Data Citation Synthesis Group (2014). In M. Martone (Ed.) *Joint Declaration of Data Citation Principles*. FORCE11. https://doi.org/10.25490/a97f-egyk

Davies, N., Field, D., Amaral-Zettler, L., Clark, M. S., Deck, J., Drummond, A., Faith, D. P., Geller, J., Gilbert, J., Glöckner, F. O., Hirsch, P. R., Leong, J.-A., Meyer, C., Obst, M., Planes, S., Scholin, C., Vogler, A. P., Gates, R. D., Toonen, R., … Zingone, A. (2014). The founding charter of the Genomic Observatories Network. *GigaScience*, *3*(1), 2. https://doi.org/10.1186/2047-217X-3-2

Deck, J., Gaither, M. R., Ewing, R., Bird, C. E., Davies, N., Meyer, C., Riginos, C., Toonen, R. J., & Crandall, E. D. (2017). The Genomic Observatories Metadatabase (GeOMe): A new repository for field and sampling event metadata associated with genetic samples. *PLOS Biology*, *15*(8), e2002925. https://doi.org/10.1371/journal.pbio.2002925

Deck, J., Gross, J., Stones-Havas, S., Davies, N., Shapley, R., & Meyer, C. (2012). Field information management systems for DNA barcoding. In W. J. Kress, & D. L. Erickson (Eds.), *DNA barcodes: Methods and Protocols* (Vol. *858*, pp. 255–267). Springer. https://doi.org/10.1007/978-1-61779-591-6_12

Gilbert, K. J., Andrew, R. L., Bock, D. G., Franklin, M. T., Kane, N. C., Moore, J.-S., & Vines, T. H. (2012). Recommendations for utilizing and reporting population genetic analyses: The reproducibility of genetic clustering using the program structure. *Molecular Ecology*, *21*(20), 4925–4930. https://doi.org/10.1111/j.1365-294X.2012.05754.x

Gratton, P., Marta, S., Bocksberger, G., Winter, M., Keil, P., Trucchi, E., & Kühl, H. (2017). Which latitudinal gradients for genetic diversity? *Trends in Ecology & Evolution*, *32*(10), 724–726. https://doi.org/10.1016/j.tree.2017.07.007

Hampton, S. E., Anderson, S. S., Bagby, S. C., Gries, C., Han, X., Hart, E. M., Jones, M. B., Lenhardt, W. C., MacDonald, A., Michener, W. K., Mudge, J., Pourmokhtarian, A., Schildhauer, M. P., Woo, K. H., & Zimmerman, N. (2015). The Tao of open science for ecology. *Ecosphere*, *6*(7), art120–13. https://doi.org/10.1890/ES14-00402.1

Kaye, J., Heeney, C., Hawkins, N., de Vries, J., & Boddington, P. (2009). Data sharing in genomics–re-shaping scientific practice. *Nature Reviews Genetics*, *10*(5), 331–335. https://doi.org/10.1038/nrg2573

Lawrence, E. R., Benavente, J. N., Matte, J.-M., Marin, K., Wells, Z. R. R., Bernos, T. A., Krasteva, N., Habrich, A., Nessel, G. A., Koumrouyan, R. A., & Fraser, D. J. (2019). Geo-referenced population-specific microsatellite data across American continents, the MacroPopGen Database. *Scientific Data*, *6*(1), 1–9. https://doi.org/10.1038/s41597-019-0024-7

Leinonen, R., Akhtar, R., Birney, E., Bonfield, J., Bower, L., Corbett, M., Cheng, Y., Demiralp, F., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Hunter, C., Jang, M., Leonard, S., Lin, Q., Lopez, R., Maguire, M., McWilliam, H., … Cochrane, G. (2009). Improvements to services at the European Nucleotide Archive. *Nucleic Acids Research*, *38*(suppl_1), D39–D45. https://doi.org/10.1093/nar/gkp998

Leinonen, R., Sugawara, H., Shumway, M., & INSDC, (2010). The sequence read archive. *Nucleic Acids Research*, *39*(Database), D19–D21. https://doi.org/10.1093/nar/gkq1019

Liggins, L., & Arranz, V. (2018). Genetic diversity and connectivity of Beveridge Reef's marine biodiversity in an Indo-Pacific wide context (37p). *Report for the Department of Agriculture, Forestry and Fisheries of Niue*, 1–37. https://doi.org/10.6084/m9.figshare.12287069.v1

Manel, S., Guerin, P. E., Mouillot, D., Blanchet, S., Velez, L., Albouy, C., & Pellissier, L. (2020). Global determinants of freshwater and marine fish genetic diversity. *Nature Communications*, *11*(1), 692. https://doi.org/10.1038/s41467-020-14409-7

Matias, A. M. A., & Riginos, C. (2018). Revisiting the "Centre Hypotheses" of the Indo-West Pacific: Idiosyncratic genetic diversity of nine reef species offers weak support for the Coral Triangle as a centre of genetic biodiversity. *Journal of Biogeography*, *96*, 707. https://doi.org/10.1111/jbi.13376

Millette, K. L., Fugere, V., Debyser, C., Greiner, A., Chain, F. J. J., & Gonzalez, A. (2020). No consistent effects of humans on animal genetic diversity worldwide. *Ecology Letters*, *23*, 55–67. https://doi.org/10.1111/ele.13394

Miraldo, A., Li, S., Borregaard, M. K., Florez-Rodriguez, A., Gopalakrishnan, S., Rizvanovic, M., Wang, Z., Rahbek, C., Marske, K. A., & Nogues-Bravo, D. (2016). An Anthropocene map of genetic diversity. *Science*, *353*(6307), 1532–1535. https://doi.org/10.1126/science.aaf4381

Parker, M., Stones-Havas, S., Starger, C., & Meyer, C. (2012). laboratory information management systems for DNA barcoding. In W. J. Kress, & D. L. Erickson (Eds.), *DNA barcodes: Methods and Protocols*. (Vol. *858*, pp. 269–310). Springer. https://doi.org/10.1007/978-1-61779-591-6_13

Pope, L. C., Liggins, L., Keyse, J., Carvalho, S. B., & Riginos, C. (2015). Not the time or the place: The missing spatio-temporal link in publicly available genetic data. *Molecular Ecology*, *24*(15), 3802–3809. https://doi.org/10.1111/mec.13254

Powers, S. M., & Hampton, S. E. (2018). Open science, reproducibility, and transparency in ecology. *Ecological Applications*, *29*(1), 282–288. https://doi.org/10.1002/eap.1822

Quattrini, A. M., Faircloth, B. C., Dueñas, L. F., Bridge, T. C. L., Brugler, M. R., Calixto-Botía, I. F., DeLeo, D. M., Forêt, S., Herrera, S., Lee, S. M. Y., Miller, D. J., Prada, C., Rádis-Baptista, G., Ramírez-Portilla, C., Sánchez, J. A., Rodríguez, E., & McFadden, C. S. (2018). Universal target-enrichment baits for anthozoan (Cnidaria) phylogenomics: New approaches to long-standing problems. *Molecular Ecology Resources*, *18*, 281–295. https://doi.org/10.1111/1755-0998.12736

Roche, D. G., Lanfear, R., Binning, S. A., Haff, T. M., Schwanz, L. E., Cain, K. E., & Kruuk, L. E. B. (2014). Troubleshooting public data archiving: Suggestions to increase participation. *PLOS Biology*, *12*(1), e1001779. https://doi.org/10.1371/journal.pbio.1001779.g002

Salces-Castellano, A., Patiño, J., Alvarez, N., Andújar, C., Arribas, P., Braojos-Ruiz, J. J., Arco-Aguilar, M., García-Olivares, V., Karger, D. N., López, H., Manolopoulou, I., Oromí, P., Pérez-Delgado, A. J., Peterman, W. E., Rijsdijk, K. F., & Emerson, B. C. (2020). Climate drives community-wide divergence within species over a limited spatial scale: Evidence from an oceanic island. *Ecology Letters*, *23*(2), 305–315. https://doi.org/10.1111/ele.13433

Smith, B. T., McCormack, J. E., Cuervo, A. M., Hickerson, M. J., Aleixo, A., Cadena, C. D., & Brumfield, R. T. (2014). The drivers of tropical speciation. *Nature Communications*, *515*(7527), 406–409. https://doi.org/10.1038/nature13687

Stewart, L. C., Hill-Moana, T., Brown, J., Bury, S. J., & Crutchley, G. (2020). Controls on methane oxidation rates in sediments from Hikurangi margin methane seeps. *Environmental Microbiology*.

Stewart, L. C., Houghton, K., Carere, C. R., Power, J. F., Chambefort, I., & Stott, M. B. (2018). Interaction between ferruginous clay sediment and an iron-reducing hyperthermophilic *Pyrobaculum sp*. in a terrestrial hot spring. *FEMS Microbiology Ecology*, *94*(11), 33–11. https://doi.org/10.1093/femsec/fiy160

Stewart, L. C., Stucker, V. K., Stott, M. B., & de Ronde, C. E. J. (2018). Marine-influenced microbial communities inhabit terrestrial hot springs on a remote island volcano. *Extremophiles*, *22*(4), 687–698. https://doi.org/10.1007/s00792-018-1029-4

Titus, B. M., Benedict, C., Laroche, R., Gusmão, L. C., Van Deusen, V., Chiodo, T., Meyer, C. P., Berumen, M. L., Bartholomew, A., Yanagi, K.,

Reimer, J. D., Fujii, T., Daly, M., & Rodríguez, E. (2019). Phylogenetic relationships among the clownfish-hosting sea anemones. *Molecular Phylogenetics and Evolution*, *139*, 106526. https://doi.org/10.1016/j.ympev.2019.106526

Vision, T. J. (2010). Open data and the social contract of scientific publishing. *BioScience*, *60*(5), 330–331. https://doi.org/10.1525/bio.2010.60.5.2

Whitlock, M. C. (2011). Data archiving in ecology and evolution: Best practices. *Trends in Ecology & Evolution*, *26*(2), 61–65. https://doi.org/10.1016/j.tree.2010.11.006

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., & Vieglais, D. (2012). Darwin Core: An evolving community-developed biodiversity data standard. *PLoS One*, *7*(1), e29715. https://doi.org/10.1371/journal.pone.0029715

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. 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. (2019). Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Nature Ecology and Evolution*, *3*(1), 1–9. https://doi.org/10.1038/sdata.2016.18

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, 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*(5), 415–420. https://doi.org/10.1038/nbt.1823

**How to cite this article:** Riginos C, Crandall ED, Liggins L, et al. 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. *Mol Ecol Resour.* 2020;20:1458–1469. https://doi.org/10.1111/1755-0998.13269