BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 4: e59065
doi: 10.3897/biss.4.59065

OPEN ACCESS

Conference Abstract

Survey of Species Covered by DNA Barcoding Data in BOLD and GenBank for Integration of Data for Museomics

Takeru Nakazato ‡

‡ Database Center for Life Science, Mishima, Japan

Corresponding author: Takeru Nakazato (nakazato@dbcls.rois.ac.jp)

Received: 28 Sep 2020 | Published: 29 Sep 2020

Citation: Nakazato T (2020) Survey of Species Covered by DNA Barcoding Data in BOLD and GenBank for Integration of Data for Museomics. Biodiversity Information Science and Standards 4: e59065. https://doi.org/10.3897/biss.4.59065

Abstract

DNA barcoding technology has become employed widely for biodiversity and molecular biology researchers to identify species and analyze their phylogeny. Recently, DNA metabarcoding and environmental DNA (eDNA) technology have developed by expanding the concept of DNA barcoding. These techniques analyze the diversity and quantity of organisms within an environment by detecting biogenic DNA in water and soil. It is particularly popular for monitoring fish species living in rivers and lakes (Takahara et al. 2012). BOLD Systems (Barcode of Life Database systems, Ratnasingham and Hebert 2007) is a database for DNA barcoding, archiving 8.5 million of barcodes (as of August 2020) along with the voucher specimen, from which the DNA barcode sequence is derived, including taxonomy, collected country, and museum vouchered as metadata (e.g. https://www.boldsystems.org/index.php/Public_RecordView?processid=TRIBS054-16). Also, many barcoding data are submitted to GenBank (Sayers et al. 2020), which is a database for DNA sequences managed by NCBI (National Center for Biotechnology Information, US). The number of the records of DNA barcodes, i.e. COI (cytochrome c oxidase I) gene for animal, has grown significantly (Porter and Hajibabaei 2018). BOLD imports DNA barcoding data from GenBank, and lots of DNA barcoding data in GenBank are also assigned BOLD IDs. However, we have to refer to both BOLD and GenBank data when performing DNA barcoding. I have previously investigated the registration of DNA

barcoding data in GenBank, especially the association with BOLD, using insects and flowering plants as examples (Nakazato 2019). Here, I surveyed the number of species covered by BOLD and GenBank. I used fish data as an example because eDNA research is particularly focused on fish.

I downloaded all GenBank files for vertebrates from NCBI FTP (File Transfer Protocol) sites (as of November 2019). Of the GenBank fish entries, 86,958 (7.3%) were assigned BOLD identifiers (IDs). The NCBI taxonomy database has registrations for 39,127 species of fish, and 20,987 scientific names at the species level (i.e., excluding names that included sp., cf. or aff.). GenBank entries with BOLD IDs covered 11,784 species (30.1%) and 8,665 species-level names (41.3%).

I also obtained whole "specimens and sequences combined data" for fish from BOLD systems (as of November 2019). In the BOLD, there are 273,426 entries that are registered as fish. Of these entries, 211,589 BOLD entries were assigned GenBank IDs, i.e. with values in “genbank_accession” column, and 121,748 entries were imported from GenBank, i.e. with "Mined from GenBank, NCBI" description in "institution_storing" column. The BOLD data covered 18,952 fish species and 15,063 species-level names, but 35,500 entries were assigned no species-level names and 22,123 entries were not even filled with family-level names. At the species level, 8,067 names co-occurred in GenBank and BOLD, with 6,997 BOLD-specific names and 599 GenBank-specific names.

GenBank has 425,732 fish entries with voucher IDs, of which 340,386 were not assigned a BOLD ID. Of these 340,386 entries, 43,872 entries are registrations for COI genes, which could be candidates for DNA barcodes. These candidates include 4,201 species that are not included in BOLD, thus adding these data will enable us to identify 19,863 fish to the species level.

For researchers, it would be very useful if both BOLD and GenBank DNA barcoding data could be searched in one place. For this purpose, it is necessary to integrate data from the two databases. A lot of biodiversity data are recorded based on the Darwin Core standard while DNA sequencing data are sometimes integrated or cross-linked by RDF (Resource Description Framework). It may not be technically difficult to integrate these data, but the species data referenced differ from the EoL (The Encyclopedia of Life) for BOLD and the NCBI taxonomy for GenBank, and the differences in taxonomic systems make it difficult to match by scientific name description. GenBank has fields for the latitude and longitude of the specimens sampled, and Porter and Hajibabaei 2018 argue that this information should be enhanced. However, this information may be better described in the specimen and occurrence databases. The integration of barcoding data with the specimen and occurrence data will solve these problems. Most importantly, it will save the researcher from having to register the same information in multiple databases. In the field of biodiversity, only DNA barcode sequences may have been focused on and used as gene sequences. The museomics community regards museum-preserved specimens as rich resources for DNA studies because their biodiversity information can accompany the extraction and analysis of their DNA (Nakazato 2018). GenBank is useful for biodiversity studies due to its low rate of mislabelling (Leray et al. 2019). In the future, we will be

working with a variety of DNA, including genomes from museum specimens as well as DNA barcoding. This will require more integrated use of biodiversity information and DNA sequence data. This integration is also of interest to molecular biologists and bioinformaticians.

Keywords

metabarcoding, environmental DNA, data integration, semantic web

Presenting author

Takeru Nakazato

Presented at

TDWG 2020

Funding program

ROIS-DS-JOINT (004RM2017 and009RM2018), and The Life Science Database Integration Project

Conflicts of interest

The author has declared that no competing interest exists.

References

- Leray M, Knowlton N, Ho S, Nguyen B, Machida R (2019) GenBank is a reliable resource for 21st century biodiversity research. Proceedings of the National Academy of Sciences 116 (45): 22651-22656. https://doi.org/10.1073/pnas.1911714116
- Nakazato T (2018) A Challenge to Integrate Bioinformatics and Biodiversity Informatics Data as Museomics. Biodiversity Information Science and Standards 2 https://doi.org/10.3897/biss.2.26102
- Nakazato T (2019) Current situation of DNA Barcoding data in biodiversity and genomics databases and data integration for museomics. Biodiversity Information Science and Standards 3 https://doi.org/10.3897/biss.3.35165
- Porter TM, Hajibabaei M (2018) Over 2.5 million COI sequences in GenBank and growing. PLOS ONE 13 (9): e0200177. https://doi.org/10.1371/journal.pone.0200177
- Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). Molecular ecology notes 7 (3): 355-364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

- Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I (2020) GenBank. Nucleic acids research 48 (D1): D84-D86. https://doi.org/10.1093/nar/gkz956
- Takahara T, Minamoto T, Yamanaka H, Doi H, Kawabata Z (2012) Estimation of fish biomass using environmental DNA. PloS one 7 (4): e35868. https://doi.org/10.1371/journal.pone.0035868