COMMUNITY PAGE

The Genomic Observatories Metadatabase (GeOMe): A new repository for field and sampling event metadata associated with genetic samples

John Deck[1]*, Michelle R. Gaither[2], Rodney Ewing[3], Christopher E. Bird[2,4], Neil Davies[5,6], Christopher Meyer[7], Cynthia Riginos[8], Robert J. Toonen[2], Eric D. Crandall[9]*

1 Berkeley Natural History Museums, University of California, Berkeley, California, United States of America, **2** Hawaii Institute of Marine Biology, University of Hawaii, Kaneohe, Hawaii, United States of America, **3** Biocode, LLC, Junction City, Oregon, United States of America, **4** Texas A&M University, Corpus Christi, Texas, United States of America, **5** Gump South Pacific Research Station, University of California, Moorea, French Polynesia, **6** Berkeley Institute for Data Science, University of California, Berkeley, California, United States of America, **7** National Museum of Natural History, Smithsonian Institution, Washington, DC, United States of America, **8** University of Queensland, St Lucia, Queensland, Australia, **9** School of Natural Sciences, California State University, Monterey Bay, Marina, California, United States of America

* jdeck88@gmail.com (JD); ecrandall@csumb.edu (EC)

Citation: Deck J, Gaither MR, Ewing R, Bird CE, Davies N, Meyer C, et al. (2017) The Genomic Observatories Metadatabase (GeOMe): A new repository for field and sampling event metadata associated with genetic samples. PLoS Biol 15(8): e2002925. https://doi.org/10.1371/journal.pbio.2002925

Published: August 3, 2017

Funding: National Science Foundation https://www.nsf.gov/ (grant number DEB-1457848; DIPnet RCN). Received by EDC, CR, RT, MRG, CEB, JD, ND. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Gordon and Betty Moore Foundation https://www.moore.org/ (grant number GBMF1619 Moorea Biocode). Received by CM, ND, JD. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. National Oceanic and Atomospheric Administration http://www.noaa.gov/ (grant number NMFS-PIRO-NA15NMF4270330). Received by CEB. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. National Science Foundation https://

Abstract

The Genomic Observatories Metadatabase (GeOMe, http://www.geome-db.org/) is an open access repository for geographic and ecological metadata associated with biosamples and genetic data. Whereas public databases have served as vital repositories for nucleotide sequences, they do not accession all the metadata required for ecological or evolutionary analyses. GeOMe fills this need, providing a user-friendly, web-based interface for both data contributors and data recipients. The interface allows data contributors to create a customized yet standard-compliant spreadsheet that captures the temporal and geospatial context of each biosample. These metadata are then validated and permanently linked to archived genetic data stored in the National Center for Biotechnology Information's (NCBI's) Sequence Read Archive (SRA) via unique persistent identifiers. By linking ecologically and evolutionarily relevant metadata with publicly archived sequence data in a structured manner, GeOMe sets a gold standard for data management in biodiversity science.

The missing metadata

Documenting patterns of global biodiversity and understanding how that diversity is generated and maintained are important steps towards mitigating the effects of anthropogenic stressors [1–3], whether local or global. Genetic data are key to this effort as these data can be used to: (a) identify cryptic diversity, (b) define population structure and associated management units, (c) identify hot spots of genetic diversity for the conservation of adaptive potential, (d) study the mechanisms driving patterns of biodiversity to identify regions of high evolutionary potential [4,5], and (e) monitor the flux of both intra- and interspecific genetic diversity at

www.nsf.gov/ (grant number DBI-0956426; Biological Science Collections Tracker). Received by ND, JD, CM. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. National Science Foundation https://www.nsf.gov/ (grant number EF-0905606; NESCent catalysis and working group funding). Received by EDC, CR, RT. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: BLAST, Basic Local Alignment Search Tool; CC0, Creative Commons Zero license; CSV, comma-separated value; EMBL-EBI, European Bioinformatics Institute; FASTA, Fast Alignment Search Tool-All; FASTQ, Fast Alignment Search Tool-Quality; FIMS, Biocode Field Information Management System; GBIF, Global Biodiversity Information Facility; GeOMe, Genomic Observatories Metadatabase; GO Network, Genomic Observatories Network; GSC, Genomic Standards Consortium; IPT, Integrated Publishing Toolkit; KML, keyhole markup language; LIMS, Laboratory Information Management System; MIxS, GSC's Minimum Information about any (x) marker Sequence; NCBI, US National Center for Biotechnology Information; NSF, US National Science Foundation; RCN, NSF Research Coordination Network; RDF, Resource Description Framework; REST, representational state transfer; SRA, NCBI's Sequence Read Archive; TDWG, Biodiversity Information Standards Organization; URI, Uniform Resource Identifiers.

Provenance: Not commissioned; externally peer reviewed

a particular site or within a particular region [6]. Whereas there have been several coordinated efforts to document patterns of species diversity (e.g., Global Biodiversity Information Facility [GBIF, http://www.gbif.org/; see Table 1 for acronym definitions], Ocean Biogeographic Information System [http://www.iobis.org/]), there have been fewer attempts to document and archive global patterns of genetic diversity. Notable efforts in this direction, however, include the Earth Microbiome Project [7,8] and Ocean Sampling Day [9], focusing on microbes, the Genomic Observatories Network (GO Network) of research sites focusing on entire ecosystems [10,11], and analyses of data archived in public repositories [12].

While granting agencies and publishers enforce data accessibility and open access requirements for genetic data, they do not always require standardized metadata [13–15]. The public genetic repositories, such as NCBI and the European Bioinformatics Institute (EMBL-EBI), were established to store large volumes of sequence data. With vast capacity for storage and curation of genetic data, their role as repositories for the growing volume of genetic data is crucial; however, NCBI, for example, encourages but does not require the standardized metadata needed for ecological- or evolutionary-level analyses. Yet standards do exist for such metadata, notably thanks to the efforts of the Genomic Standards Consortium (GSC) [16] and the Biodiversity Information Standards Organization (known as "TDWG," http://www.tdwg.org/). The GSC's Minimum Information about any (x) marker Sequence (MIxS) standard [17] specifies a set of metadata standards for genetic data. Likewise, TDWG's Darwin Core is a body of standards for describing and sharing biodiversity information [18]. However, neither NCBI nor EMBL-EBI currently enforces these standards or offers a portal for searching MIxS-compliant data. The problem is not only with the genetic repositories. The Dryad Digital Repository is an important resource that links data to their associated scientific publications and makes those data citable, yet Dryad does not enforce set standards or metadata requirements.

New databases and repositories that accommodate specific disciplines and subfields are coming online, e.g., http://reefgenomics.org/ [19], but there remains no central cross-disciplinary repository that enforces MIxS standards for sequence data and requires submission of the associated metadata describing the ecological and geographic context of source tissues. This "metadata gap" means that vital information about sampling events, such as sampling location, date, habitat, and organism life history, are rarely reported. Instead, most of this information is left unpublished, greatly diminishing the potential value (reuse) of the data [13,14,20].

Filling the metadata gap: GeOMe

To fill the metadata gap for genetic sequence data, we have developed a web-based database and infrastructure to aid collaboration and the cross dissemination of published genetic data (http://geome-db.org/). GeOMe can be easily expanded as necessary to accommodate an increasing diversity of data from various research communities. Early development began as part of the Moorea Biocode Project (http://biocode.berkeley.edu/, Moore Foundation) and subsequently the National Science Foundation (NSF) Biological Science Collections Tracker project (http://biscicol.blogspot.com/). Development continued under a NSF Research Coordination Network (RCN) grant [16], which led to the establishment of the GO Network [10, 11] as a joint initiative of GSC and the Group on Earth Observations Biodiversity Observation Network [21]. The resulting informatics stack (Biocode Commons) reached its current level of development under the auspices of another NSF RCN (the Diversity of the Indo-Pacific Network, http://diversityindopacific.net/) and is now being expanded for the broader scientific community as GeOMe.

Table 1. Acronym definitions.

Category	Acronym	Name
Databases	Dryad	Dryad Digital Repository
	GBIF	Global Biodiversity Information Facility
	GeOMe	Genomic Observatories Metadatabase
	SRA	NCBI's Sequence Read Archive
Organizations	EMBL-EBI	European Bioinformatics Institute
	EMP	Earth Microbiome Project
	GO Network	Genomic Observatories Network
	GSC	Genomic Standards Consortium
	NCBI	US National Center for Biotechnology Information
	NSF	US National Science Foundation
	RCN	Research Coordination Network
	TDWG	Biodiversity Information Standards Organization, aka Taxonomic Databases Working Group
Standards	MIxS	GSC's Minimum Information about any (x) marker Sequence
	RDF	Resource Description Framework
	DwC	Darwin Core, TDWG's body of standards for sharing information about biological diversity
File formats	FASTA	Fast Alignment Search Tool-All
	FASTQ	Fast Alignment Search Tool-Quality
Tools	EZID	Tool for creating and managing globally-unique, long-term identifiers for data
	FIMS	Biocode Field Information Management System

https://doi.org/10.1371/journal.pbio.2002925.t001

The suite of tools provided through GeOMe provides a platform for investigators to publish standardized metadata that captures the temporal, environmental, geospatial, and even scholarly context for each sample and its derivative genetic data. GeOMe's user-friendly, web-based interface allows users, from student and single investigator–driven projects to large scientific consortia, to customize metadata templates using the Biocode Field Information Management System (FIMS) [22]. Users select from a set of fields constructed from standard Darwin Core terms (http://rs.tdwg.org/dwc/) to create a metadata template that best reflects their needs and can be reused across multiple projects within or between labs (Fig 1). Data field options include hypotheses about the taxon (if an individual organism) or taxa in the sample (e.g., bacteria) and information on sampling habitat, life history (if an individual organism), details of sampling location and time, and publications deriving from the data. GeOMe provides a set of customizable project-level metadata validation rules, which ensures that metadata are

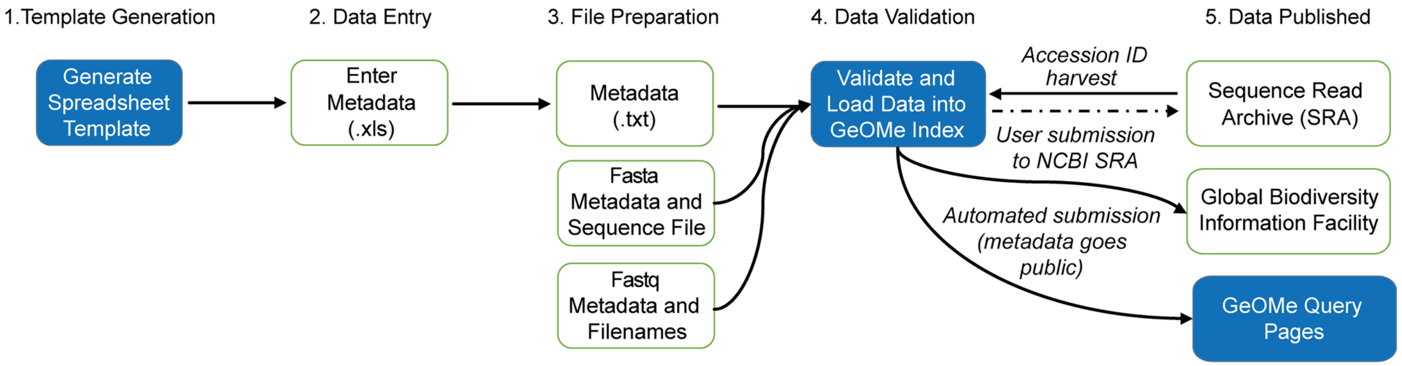

1.Template Generation
2. Data Entry
3. File Preparation
4. Data Validation
5. Data Published
Generate Spreadsheet Template
Enter Metadata (.xls)
Metadata (.txt)
Fasta Metadata and Sequence File
Fastq Metadata and Filenames
Validate and Load Data into GeOMe Index
Accession ID harvest
User submission to NCBI SRA
Sequence Read Archive (SRA)
Automated submission (metadata goes public)
Global Biodiversity Information Facility
GeOMe Query Pages

Fig 1. The Genomic Observatories Metadatabase (GeOMe) workflow. Steps in blue are those conducted within the Field Information Management System (FIMS) of GeOMe while those in white are independent of GeOMe.

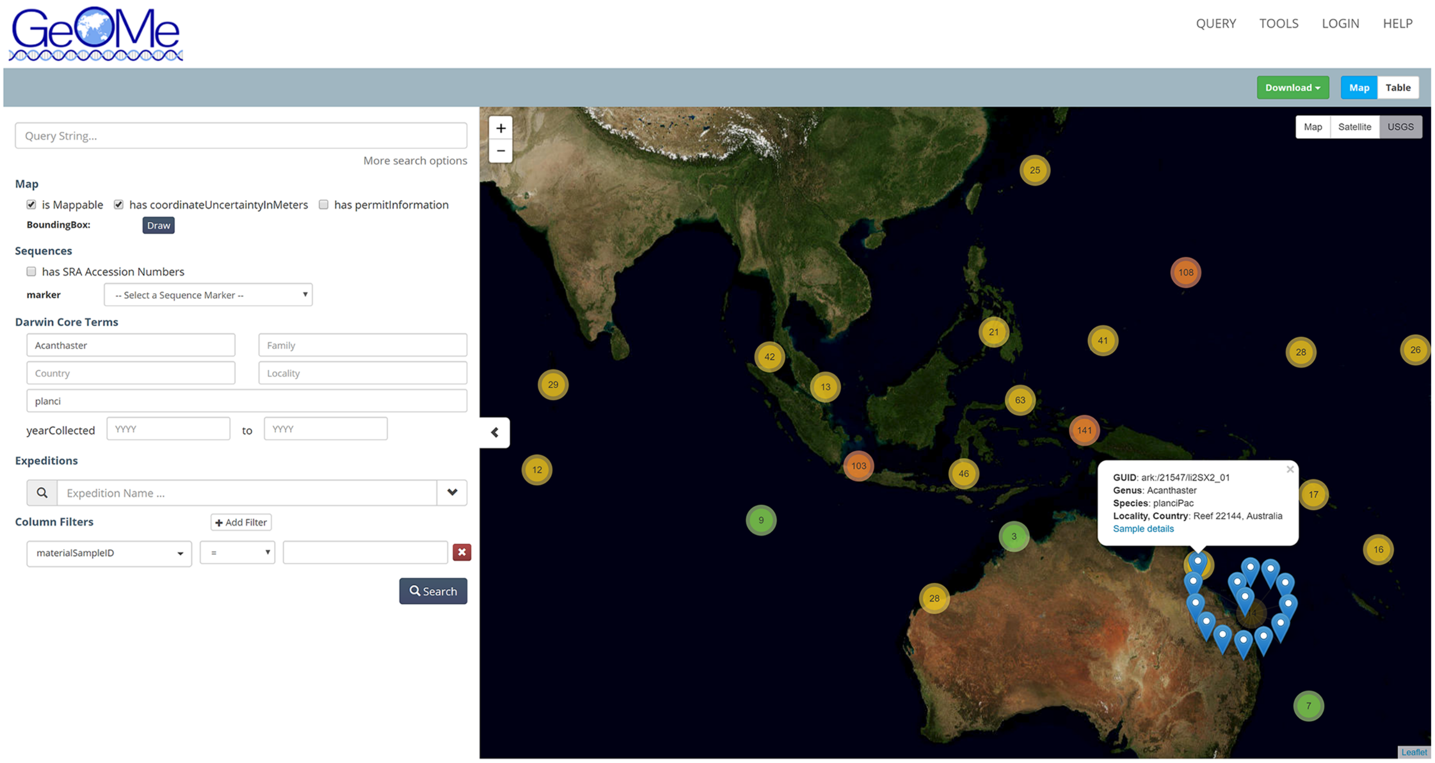
https://doi.org/10.1371/journal.pbio.2002925.g001

compliant with both Darwin Core and MIxS standards (i.e., each sample has a unique identifier and required fields are provided). Thus, research communities can easily design their own templates and validation rules to describe, for example, an environmental sample used in metagenomics, tissues associated with transcriptomics, or an individual organism's genomic sequence. Once the metadata template has been created, no internet connection is required for template editing until the data are uploaded, and therefore the system can be used in remote locations and with any personal computer that employs spreadsheet software (e.g., Microsoft Excel or comma-separated value [CSV] formats are supported).

The FIMS architecture (https://github.com/biocodellc) draws on community vocabularies (Darwin Core and MIxS) with terms stored internally as Uniform Resource Identifiers (URIs) and as specified by the Resource Description Framework (RDF) model. Most user-supplied data are stored as attributes of a core "sample" and are joined to either Sanger-based sequence data (including the marker name and actual sequence) or high throughput sequence data (storing metadata associated with sequence data stored on NCBI's SRA). RDF-based attributes and class names for samples and sequences are then indexed in a document-store database (ElasticSearch, http://www.elastic.co/) for fast retrieval.

To submit data to GeOMe (Fig 1), contributors upload a tab-delimited text file together with a Fast Alignment Search Tool-All (FASTA) file (for Sanger sequence data) or a list of Fast Alignment Search Tool-Quality (FASTQ) file names (for high throughput sequence data, in which FASTQ files contain data from an individual sample). GeOMe then validates the dataset, checking to ensure that a set of minimum required fields are complete (following project-specific rules) and that sequence identifiers match metadata identifiers. When rules are violated, an informative and easy-to-interpret error message appears, prompting the user to fix the issue before proceeding. The contributor is also presented with a map of sampling localities to allow them to verify the geospatial information. Once validated, GeOMe assigns persistent, universally unique identifiers to each sample (EZID: California Digital Library; http://ezid.cdlib.org/), which are used for linking samples between GeOMe, NCBI, and other repositories. Sanger sequence data are stored as a text field within the database. For high throughput sequence data, GeOMe provides the data contributor with a completed batch metadata file for NCBI's SRA and a SRA BioSample file to facilitate submission of the data to the NCBI SRA. Once the data are uploaded to the SRA, GeOMe harvests the NCBI accession numbers, thereby creating a direct link between the genetic data, the sample EZID, and associated metadata. To maximize open access, metadata are available under a Creative Commons Zero license (CC0) and are automatically pushed to GBIF using a dedicated Integrated Publishing Toolkit (IPT, http://www.gbif.org/ipt) installation [23]. Finally, users can choose to embargo their uploaded datasets from public view for a period of up to 2 years from the date of submission. While we encourage all users to make their data immediately public and CC0 on upload, we recognize that GeOMe is useful in preparing and processing research outputs and, consequently, data may not be ready for public release.

GeOMe is designed for flexibility and persistence using representational state transfer (REST) web services for communication between the database and the interface, while enabling potential third party applications to interact with services, as well. GeOMe's web interface enables flexible searches based on any field and/or a geospatial bounding box (Fig 2). The GeOMe database may also be queried with a dedicated R package (geomedb; https://github.com/DIPnet/fimsR-access). GeOMe has also been designed so that it can be used in conjunction with the Biocode Laboratory Information Management System (LIMS; http://software.mooreabiocode.org) for the Geneious software platform (Biomatters, Incorporated). Sanger sequence data are available for download in FASTA format, while high throughput sequence data are provided as a list of SRA accession numbers. Associated metadata can then



Query String...

Map

☑ is Mappable ☑ has coordinateUncertaintyInMeters □ has permitInformation

Sequences

□ has SRA Accession Numbers

marker: -- Select a Sequence Marker --

Darwin Core Terms

Acanthaster | Family

Country | Locality

planci

yearCollected YYYY to YYYY

Expeditions

Expedition Name ...

Column Filters + Add Filter

materialSampleID =

Search

GUID: ark:/21547/li2SX2_01
Genus: Acanthaster
Species: planciPac
Locality, Country: Reef 22144, Australia
Sample details

Fig 2. Screen shot of the Genomic Observatories Metadatabase (GeOMe) query system for *Acanthaster planci*, the crown of thorns sea star. Each number indicates the number of specimens in the database from that location. When a group of specimens is selected, distinct samples are visible as a spiral radiating from the chosen location, and individual records report summary information about each sample.

https://doi.org/10.1371/journal.pbio.2002925.g002

be downloaded in CSV and keyhole markup language (KML) formats. Already, the database contains metadata for >35,000 Sanger sequences across 233 species supplied from >50 participating laboratories. It has recently begun accepting metadata for high throughput FASTQ datasets. By using the FIMS architecture for metadata but continuing to store genetic sequence data at NCBI, we are helping to ensure long-term persistence of links between sequence data and its associated metadata while keeping the data searchable with NCBI's Basic Local Alignment Search Tool (BLAST). We believe that this flexibility enables maximum integration with similar regional or discipline-specific data archival initiatives.

Conclusion

A major challenge for biodiversity genomics research is the need to carry out physical sampling in the field (nucleotide sequences cannot be obtained remotely) and then to link biologically and ecologically important metadata with downstream data products, notably, published genetic sequences. No existing federated database provides this functionality. Yet, maintaining linkages among these data types is vital for data integration and analysis. Publicly archiving these metadata is essential to ensure scientific reproducibility and synthesis as well as to maximize potential reuse of sequence data as new techniques develop. Here, we provide a solution to the metadata gap: GeOMe. A bottom-up effort with buy-in from over 50 laboratories, our database is growing and adding new capacity while also setting the industry standard for metadata publication.

Acknowledgments

This paper is dedicated to Dr. Steve Karl, an early supporter and steering committee member for the Diversity in the Indo-Pacific Network, who helped to shape an early version of GeOMe and who passed away while this paper was in revision. The authors also thank the other enthusiastic participants in a catalysis meeting and working group on the Molecular Ecology and Evolution in the Indo-Pacific, hosted at the National Evolutionary Synthesis Center (http://www.nescent.org/) from 2012 to 2015.

References

1. Brooks TM, Mittermeier RA, da Fonseca GA, Gerlach J, Hoffmann M, Lamoreux JF, et al. Global biodiversity conservation priorities. Science. 2006; 313(5783): 58–61. https://doi.org/10.1126/science.1127609 PMID: 16825561
2. Halpern BS, Walbridge S, Selkoe KA, Kappel CV, Micheli F, D'Agrosa C, et al. A global map of human impact on marine ecosystems. Science. 2008; 319(5865): 948–52. https://doi.org/10.1126/science.1149345 PMID: 18276889
3. Barnosky AD, Hadly EA, Bascompte J, Berlow EL, Brown JH, Fortelius M, et al. Approaching a state shift in Earth's biosphere. Nature. 2012; 486(7401): 52–8. https://doi.org/10.1038/nature11018 PMID: 22678279
4. Harrisson KA, Pavlova A, Telonis-Scott M, Sunnucks P. Using genomics to characterize evolutionary potential for conservation of wild populations. Evolutionary Applications. 2014; 7(9): 1008–25. https://doi.org/10.1111/eva.12149 PMID: 25553064
5. Moritz C. Strategies to protect biological diversity and the evolutionary processes that sustain it. Syst Biol. 2002; 51(2): 238–54. https://doi.org/10.1080/10635150252899752 PMID: 12028731
6. Beger M, Selkoe KA, Treml E, Barber PH, von der Heyden S, Crandall ED, et al. Evolving coral reef conservation with genetic information. Bull Mar Sci. 2014; 90(1): 159–85.
7. Gilbert JA, Jansson JK, Knight R. The Earth Microbiome Project: successes and aspirations. BMC Biol. 2014; 12(1): 69.
8. Gilbert JA, Meyer F, Antonopoulos D, Balaji P, Brown CT, Brown CT, et al. Meeting report: the terabase metagenomics workshop and the vision of an earth microbiome project. Standards in Genomic Sciences. 2010; 3(3): 243. https://doi.org/10.4056/sigs.1433550 PMID: 21304727
9. Kopf A, Bicak M, Kottmann R, Schnetzer J, Kostadinov I, Lehmann K, et al. The Ocean Sampling Day Consortium. Gigascience. 2015; 4(1): 27.
10. Davies N, Field D, Amaral-Zettler L, Clark MS, Deck J, Drummond A, et al. The founding charter of the Genomic Observatories Network. GigaScience. 2014; 3(1): 2. https://doi.org/10.1186/2047-217X-3-2 PMID: 24606731
11. Davies N, Meyer C, Gilbert JA, Amaral-Zettler L, Deck J, Bicak M, et al. A call for an international network of genomic observatories (GOs). GigaScience. 2012; 1(1): 5. https://doi.org/10.1186/2047-217X-1-5 PMID: 23587188
12. Miraldo A, Li S, Borregaard MK, Flórez-Rodríguez A, Gopalakrishnan S, Rizvanovic M, et al. An Anthropocene map of genetic diversity. Science. 2016; 353(6307): 1532–5. https://doi.org/10.1126/science.aaf4381 PMID: 27708102
13. Pope LC, Liggins L, Keyse J, Carvalho SB, Riginos C. Not the time or the place: the missing spatio-temporal link in publicly available genetic data. Mol Ecol. 2015; 24(15): 3802–9. https://doi.org/10.1111/mec.13254 PMID: 26033415
14. Gratton P, Marta S, Bocksberger G, Winter M, Trucchi E, Kühl H. A world of sequences: can we use georeferenced nucleotide databases for a robust automated phylogeography? J Biogeogr. 2016.
15. Keyse J, Crandall ED, Toonen RJ, Meyer CP, Treml EA, Riginos C. The scope of published population genetic data for Indo-Pacific marine fauna and future research opportunities in the region. Bull Mar Sci. 2014; 90(1): 47–78.
16. Wooley JC, Field D, Glöckner F-O. Extending standards for genomics and metagenomics data: a Research Coordination Network for the Genomic Standards Consortium (RCN4GSC). Standards in genomic sciences. 2009; 1(1): 87. https://doi.org/10.4056/sigs.26218 PMID: 21304642
17. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol. 2011; 29(5): 415–20. https://doi.org/10.1038/nbt.1823 PMID: 21552244

18. Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, et al. Darwin Core: an evolving community-developed biodiversity data standard. PLoS ONE. 2012; 7(1): e29715. https://doi.org/10.1371/journal.pone.0029715 PMID: 22238640

19. Liew YJ, Aranda M, Voolstra CR. Reefgenomics.org-a repository for marine genomics data. Database. 2016; 2016: baw152.

20. Sidlauskas B, Ganapathy G, Hazkani-Covo E, Jenkins KP, Lapp H, McCall LW, et al. Linking big: the continuing promise of evolutionary synthesis. Evolution. 2010; 64(4): 871–80. https://doi.org/10.1111/j.1558-5646.2009.00892.x PMID: 19895550

21. Scholes R, Mace G, Turner W, Geller G, Jürgens N, Larigauderie A, et al. Toward a global biodiversity observing system. Science. 2008; 321(5892): 1044–5. https://doi.org/10.1126/science.1162055 PMID: 18719268

22. Deck J, Gross J, Stones-Havas S, Davies N, Shapley R, Meyer C. Field information management systems for DNA barcoding. DNA Barcodes: Methods and Protocols. 2012: 255–67.

23. Robertson T, Döring M, Guralnick R, Bloom D, Wieczorek J, Braak K, et al. The GBIF integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the internet. PLoS ONE. 2014; 9 (8): e102623. https://doi.org/10.1371/journal.pone.0102623 PMID: 25099149