Data Release Policy

POLICY TITLE: DATA RELEASE
FIRST APPROVAL: MARCH 14, 2023
LAST REVISION APPROVAL: MAY 2026
NEXT SCHEDULED REVIEW: MAY 2027

Purpose

The Centre for Biodiversity Genomics (CBG) serves members of the international research community who are employing DNA-based identification systems. It is resolutely committed to the rapid release and sharing of sequence data and collateral specimen information. This policy document outlines the CBG's approach to ensuring data generated at the CBG are available to the research community in a timely way.

Guidelines

As the CBG is a recipient of funds from Genome Canada, this policy aligns with Genome Canada Data Sharing Policies. Public data release means that data generated at the CBG is made accessible in an appropriate online platform or data repository:

a) DNA barcode data is deposited in the Barcode of Life Data Systems (BOLD) where it is made publicly available after an embargo period (see below). Any individual can create a BOLD user account which provides access to both the BOLD Public Data Portal and its APIs. Public releases include sequence data, specimen provenance, and taxonomic information and may be used by BOLD's data aggregation, analytical, and reporting tools.

b) Metabarcoding data is deposited in the Metabarcoding Research and Visualization Environment (mBRAVE) platform and is made publicly accessible by uploading to the NCBI Sequence Read Archive (SRA) after the embargo period.

Data definitions

DNA barcode data is stored in BOLD as specimen records, each consisting of two components: **(i) specimen data** – information and images for the physical specimen, each designated by a unique Sample ID, and **(ii) sequence data** – information resulting from and related to the molecular analysis of a specimen, each designated by a unique Process ID. Specimen records for animals usually have a third component: **(iii) Barcode Index Number (BIN)** – BOLD's assignment of each qualifying record to a persistent operational taxonomic unit known as a BIN, each designated by a unique identifier. More information on these data elements and definitions is provided in the BOLD Handbook.

Non-target sequence (NTS) records are DNA barcode sequences that are not the intended subject of the sequencing process. NTS records reflect the complexity of biological samples and the sensitivity of modern sequencing technologies and are automatically generated as part of the DNA barcoding workflow. NTS records might reflect genuine biological associations (e.g., parasitic, mutualistic, commensal, predator/prey), or unrelated co-occurrences (e.g., random contamination, nuclear mitochondrial pseudogenes, sequencing artifacts). While our bioinformatic algorithms attempt to identify and remove artifactual NTS prior to upload to BOLD, it is the responsibility of the user of BOLD NTS records to determine their validity. However, these records are only retained on BOLD with explicit permission (through Biological Material Transfer and Data Policy Agreement).

Method of data release

Some components of a nascent DNA barcode record are **immediately available** on BOLD at the time of submission through various tools and pages. For example, upon submission of a specimen to BOLD, its provenance information and image(s) become available to the public through the BOLD Taxonomy Browser. This information is used to generate summary statistics and illustrative distribution maps but does not disclose the content of individual private research projects or specimen data until expiry of the embargo period.

Upon successful sequencing, a specimen and its sequence data enter the BOLD Identification Engine on a **weekly** basis. Reports generated by this tool include similarity scores and tree-based identification with branch labels containing taxonomy and basic geographic information (country, province. Information on individual specimens (e.g., museum catalogue number and voucher repository) and their detailed geographic origin is not disclosed through the BOLD Identification Engine.

Finally, new specimen records are assigned to a BIN **monthly**, and this data is added to the associated BIN page that aggregates information on its members (see example at https://portal.boldsystems.org/bin/BOLD:AAA9566).

BIN Data is batch released to the public on BOLD on a **twice-per-year schedule**, December 15th and July 15th. These data releases are accessible in various formats on BOLD's Data Packages page. Data may then also be uploaded to GenBank and/or GBIF if initiated by the project manager.

Metabarcoding sequencing data are released as BioProject on the NCBI SRA and occurrence records as extended Darwin Core Archive on the Global Biodiversity Information Facility (GBIF).

Embargo period

The embargo period is the amount of time following data upload to BOLD that the data remains private to the primary user. This period enables the primary user to interpret and publish their results before they become available to the entire user community.

Data generated by the CBG is released to the public through the BOLD Data Portal, BOLD Data Packages, and their associated tools twice per year. The official embargo period is 12 months from the date of sequence upload. Data is not released publicly until the embargo period has expired.

Option for data to stay private

The CBG enters into some fee-for-service contractual agreements where it commits to hold client data private. The CBG reviews these arrangements on a case-by-case basis for justification. Such services are only provided by the CBG on a full-cost recovery basis.

Option for immediate data release

Some funders (e.g. Wyss Foundation) might require immediate public release. In such cases CBG will release all data immediately after data validation.

Data release preparation period

The data release preparation period is the interval required for the preparation and validation of data prior to public release. Barcode and NTS data require active curation and thorough validation prior to release to ensure accuracy so it can be used with confidence by the research community. In some instances, data preparation may take longer than the official embargo period.

Data which do not pass quality standards are not released publicly (i.e., failure to recover a sequence; sequencing errors).

Data revision

BOLD is a live database that is actively used and updated by registered users. This implies that records on BOLD are subject to iterative validation processes and other curatorial activities. Any field including the sequence may be updated by CBG and external curators. Both public and private records may be updated through this work. In cases where such revisions affect records from an external project, BOLD notifies the project manager via email.

Acknowledgement

The CBG expects users of its data and resources to acknowledge this usage in publications and to abide by any terms and conditions of use.

These terms and conditions are made explicit on BOLD's Data Packages page.