BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 8: e137771
doi: 10.3897/biss.8.137771

OPEN ACCESS

Conference Abstract

Aspects of NCBI GenBank as a Biodiversity Information Resource

Takeru Nakazato ‡

‡ National Institute of Technology and Evaluation, Tokyo, Japan

Corresponding author: Takeru Nakazato (nakazato.tkr@gmail.com)

Received: 24 Sep 2024 | Published: 24 Sep 2024

Citation: Nakazato T (2024) Aspects of NCBI GenBank as a Biodiversity Information Resource. Biodiversity Information Science and Standards 8: e137771. https://doi.org/10.3897/biss.8.137771

Abstract

DNA sequencing of museum specimens, also known as museomics, provides new insights into the study of biodiversity, including taxonomy, phylogeny, and environmental studies. Also, sequencing specimens have led to the rediscovery of extinct species (Suzuki et al. 2016), identification of related species (Waku et al. 2016), and analysis of ancient DNA (Kanzawa-Kiriyama et al. 2016). Nucleotide sequence data have been collected for more than 30 years under the framework of the International Nucleotide Sequence Database Collaboration (INSDC) by three institutes, namely, National Center for Biotechnology Information, US (NCBI), European Bioinformatics Institute (EBI), and DNA Data Bank of Japan (DDBJ) (Arita et al. 2020). NCBI has collated a database of sequence data, GenBank, which contains approximately 494 million sequences as of April 2022 (Sayers et al. 2021). In fact, GenBank is designed with qualifiers to describe various types of biodiversity information such as "/specimen_voucher", "/lat_lon" (latitude and longitude) and "/collection_date". Also, INSDC now requires that all submissions include the sampling location and date (INSDC 2023). I surveyed the biodiversity information assigned to GenBank records to determine the potential of GenBank as a biodiversity resource.

I downloaded all GenBank data as of August 2023 from the FTP site. The “/specimen_voucher” qualifier was introduced to describe specimen ID in Release 104 in December 1997. This qualifier was designed to fill the value in free text: for example, /specimen_voucher="Smith s. n. 4-IV-1995 (U. S. Natl. Herbarium)". After Release 162 in

October 2007, a method of writing with a structured value of "[<institution-code>:[<collection-code>:]] <specimen_id>" was added (institution-code and collection-code are optional). There are 527,215 records (37.8%) with "/specimen_voucher" qualifier for fish, 3,096,112 records (40.3%) for insects, 1,505,556 records (39.0%) for flowering plants. But fewer than 10% of records have specimen IDs listed using this structured description. To utilize these ambiguous specimen IDs in GenBank, these IDs may need to be cleansed using databases such as NCBI BioCollections, GRSciColl (Global Registry of Scientific Collections) or AI to map them to IDs in databases rich in specimen information such as those of the Global Biodiveristy Information Facility (GBIF) and Barcode of Life System (BOLD). In GenBank, the BOLD ID is listed in the /db_xref qualifier in the “Features” field as the ID of the external database. The 70% of insect sequence data with a specimen ID in the /specimen_voucher qualifier are also assigned a BOLD ID (Nakazato and Jinbo 2022). The correspondence between specimen IDs in biodiversity information databases such as GBIF and specimen IDs in GenBank is expected to further enhance the value of museum specimens.

In addition, GenBank provides the /type_material qualifier for describing the type of voucher (e.g., holotype of *Asphondylia bursicola*). In GenBank insect data, there were over 2,000 records for type material, and approximately 450 species were mentioned, including 269 for holotypes. We found approximately 3,000 records with type information by including “/notes” and “/specimens_voucher” qualifiers in addition to “/type_material”.

Thus, GenBank has potential as a biodiversity information resource, but for more effective use, data mining and linkage with other specimen-based biodiversity databases are essential.

Keywords

sequencing data, DNA barcoding, data integration, BOLD, GBIF, occurrence data, museum specimen, museomics

Presenting author

Takeru Nakazato

Presented at

SPNHC-TDWG 2024

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Arita M, Karsch-Mizrachi I, Cochrane G (2020) The international nucleotide sequence database collaboration. Nucleic Acids Research 49 https://doi.org/10.1093/nar/gkaa967
- INSDC (2023) INSDC spatiotemporal metadata – missing values update (03-04-2023). https://www.insdc.org/news/insdc-spatiotemporal-metadata-missing-values-update-03-04-2023/. Accessed on: 2024-9-22.
- Kanzawa-Kiriyama H, Kryukov K, Jinam TA, Hosomichi K, Saso A, Suwa G, Ueda S, Yoneda M, Tajima A, Shinoda K, Inoue I, Saitou N (2016) A partial nuclear genome of the Jomons who lived 3000 years ago in Fukushima, Japan. Journal of Human Genetics 62 (2): 213-221. https://doi.org/10.1038/jhg.2016.110
- Nakazato T, Jinbo U (2022) Cross-sectional use of barcode of life data system and GenBank as DNA barcoding databases for the advancement of museomics. Frontiers in Ecology and Evolution 10 https://doi.org/10.3389/fevo.2022.966605
- Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry S, Karsch-Mizrachi I (2021) GenBank. Nucleic Acids Research 50 https://doi.org/10.1093/nar/gkab1135
- Suzuki M, Segawa T, Mori H, Akiyoshi A, Ootsuki R, Kurihara A, Sakayama H, Kitayama T, Abe T, Kogame K, Kawai H, Nozaki H (2016) Next-Generation Sequencing of an 88-Year-Old Specimen of the Poorly Known Species *Liagora japonica* (Nemaliales, Rhodophyta) Supports the Recognition of *Otohimella* gen. nov. PLOS ONE 11 (7). https://doi.org/10.1371/journal.pone.0158944
- Waku D, Segawa T, Yonezawa T, Akiyoshi A, Ishige T, Ueda M, Ogawa H, Sasaki H, Ando M, Kohno N, Sasaki T (2016) Evaluating the Phylogenetic Status of the Extinct Japanese Otter on the Basis of Mitochondrial Genome Analysis. PLOS ONE 11 (3). https://doi.org/10.1371/journal.pone.0149341