



BOLD v4: A Centralized Bioinformatics Platform for DNA-Based Biodiversity Data

Sujeevan Ratnasingham, Catherine Wei, Dean Chan, Jireh Agda, Josh Agda, Liliana Ballesteros-Mejia, Hamza Ait Boutou, Zak Mohammad El Bastami, Eddie Ma, Ramya Manjunath, Dana Rea, Chris Ho, Angela Telfer, Jaclyn McKeowan, Miduna Rahulan, Claudia Steinke, Justin Dorsheimer, Megan Milton, and Paul D. N. Hebert

Abstract

BOLD, the Barcode of Life Data System, supports the acquisition, storage, validation, analysis, and publication of DNA barcodes, activities requiring the integration of molecular, morphological, and distributional data. Its pivotal role in curating the reference library of DNA barcodes, coupled with its data management and analysis capabilities, makes it a central resource for biodiversity science. It enables rapid, accurate identification of specimens and also reveals patterns of genetic diversity and evolutionary relationships among taxa.

Launched in 2005, BOLD has become an increasingly powerful tool for advancing the understanding of planetary biodiversity. It currently hosts 17 million specimen records and 14 million barcodes that provide coverage for more than a million species from every continent and ocean. The platform has the long-term goal of providing a consistent, accurate system for identifying all species of eukaryotes.

BOLD's integrated analytical tools, full data lifecycle support, and secure collaboration framework distinguish it from other biodiversity platforms. BOLD v4 brought enhanced data management and analysis capabilities as well as novel functionality for data dissemination and publication. Its next version will include features to strengthen its utility to the research community, governments, industry, and society-at-large.

Key words Biodiversity, DNA barcoding, Bioinformatics, Research collaboration

1 Introduction

Efforts to document biodiversity changes rely heavily on access to robust data sources and tools. Essential reference data is increasingly sourced from the GenBank, GBIF, Catalogue of Life, and BOLD databases, each with a distinct mission and operational mode. GenBank is a repository of nucleotide sequences spanning

all organisms and genes [1] while the Catalogue of Life provides an authoritative taxonomic hierarchy [2]. By contrast, GBIF facilitates access to species occurrence data, predominantly through the aggregation and mediation of specimen records from natural history collections and observational data [2, 3]. BOLD interfaces with all three types of data and functions both as an aggregator and a repository. Like GenBank, it assembles sequence information from diverse taxa but focuses on standardized gene regions [4]. The significance of BOLD lies not only in its large DNA barcode reference library but also in its potential to transform the manner in which species are defined—through genic analysis [5]. BOLD's objective is to simplify the DNA barcoding process by making data more accessible and easily interpretable for a wider audience [4].

The BOLD identification engine further enhances the platform's utility [6]. It allows users to upload unknown DNA barcodes, which are then compared against BOLD's reference sequences to generate an identification. Users receive the closest species match, effectively automating the identification process.

1.1 A Central Database

BOLD is a centralized database for DNA barcode records, which allows researchers/stakeholders to access, store, and analyze DNA barcode data. This centralization brings several important benefits. First, it enables easy access and comparison of DNA barcode data across geographic regions, ecosystems, and taxonomic groups. Second, it facilitates collaboration among researchers and institutions by promoting data sharing.

BOLD further aids the collection of data by employing standardized formats, enhancing the accuracy and reliability of the DNA barcode records employed for species identification and classification. The centralization of DNA barcode data makes it possible to track changes in biodiversity over time, detect and monitor invasive species, and identify key species and habitats that require protection [7]. The availability of comprehensive, reliable data further supports practical conservation efforts and informed decision-making.

1.2 Data Standards

The application of DNA barcodes for specimen identification requires the construction of reference libraries that connect sequences from designated barcode regions (e.g., COI for the animal kingdom) to carefully identified voucher specimens. Reference libraries integrate molecular data with taxonomic information and geo-temporal data. This combination enables algorithmic species identification [8] while also providing high-quality occurrence data.

To qualify for recognition as a DNA barcode, the gene sequence must derive from one of the regions designated by the research community as the barcode marker for a particular domain of life: COI for animals; matK, rbcL, and trnH-psbA for plants [9];

and COI and ITS2 for fungi [10, 11] In addition, records must satisfy three criteria: (i) $< 1\%$ ambiguous bases (= Phred score < 20) [12]; (ii) primer sequences; and (iii) the country of origin. As well, the sequences must be $>75\%$ of the length of the barcode region (i.e., 487 bp for COI) [9–11]. While these criteria must be met to gain formal barcode status, records failing to meet any of these criteria can still be stored in BOLD.

Data submitters mainly drive the status of records within BOLD, but it supports them in maximizing data quality. Various tools help to identify anomalies or low-quality records. For example, submitted sequences of COI are translated automatically into amino acids and compared with a Hidden Markov Model (HMM) [13] of the COI protein to verify that they derive from this gene. Sequences are also examined for stop codons to aid the detection of sequencing errors, pseudogenes, or chimeras. Contamination checks are performed by comparing barcode genes (COI, matK, rbcL, and ITS2) against known laboratory contaminants. For animals, this includes several vertebrates (human, mouse, cow, and pig) and known endosymbionts, among others. The record is flagged if a close match to a contaminant is found. BOLD also captures the event history of records to ensure that erroneous modifications can be identified and reversed.

1.3 DNA-Based Species Identification

BOLD's identification engine is a key feature as it allows users to identify unknown specimens based on their barcode sequence(s). This functionality employs sequence similarity thresholds to generate identifications. Although more rigorous methods are available [14, 15], BOLD employs similarity methods because they are more computationally efficient, and their accuracy is adequate for most use cases [16].

BOLD employs two methods to deliver identifications. For animals, a two-stage process is used to assign an identification. Query sequences are first translated and aligned using a profile Hidden Markov Model [17] of the COI protein. Identifications are then generated by calculating distances between each query sequence and all those in the reference library. The 100 records in the reference library with the closest match are returned along with an assignment to a taxon at a given rank based on similarity thresholds. A single-stage process, based on the BLAST+ [18] algorithm, is used for all other barcode markers.

The reference sequence library for each gene region is updated weekly, incorporating new submissions to BOLD and NCBI. Most records on BOLD are public but some are private. Specific data elements are redacted from private records to safeguard the rights of the data owners, ensuring that they can publish at a reasonable pace. However, these private records are immediately included within the identification engine so all users benefit from the most recent data and aid the detection of erroneous records.

Table 1
Reference libraries on BOLD and their target use cases

Barcode library	Conditions for inclusion	Target use cases
Species level	Records meet the barcode standard. Source specimens have a species designation.	This library delivers species-level matches. Some matches are to private records.
Public	Records meeting the barcode standard which are public.	This library should be used when access to the records underlying an identification is required (e.g., regulatory or forensic uses).
Species level full length	Records with a species-level taxonomic assignment and a full-length barcode (> 640 bp for COI).	This is the highest quality library on BOLD as it only includes records with species designations and full-length barcode sequences. This library is very effective for the identification of short barcode sequences.
All	All records meet the barcode standard.	This library is the most comprehensive, but many records possess limited taxonomic resolution (e.g., identified to a family).
Historical	A snapshot of the records in the library at the end of each year.	These libraries focus on reproducibility and support the re-evaluation of prior studies.

The BOLD Identification Engine is accessible from three points on the platform. The commonest access point is from the Identification Page (http://www.boldsystems.org/index.php/IDS_OpenIdEngine), which is open to any user. A more scalable option is available through the Workbench. Its use requires registration, but it supports the batch identification of hundreds of sequences. The final option is via the Application Programming Interface (API), which allows users to programmatically engage a service endpoint to request identifications, enabling them to integrate identification services into a pipeline or custom application.

Five reference sequence libraries are available (*see* Table 1) to support different needs. Four are updated weekly while the historical libraries represent an annual snapshot.

1.4 Workbench Supporting Data Collection and Publication

BOLD manages DNA barcode data from the earliest point in its lifecycle. Its primary function is to provide secure, centralized storage of DNA barcode sequences and their associated metadata. It employs strict data standards to enable easy retrieval and comparison of records from various sources. BOLD also provides analytical and quality control tools to help researchers explore and analyze the data, streamlining the overall management process.

BOLD is designed to facilitate collaboration among researchers working with DNA barcodes. It meets this goal by providing a simple path to create projects and assigning roles and permissions to team members, enabling seamless data sharing. Furthermore, it supports data visualization and export, helping researchers communicate their results to a broader audience.

In addition to fostering collaboration, BOLD supports the publication of DNA barcode records. Researchers can publish records within BOLD, enabling other researchers to integrate them into broader analyses. As well, it eases the submission of

data to GenBank by handling all aspects of the submission process, including data reformatting and deposition. These functionalities make BOLD an invaluable tool for researchers, as it enables the sharing and dissemination of data in a standardized format.

1.5 Registry for Dark Taxa

The Barcode Index Number (BIN) system [5] is a key feature of BOLD. Its development was propelled by a need to address dark taxa [19], the millions of species lacking a scientific name. By providing a standardized method for clustering sequences into persistent operational taxonomic units (OTUs), the BIN system brings objectivity and reproducibility to species delineation and identification, bridging the gap between traditional and molecular taxonomy.

The BIN system employs an algorithm that combines single linkage and Markov clustering to group DNA barcode sequences into OTUs based on their similarity. Clustering is performed primarily on COI sequences, most derived from animals. The system utilizes the refined single linkage (RESL) algorithm [5], which considers both cohesion within a group of sequences and their stability to enable a highly reliable method for species delineation.

The BIN registry is updated monthly in response to new data and modifications to existing records. When a new BIN is recognized, it is assigned a unique alphanumeric identifier, and a corresponding webpage is established on BOLD (*see* Fig. 1). This page assembles all records for the BIN and aggregates information on their taxonomic assignments, images, and collection points.

The effectiveness of the BIN system as a proxy for species has been demonstrated in diverse taxonomic groups [20–23]. It employs uniform resource identifiers (URIs) and web service functionality to support integration with other databases, such as GenBank and GBIF. Deeply embedded in BOLD, BINs are readily accessed via consoles and analytical tools.

Since its launch, the BIN system has gained broad adoption, being cited in more than 2000 publications. It plays an increasingly important role in accelerating taxonomic research and species descriptions [24, 25]. It also makes it possible to evaluate patterns of biodiversity in groups that have received little taxonomic attention. With continued application and further development, the BIN system promises to fast track the registration of all animal species.

2 Data Model

BOLD employs a unique data model (*see* Fig. 2), one combining DNA sequence data with fields traditionally captured by museums (image and site/date of specimen collection). The initiation of each record requires two data elements (a unique Sample ID (e.g.,

Enter search terms BINS

Specimens: [DWC](#) [XML](#) [TSV](#)
 Sequences: [FASTA](#) [TRACE](#)
 Combined: [XML](#) [TSV](#)

Barcode Index Number Registry For BOLD:AAC1842

Go to: [Public records in this BIN](#)

BIN DETAILS

BIN URI:	BOLD:AAC1842	Average Distance:	0.08% (p-dist)
DOI:	dx.doi.org/10.5883/BOLD:AAC1842	Maximum Distance:	0.18% (p-dist)
Member Count:	23 [23 Public]	Distance to Nearest Neighbor:	1.2% (p-dist)
Barcode Compliant Members:	21		
Founding Record:			

NEAREST NEIGHBOR (NN) DETAILS

Nearest BIN URI:	BOLD:ACE3418	Average Distance:	0.29% (p-dist)
Member Count:	4	Maximum Distance:	0.48% (p-dist)
Nearest Member:	FAHAA3657-18	Distance Variance:	0.03% (p-dist)
Nearest Member Taxonomy:	Arthropoda, Insecta, Lepidoptera, Hesperilidae, Eudaminae, Eudamini, <i>Astraptus</i> , <i>Astraptus enotrus</i>		


TAXONOMY

- Phylum:** Arthropoda [23]
- Class:** Insecta [23]
- Order:** Lepidoptera [23]
- Family:** Hesperilidae [23]
- Subfamily:** Eudaminae [23]
- Genus:** *Astraptus* [23]
- Species:** *Astraptus janeiraDHJ02* [23]
- Subspecies:**

DISTANCE DISTRIBUTION

SPECIMEN IMAGES:

BIN COMPLIANT WITH METADATA REQUIREMENTS

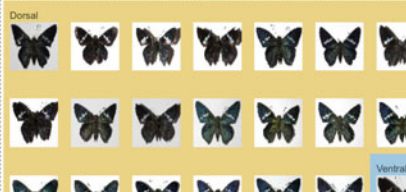


MHMYE1437-09 (*Astraptus janeiraDHJ02*)

License: Creative Commons Attribution NonCommercial ShareAlike (2009) [\[CC BY-NC-SA\]](#)

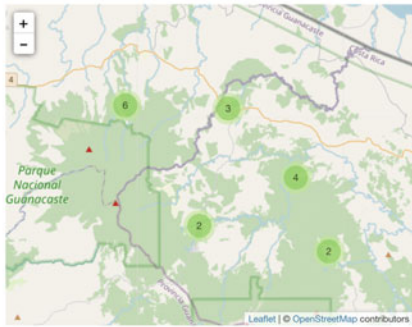
License Holder: Daniel H. Janzen, Guanacaste Dry Forest Conservation Fund

Dorsal



Ventral

COLLECTION SITE:



ATTRIBUTION

Specimen Depositories:

- Research Collection of D. H. Janzen & W. Hallwachs - [5]
- University of Pennsylvania - [18]

Fig. 1 BIN page for BOLD:AAC1842, which represents an interim species, *Astraptus janeiraDHJ02*

Specimen			Sequence	
			Process ID	
			Genbank accession number	
			Sequence upload date	
			ProcessID minted date	
			Recordset code array	
Attributes	Taxonomy	Collection Location	Marker Sequence	Traces
Sex Life stage Extra info Tissue type Reproduction Storing institution Voucher type Sampling protocol Associated specimens Associated taxa Funding source Specimen linkout	Kingdom * Phylum Class Order Family Tribe Genus Species Subspecies References Identified by Identification method Identifier email Notes	Country Country code Province Region Sector Site GPS coordinates Elevation/Depth Location accuracy Collectors Collection date Collection time Collection accuracy Collection code Site code Collection event ID Notes	Gene Sequence Marker code	Quality scores
Images			Primers	BINS
Orientation Attribution Licensing			Forward primer sequence Forward primer code Reverse primer sequence Reverse primer code	Unique identifier Nearest Neighbour Sequence variation

Fig. 2 The BOLD data model with specimen-related data elements in green and sequence-related ones in orange. * Fields required to instantiate a record

BIOUG30290-H09) and a high-level taxonomic assignment (e.g., phylum). Once the Sample has entered BOLD, further data elements can be added as they become available. Moreover, the extension of existing records with new data (e.g., sequences for more genes) does not impact populated fields. In addition, BOLD supports flexible curatorial workflows where access to distinct data elements can be managed independently. For example, permissions to modify the identification of specimens can be provided to taxonomic specialists, while being withheld from molecular specialists who would, on the other hand, be granted access to modify sequence data.


2.1 Specimen Data

BOLD provides a page that consolidates all data elements associated with each specimen (*see* Fig. 3). This Specimen Page serves as an up-to-date representation of the record, instantly reflecting any modifications to the underlying data elements. To speed curation, all components of each record can be edited directly from the Specimen Page.

Modifications to records are tracked through versioning, ensuring a comprehensive audit trail. The system maintains a history of changes in two ways: (i) an activity log tracks change events and the user(s) responsible for them, offering a comprehensive chain of modifications; (ii) the Delta View displays the history of each record, highlighting changes (or deltas) between weeks, thereby enabling efficient monitoring and comparison of alterations to each record over time.

06-SRNP-2459 - HesperIIDae of the ACG 7 [MHAHG]

06-SRNP-2459
2



Daniel H Janzen
Creative Commons Attribution NonCommercial
ShareAlike (2013)
Guanacaste Dry Forest Conservation Fund
djanzen@sas.upenn.edu

Tags
Comments
New Comment

Process ID: MHAHG656-06
Identification: *Astraptes tucuti* (Williams, 1927)
Identified by: John Burns
Collected in: Costa Rica, Alajuela
by: Eida Araya
Institution Storing: University of Pennsylvania
Field ID: 06-SRNP-2459
Museum ID: 06-SRNP-2459

Edit Specimen Show Delta View

3
5

Recent Activities

25 records per page Search:

Timestamp	Who	Action
Oct-04, 2018 15:13	BOLD Data Manager	Modify-Specimen
Nov-19, 2015 11:02	BOLD Data Manager	Modify-Specimen
May-02, 2013 09:33	BOLD Data Manager	Modify-Specimen
Apr-10, 2013 14:27	Megan A. Milton	New-Image(s)
Apr-10, 2013 14:27	Megan A. Milton	New-Image(s)
Apr-10, 2013 14:27	Megan A. Milton	New-Image(s)
Apr-10, 2013 14:27	Megan A. Milton	New-Image(s)
Apr-10, 2013 14:27	Megan A. Milton	New-Image(s)
Apr-10, 2013 14:27	Megan A. Milton	New-Image(s)
Apr-10, 2013 14:27	Megan A. Milton	New-Image(s)
Apr-10, 2013 14:27	Megan A. Milton	New-Image(s)
Apr-10, 2013 14:27	Megan A. Milton	New-Image(s)
Apr-10, 2013 14:27	Megan A. Milton	New-Image(s)
Apr-10, 2013 14:27	Megan A. Milton	New-Image(s)
Apr-10, 2013 14:27	Megan A. Milton	New-Image(s)
Apr-10, 2013 14:27	Megan A. Milton	New-Image(s)

Download History: Last Week Last Month Last 6 Months

Specimen Details

Sample ID: 06-SRNP-2459	Voucher Status:
Process ID: MHAHG656-06	Tissue Descriptor:
Project: MHAHG	Sex:
Institution Storing: University of Pennsylvania	Reproduction: S
Field ID: 06-SRNP-2459	A
Museum ID: 06-SRNP-2459	Life Stage: Uretra elata
Collection Code:	Extra Info: Host: Uretra elata
Reference Link:	Associated Taxa:
Note: Urticaceae	Associated Specimens:

Taxonomy

Phylum: Arthropoda	Identification: <i>Astraptes tucuti</i> (Williams, 1927)
Class: Insecta	Rank: Species
Order: Lepidoptera	Identifier: John Burns
Family: HesperIIDae	Identification Method:
Subfamily: Eudaminae	Identifier Institution: Smithsonian Institution
Tribe: Eudamini	Identifier Email: burns@si.edu
Genus: <i>Astraptes</i>	Taxonomy Note:
Species: <i>Astraptes tucuti</i>	


Barcode Index Numbers

BIN: BOLD:AAB4612	Phylum: Arthropoda [33]
Type: Member	Class: Insecta [33]
Max Divergence in BIN: 0.17% (p-dist)	Order: Lepidoptera [33]
Distance to NN: 3.8% (p-dist)	Family: HesperIIDae [33]
	Subfamily: Eudaminae [33]
	Genus: <i>Astraptes</i> [33]
	Species: <i>Astraptes tucuti</i> [33]

Collection Data

Country: Costa Rica	Collector: Eida Araya
Province/State: Alajuela	Date Collected: 27-Apr-2006
Region/Country: Area de Conservacion Guanacaste	Date Accuracy:
Sector: Sector San Cristobal	Time Collected:
Exact Site: Camino Brasilia	Site Code: DHJACG
Lat/Lon: 10.93, -85.372	Habitat:
Elevation: 500 m	Sampling Protocol:
Elevation Accuracy:	Coord. Source:
Depth:	Coord. Accuracy:
Depth Accuracy:	
Collection Event ID:	
Collection Notes:	

Map



4
7

Comments

New Comment

Feedback

Fig. 3 This Specimen Page for an individual of *Astraptes tucuti* (HesperIIDae) shows the seven standard elements: (1) Identifiers and attributes; (2) Photograph(s); (3) Image annotations as tags and comments; (4) Collection site; (5) Specimen details, including links for editing specimen information and accessing the Delta View; (6) Activity report; and (7) Comments. All elements on each Specimen Page can be modified via the Edit Specimen function available below the specimen details

Directly modifying fields is just one way to manipulate data on BOLD; each specimen record can also be annotated with tags and comments. For instance, if a user believes the identification of a specimen is incorrect, a tag can be added, such as “Suspicious Identification,” as an alternative to removing the present taxonomic assignment. This function allows for further contextualization and collaboration among users.

By offering direct editing capabilities, comprehensive change tracking, and annotation, the Specimen Page ensures efficient data curation, enhancing the overall usability and reliability of the system.

Specimen images, although optional, have become a critical component of records on BOLD as they play a crucial role in quality assurance and taxon profile construction. They also enable the inspection of morphological characters, enhancing the credibility of taxonomic assignments.

Guidelines are provided to ensure the comparability of images. Employing strict image orientation and standard aspect ratios enables the simultaneous comparison of thousands of images by constructing image arrays. Images are often uploaded to BOLD before sequence analysis, enabling their use in sequence validation.

Image licensing is determined by the image owner and can be modified. The International Creative Commons–Attribution–Share Alike license is advocated as it balances protection and accessibility. It is used as the default license when none is provided.

2.2 Sequence Data

The Sequence Data page provides a consolidated view of all fields of a sequence record together with supporting information (*see* Fig. 4). Accordingly, it displays the code(s) for the targeted gene region(s), the sequences, the PCR primers used to amplify them, as well as the trace files. When multiple markers are submitted for the same specimen, the data for each marker is displayed separately. Similar to the Specimen Page, users can modify every element of the sequence data directly from the Sequence Page. Modifications are tracked through an Activity Log and Deltas.

Sequences frequently feature associated electropherograms (traces) and, if available, quality scores (e.g., Phred). With the broad adoption of second and third-generation sequencing, the support for the reads in FastQ format [26] underlying their consensus sequences will soon be added. While these files are necessary for records to attain barcode status, they also provide valuable information about sequence quality. BOLD stores failed traces for record keeping, while other results are placed in three quality classes (Low, Medium, and High). Prior to uploading traces, specimen records must be established, and the primers used to generate the sequences must be registered in the BOLD Primer Database.

Sequence View for Process ID: MHAHG656-06

2
Activity Report
Show Delta View

Specimen Details Current

Sample ID: 06-SRNP-2459
Process ID: MHAHG656-06
Project: MHAHG
Tax Names: Arthropoda, Insecta, Lepidoptera, Hesperidae, Eudaminae, Eudamini, Astraptes, Astraptes tucuti
Taxon: Astraptes tucuti
Rank Name: species
Sampling Protocol: N/A
BIN URI: BOLD:AAB461.2
Kingdom: Animalia

Marker Summary

Marker Code	Sequence Length	GC	Ambiguous	Trace Count
COI-5P	658	30.1%	0%	2

COI-5P

Illustrative Barcode

Nucleotide Sequence

```

NNAACTTATATATTTATTTTGGAAATTCAGCAGGATTAATGGAACTCTCTTAAGATTACTATCCGAACTGAATFAGGAATCCAGGATCTTAAATGGAGAT
GATCAAAATTTATATCTATTTTACAGCCGATCAATATATATATTTTATGATTTTATATATATGAGGATTTGGATTTGGATTTGATCTTGGCTTGGCTTGA
ATAATAGGATGCCCCGATATAGCTTATCCCGCATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATAT
GATCTCCGACCCGATGAGCAATTTATCCCGCATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATAT
TCATCAATTTCTGGAGCTATTAATTTTATACAGAAATTTAAATATGCGAAATTAATAGATATCTTTTGATCAAAATFACCTTATTTGATGAGCTTGGAAAT
ACAGCATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATAT
GGAGATCCAAATTTATACCAACTTATTT
                    
```

Amino Acid Sequence

```

TLVPIFGIMAGLIGTSLILLITELGTPGSLIGDQIYNTVITAFIMIFPHVVDIMIGGQNLVPLHMGAPDHAFFRNNHNFVLLPPELTLISSEIVENG
AOTMVTYVPLSENIHQGASVDAIFSLRLEAGISSILGAINFITTIINRINSEISFDQMLFVNAVGTALLLLSLPLVLAGATMLTDRHMLTSTFFPDAGG
DPILVQLHF
                    
```

Sequence Metadata

Genbank: GU151178
Accession:
Translation Matrix: Invertebrate Mitochondrial
Last Updated: 2006-10-06 14:51:04.340945
Sequence Runsite: Centre for Biodiversity Genomics

Identify Sequence:

[Full DB](#) [Species DB](#) [Published DB](#) [Full Length DB](#)

COI-5P Tags & Comments

+/-

New Comment

Sequencing Date: 2006-09-26 22:41:15
Trace Direction: F
Forward Primer: LepF1
Reverse Primer: LepR1
Sequence Primer: LepF1
Status: high qual
Trace Runsite: Biodiversity Institute of Ontario

Trace Tags

+/-

Trace Comments

New Comment

Fig. 4 Sequence Page for a representative of *Atraspes tucuti* (Hesperiidae) with the following information: (1) Specimen details; (2) Links to Activity Report on the sequence and a Delta View; (3) Marker(s) summary: when there are additional markers, their information is displayed independently; (4) Marker(s) information details, including the nucleotide and amino acid sequence and its metadata. As with the summary, when there is more than one marker, their information is displayed independently; (5) Trace file(s) are displayed and available for download

The Primer Database is a registry of all primers employed to amplify a barcode region (e.g., COI, matK, rbcL, and ITS2). It is mandatory to register primers before assembling the sequence submission package. When an uploaded primer sequence matches an existing sequence with a different name, BOLD reuses the original sequence and provides its registered code for association with sequence records.

Users can upload and update sequence records at any time. Once on BOLD, sequences are instantly viewable, downloadable, and ready for analysis.

3 Data Management

BOLD employs a flexible data management and access model to support a high volume of data and diverse workflows. The system utilizes a hierarchical structure to support the organization of records (*see* Table 2). After login, users are directed to the Main Console, a command center that facilitates data monitoring and management. From it, users can create new projects or access existing ones. Since projects serve as the primary management structure on BOLD, all new records require a designated home project. Each project is equipped with an access control list (ACL) under the control of the Project Manager (the user who created the

Table 2
Data management structures supported by BOLD

Structure	Definition	Use cases
Project	As the fundamental management unit, it holds records and supports access control. Each project has a manager who controls access to it. All new records require a project for deposition and maintain this affiliation through a shared code.	Data generation and validation. Collaboration with a small team or within a single laboratory.
Folder/ container	Folders have similar interfaces to projects but contain projects instead of records. Containers have separate access control allowing data to be shared with users operating across a set of projects.	Data lifecycle management. Collaboration across small networks. Reporting and auditing.
Campaign	Provides an organizational structure that groups folders and projects based on theme or initiative. This structure does not directly support access control but provides a way to affiliate projects for reporting or analysis.	Reporting. Collaboration across large networks.
Dataset	Provides a container for records deriving from multiple projects with a taxonomic, geographic, or thematic focus. Datasets are typically used to support publications due to their ability to mint DOIs.	Publication. Data curation.

Table 3
Five options that control access to varied domains of the records in a project

Access	Definition	Appropriate use
Analyze	Allows users to execute various analyses on the sequences in a project.	The option should be made available to all users on a project.
View & download	Allows users to view sequences and download them.	This option should be made available to users who require the capacity to export data for analysis using third-party tools or for sharing outside BOLD.
Edit sequences	Allows users to edit, overwrite, and delete sequences.	This option should be made available to those generating sequences as it allows them to upload and update sequences. It also allows data curators to repair or excise erroneous sequences.
Edit specimens	Allows users to modify any data element associated with the specimen record including taxonomy and geography.	This option should be made available to those assembling specimen records and to those with taxonomic expertise.
Add to dataset	Allows users to add new records into a dataset under their control, giving them the ability to share or publish the records.	This option should only be made available to users who require the capacity to add records into datasets for publication.

Each option can be activated independently except Add to Dataset as it also includes the Analyze and View & Download options.

project). The ACL supports multiple access levels (*see* Table 3), ranging from the most restrictive (read-only) to the most permissive (modify any data field).

While individual projects often meet the needs of a single researcher or small teams, large-scale projects and research networks necessitate additional structures to support complex data ownership, lifecycles, and reporting. Creating multiple projects and organizing them using the Folder or Campaign structure on BOLD is advisable in such cases. The folder structure aids project organization and introduces an additional ACL independent of its contained projects. This functionality is particularly useful when periodic review is required by individuals who are not directly involved in the routine management of the data.

Research networks often involve dozens of researchers, and effective data management necessitates multiple projects and folders. In such cases, reporting becomes challenging as the project count can exceed 100. The campaign structure was explicitly designed to address this situation. Although lacking an ACL, the campaign structure includes an Application Programming Interface (API) that provides data feeds to national or network databases and dashboards external to BOLD.

The final structure (dataset) is available to assemble records from diverse projects, often in support of a publication. Although

datasets fall outside the hierarchy of the other data management structures, they employ an ACL. Unlike other structures, they can pool records from multiple projects and folders to provide maximum flexibility and integration.

3.1 Projects

Projects are the core management unit within BOLD as they provide a secure means of data management. This organizational level streamlines workflows by providing an efficient environment for loading, storing, and tracking records. Each project is identified by a unique code of three to five letters that also forms the prefix for the Process ID affiliated with every record in it.

The Project Management console (*see* Fig. 5) grants real-time access to key components of the project. It enables immediate

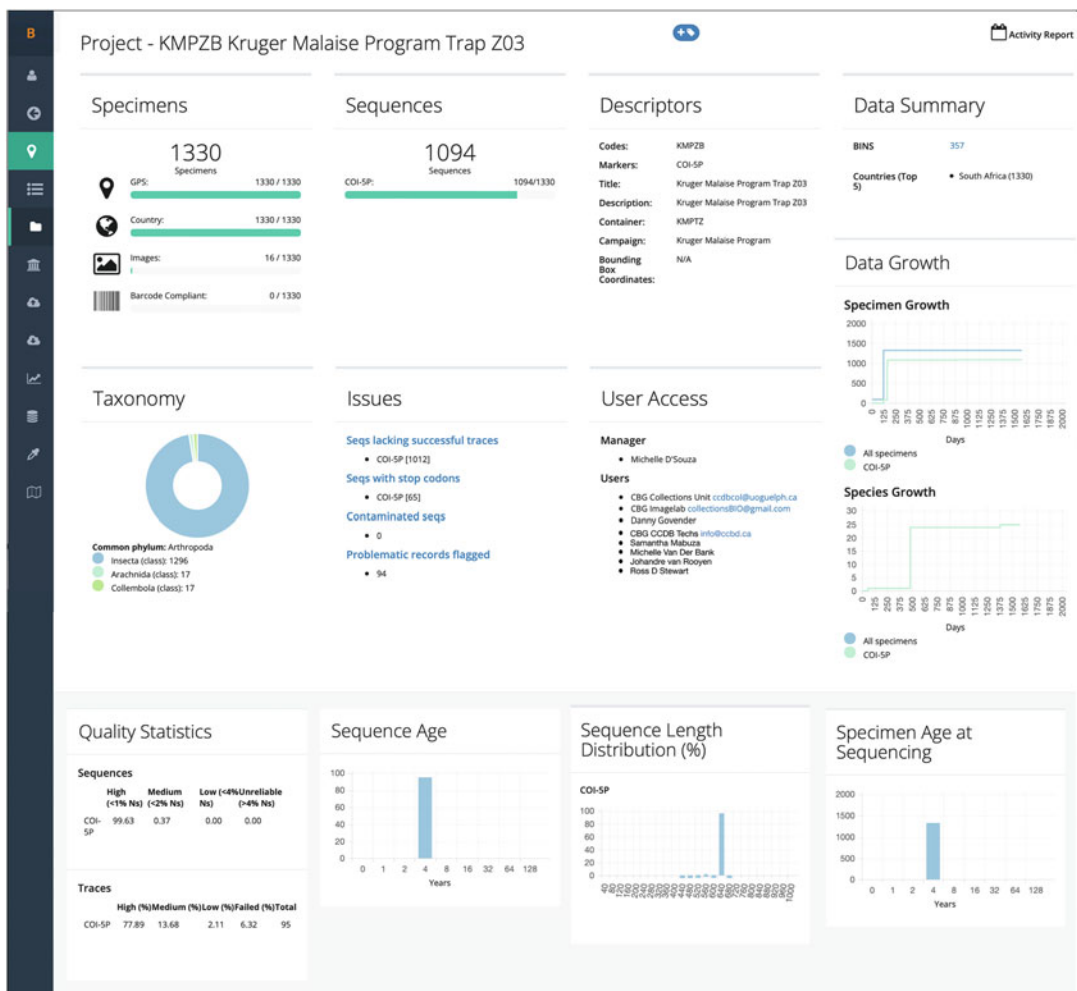


Fig. 5 The Project Console provides an overview of key metrics including the number of specimens, sequences, and BINs associated with each project. It also displays important statistics such as the success in sequence recovery, the BIN discordance rate, and a histogram of sequence lengths. Finally, it highlights issues associated with records in the project

monitoring of newly added records, verifies the completeness of data, and swiftly detects quality issues. Importantly, this console does not simply identify error identification; it allows their correction. Upon identifying data quality concerns, it generates clickable links that direct users to problematic records, expediting the resolution of errors.

Access to the Project Console and the data within each project is strictly controlled via an ACL that serves a dual function. It upholds data integrity by restricting access to authorized users and lists them together with their respective access levels. Project managers possess the capability to revise these access rights, allowing them to alter the ACL to meet new requirements for the project whether this involves revoking access rights to some existing users or conferring permissions to new ones.

An essential feature of each project is the production of an Activity Report. This report provides a detailed log of all data uploads and modifications made to the records within it. Offering a transparent audit trail guarantees traceability. The Project Console includes a range of analytical tools accessible through the main menu to facilitate data exploration and validation.

BOLD provides a drill-down function termed the Record Console (Fig. 6) for detailed data inspection. This feature allows users to view and interact with individual records within a project. Records can be accessed via links leading to their Specimen or Sequence Pages. Links are also available to BIN pages when a record has been assigned to one. Additionally, each record is annotated with error and compliance flags, which assist in identifying issues such as contamination, the presence of stop codons (for protein-coding sequences), misidentification, and non-compliance with barcode standards.

The Record Console uniquely allows users to select a subset of records within a project for download, for analysis using the integrated tools, or for addition to a dataset.

3.2 Datasets

The dataset feature on BOLD significantly enhances data organization and accessibility. It shares some functionalities with those available in projects but several key differences make it a valuable supplement to BOLD's data management system.

In contrast to projects, which function as the primary record containers, datasets operate as independent collections. They allow users to categorize and assemble records based on multiple criteria or to support specific research goals. Although any user can create a dataset with published records, the inclusion of private records requires permission from the manager of the project containing those records.

Once a dataset is created, its manager can share it with other users. Dataset access is governed by an ACL, akin to that used for projects. The dataset ACL manages who can access and interact

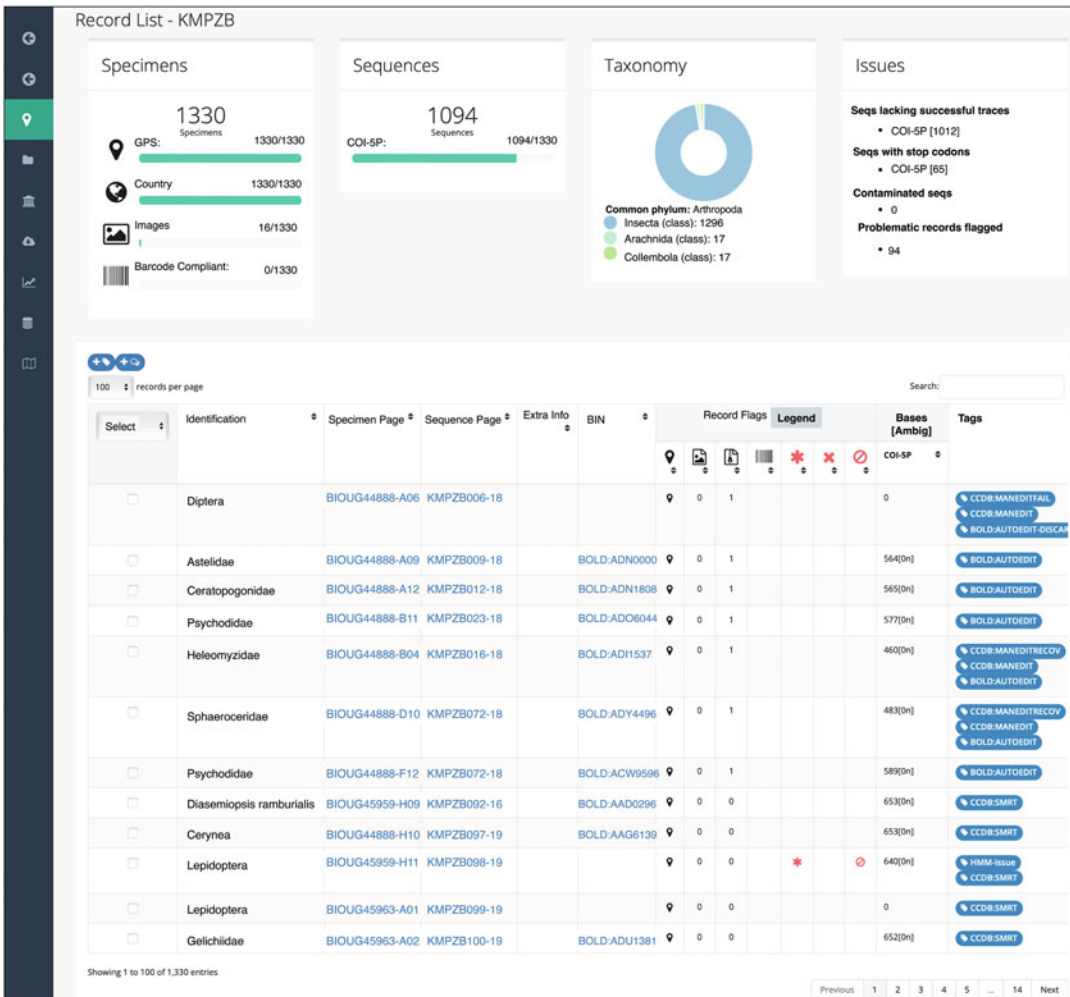


Fig. 6 The Record Console provides general statistics about the composition and completeness of the records in a project. In addition, a summary of detected issues is highlighted. A data table displays individual records and their features (presence of GPS coordinates, images, and trace files if they are barcode compliant, stop codon, contamination, flagged). It also displays the tags (if any) associated with each record. Records and associated BINs are accessed by clicking on their identifiers

with the dataset, reinforcing the platform's dedication to data integrity and secure collaboration.

The dataset feature incorporates functionalities available to conventional projects, including analysis tools and drill-down capability. Each dataset is accessed via the Dataset Console, similar to the Project Console but with additional functionality. One key feature is the ability to generate Digital Object Identifiers (DOIs) to facilitate data packaging for publications. The resultant DOIs can be directly cited in the manuscript, offering a straightforward, robust method for linking to the data. Each DOI provides an immediate link to the dataset, allowing reviewers to view and

download the relevant data. The integration of DOIs within datasets enhances data discoverability and traceability (Mannheimer 2016), making it an invaluable tool for scientific publications and data sharing.

4 Data Lifecycle

BOLD incorporates a life cycle approach to data management [14]. It facilitates the integration of specimen information and molecular data, such as sequences, primers, and traces along with other elements like images and maps in a centralized repository. This structured, unified approach helps to eliminate data silos, enhances data sharing and interoperability, and increases resilience [15] (Gilman et al., 2009). Notably, each phase of the record life cycle can be performed at any time, supporting dynamic data transformations based on updates and evolving requirements. Detailed discussions about the data life cycle (*see* Fig. 7) follow this section.

4.1 Specimen Submission

The assembly of a complete record involves multiple stages but commences with specimen submission. To establish a record, certain minimum data elements are required, including (i) a unique Sample ID; (ii) Field ID and/or Museum ID, the identifiers assigned to a specimen in the field or upon institutional accessioning; (iii) the name of the storing institution; (iv) the specimen's phylum; and (v) the country of collection. Additional elements can be incorporated/updated throughout the record's lifespan.

BOLD accepts submissions of specimen data through an online form for single or small batches of records or via a spreadsheet (<http://www.boldsystems.org/index.php/Resources/Templates>) for larger volumes. Although registered users can upload information to create BOLD records, a project must already exist or be created to house the data. As soon as records are generated, they are

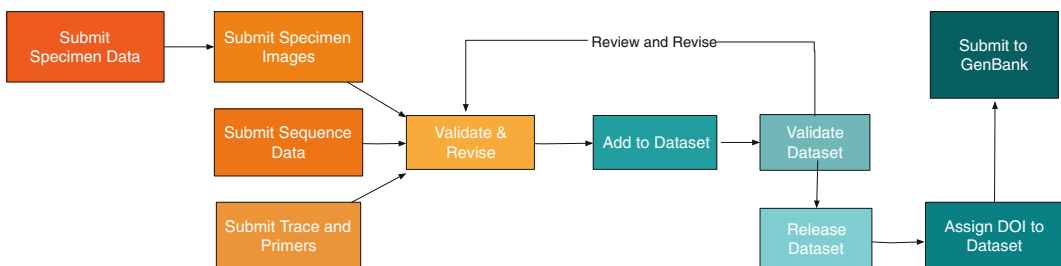


Fig. 7 Data Lifecycle The data lifecycle on BOLD starts with specimen data and ends with the publication of records and submission to public repositories

instantly viewable and editable by those with appropriate access to the project.

BOLD's ability to extend and update records with taxonomic, geographic, and other specimen data can be employed at any stage after record creation. The submission workflow allows extensions and updates in three groups: specimen descriptors, taxonomic assignments, and collection information. Updates submitted to BOLD replace the existing information with new data, but a historical record persists. BOLD tracks the details of each update (who, what, and when) to ensure a comprehensive audit trail.

Most submissions to BOLD proceed smoothly, but inaccurate or inconsistent data will prevent the upload process. BOLD conducts checks on all submissions to ensure the data is of high quality and fit for purpose. Data issues fall into three categories: consistency, completeness, and compliance (Table 4), and have prescribed remedies.

Submissions introducing terms new to BOLD require communication with BOLD staff. For example, adding new taxonomic names may lead to a request for supporting information. While such cases pause the process, they expand BOLD's controlled vocabularies or libraries, ensuring future submissions proceed smoothly. However, most triggers for investigation result from typographic or data entry errors, which must be corrected before submission proceeds. These checks are critical to maintaining data quality.

4.2 Image Submission

Images are a very important part of BOLD records because of their value in validating taxonomic assignments [27, 28] and as a source for trait data [29]. To maximize their utility, images should be uploaded early in the record assembly process. BOLD makes use of the images in two ways: (i) quality assurance, supporting the verification of taxon assignments and (ii) constructing taxon profiles, with BOLD displaying a subset of images in the Taxonomy Browser and on BIN pages. Plans are underway to leverage these images to facilitate taxonomic classification through AI methods [30].

Images may be submitted individually or in batches. Batch submissions require a Zip package containing an Excel file with the necessary information (Table 5) and a collection of JPEG images. Batch submissions can be performed via the Main Console or the Project Console. BOLD accepts high-resolution images of up to 20 megapixels with a maximum of 200 images per package. Larger submissions are permissible but require prior contact with BOLD's support team (accessible via support@boldsystems.org).

Each image requires a license with the default being the Creative Commons–Attribution–Share Alike, which deftly balances protection and accessibility. In the absence of a provided license, the default option is applied automatically.

Table 4
A list of common issues encountered during the submission of specimen data and their resolution

Categories	Checks	Resolutions
Consistency	<p>GPS coordinates in the correct format. Geographic names are consistent and match the registry. Date in the correct format. Specimen depository matching one registered on BOLD.</p> <p>Taxonomic nomenclature must match the BOLD backbone taxonomy.</p> <p>The identifier (individual providing taxonomic assignment) must be registered on BOLD with an affiliation.</p>	<p>Format errors are corrected where possible and the submission proceeds. If a correction is not possible, the submitter is contacted.</p> <p>New depositories are added once registration information is acquired and the submission proceeds. If information is unavailable, the submission is paused until the submitter provides it.</p> <p>New scientific names are verified by searching the literature. If found, they are added to BOLD, and the submission proceeds. If not, the submitter must provide a reference or withdraw the name.</p> <p>The Identifier's affiliation is ascertained and the submission proceeds. If this is not possible, the submission is paused until the submitter provides this information.</p>
Completeness	<p>GPS coordinates provided. Collection dates provided.</p> <p>Taxonomic names are provided with no gaps in the hierarchy.</p> <p>Specimen identifiers (Sample ID, Field ID, Museum ID) are unique and of sufficient length to maintain uniqueness.</p>	<p>Submission continues, but the record is annotated to indicate that key information is missing.</p> <p>Gaps are filled where possible based on BOLD's backbone taxonomy. Where a conflict is encountered, the submission is paused and the submitter is notified.</p> <p>Submissions failing to meet this criterion proceed only once addressed.</p>
Compliance	<p>Unique primary sample identifier (Sample ID). Country specified. Specimen depository specified. Collection coordinates at minimum precision (two decimals).</p>	<p>Submissions are paused until these compliance criteria are met. Exceptions are made in some cases (e.g., obfuscation of GPS for endangered species).</p>

BOLD provides a standard guide for photographing specimens. Although dorsal, lateral, and ventral views are ideal for animals, candid or unoriented specimen images are of considerable value. BOLD supports the inclusion of captions or other descriptors for each image. Photographs where diagnostic features are measured or displayed can often prove beneficial for identification. For plants, full-sheet photographs or scans of herbarium sheets with close-ups of diagnostic features are ideal (*see* Fig. 8).

Table 5
Metadata associated with images on BOLD

Data types	Field name	Definition
Identifiers	Image name ^a	Name of the file with the .Jpg extension. It should be identical to the image file name.
	Sample ID	Identifier of the specimen record previously created in BOLD.
	Process ID	Identifier of the record sequence in BOLD.
Attributes	Original specimen ^a	“Yes” if the image shows the actual specimen of the record. “No” if the photograph is a representative specimen of the same species, a photograph of the collection site, etc.
	Orientation ^a	Position of the specimen in the image. For example: Dorsal, ventral, or lateral.
	Caption	Free text description of the image with a limit of 400 characters.
	Measurement	Any relevant measurement in metric units.
	Measurement type	Feature that was measured.
License	License holder	The primary individual who holds the license.
	License ^a	BOLD supports eight (8) types of image license: Copyright No rights reserved CreativeCommons attribution CreativeCommons attribution share-alike CreativeCommons attribution no derivatives CreativeCommons attribution non-commercial CreativeCommons attribution non-commercial share-alike CreativeCommons attribution non-commercial no derivatives
	License year	The year of the license declaration. Limit of four characters.
	License institution ^a	Institutional affiliation of the primary license holder. If the individual holder is unresponsive/unreachable, decisions regarding the use of material fall to the institution.
	License contact	Contact information for the license holder. Email address, mailing address, phone number, or all the above. Limited to 128 characters.
	Photographer	The individual or team responsible for photographing the images. Limited to 128 characters.
	Attribution	

^a Denotes mandatory fields

Up to ten images can be uploaded for each specimen, allowing users to associate their records with images of live or mounted organisms, distinct features, hosts, or habitats.

Submission templates are available at <http://www.boldsystems.org/index.php/Resources/Templates>

4.3 Sequence Submission

The addition of sequence data is the final step in the assembly of a record. In order for sequences to be uploaded, their source specimen(s) must already be registered. The sequence region or marker should also be registered in BOLD’s marker database. New markers can be registered by contacting BOLD’s support team.

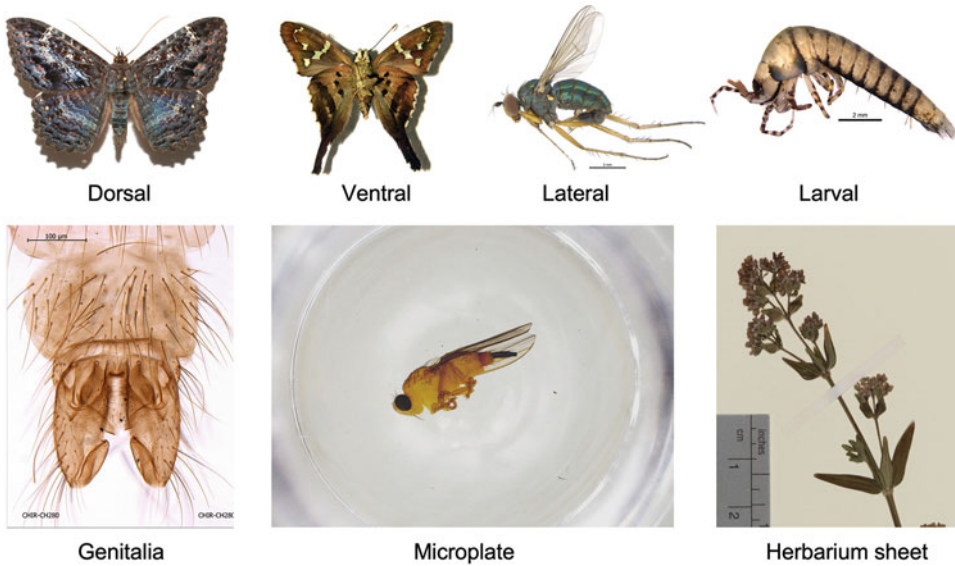


Fig. 8 Example specimen images and commonly utilized orientations. Dorsal, Ventral, Lateral, and Larval, Genitalia, Microplate, and Herbarium sheet orientations are depicted

Upon submission, sequences are immediately associated with records. In addition, BOLD automatically translates protein-coding sequences, flagging those producing stop codons. It also compares uploaded sequences for barcode regions (e.g., COI), against a contaminant library, such as human, mouse, cow, and pig. If the system detects a close match, it flags the records for review.

Sequence data can be uploaded via two access points: (i) The first and most frequently used is via the Project Console. (ii) The second entry point is via the Main Console, where a sequence upload option appears in the dashboard. Uploads require specific metadata (Table 6), including the marker being uploaded and the laboratory that generated the sequences. BOLD accepts up to 2000 sequences in a single upload. BOLD also allows for sequence editing, deletion, and replacement directly from the sequence page. To modify a sequence, navigate to the Sequence Page for the target sequence and click the Edit Sequence button.

4.4 Trace and Primer Submission

Trace files and primers help to ensure the verification and reproducibility of sequences submitted to BOLD. In addition, they fulfill a compliance requirement set by the BOLD data standard. Prior to the upload of trace files or primers, specimen records must be established.

Given their interrelated nature, trace files and primers associated with sequence records are uploaded in tandem. To register new primers, users need to complete a form with the information outlined in Table 7, which is accessible via the Data Uploads menu in the Main Console.

Table 6
Sequence information

Field	Definition
Identifier	Each sequence requires a unique identifier. Two options are available: Sample ID and Process ID.
Markers	The gene region from which the sequence originates. The name should be registered in BOLD's marker database (255 markers currently registered). Sequences from only one marker can be uploaded at a time. If the marker is not provided, the system will use the default value—The primary marker associated with the project.
Run site *	Name of the facility responsible for generating the sequence(s). If the facility is new to BOLD, it must be registered. Sequences from only one facility can be uploaded at a time.
Sequence *	One or more sequences in FASTA format. Ideally, sequences should be aligned, but this is not required. The FASTA define must conform to a strict format. It should start with ">", followed by the Process ID or Sample ID. Other data may be included in the define but must be preceded by a pipe (" ") to mark the termination point of the Process ID. For example: >Process ID other information ATGGCCTTTAATCG...

Data associated with sequence upload. Mandatory fields are denoted with an asterisk (*).

Table 7
Primer information. This table displays the data required for primer registration

Field name	Definition
Primer code	Code for the primer. Between 3 and 12 characters.
Marker	Gene region that the primer set amplifies. It should be one of the registered markers in the BOLD database.
Cocktail	Standard combination of multiple primers used in one PCR reaction. Two options: "Yes" or "No".
Primer sequence (5' to 3')	DNA sequence for the primer in the 5' to 3' direction.
Direction	Two choices: "Forward" or "reverse."
Reference/citation	Publication associated with the primer if one is available.

It is best to confirm that the primers linked to the trace files are registered in BOLD's Primer database prior to uploading. There are two main methods to submit this data to BOLD: (i) Single Record Trace File Upload: Available from the Sequence Page, this option allows you to upload trace files for a single record directly. Ideal for a small batch of trace files, this feature accommodates a maximum of four electropherograms and four optional Phred/score files per locus for each record. The association between traces and the records is immediate upon submission. (ii) Batch Trace File Upload: This method requires grouping trace files into a folder,

Table 8
Trace file information. Data required for trace file submission

Field Name	Definition
Trace file	Complete name, including extension (“ab1” or “scf”) identical to the file name in the folder.
Score file	Complete name, including extension (.phd.1) identical to the file name in the folder (if available).
PCR primers Fwd/rev	Primer codes for the forward and reverse primers. Both fields must be present. Primer codes are case sensitive.
Sequencing primer direction	Forward or reverse direction.
Sequencing primer	Primer used for sequencing.
Process ID	Identifier of the sequence in BOLD.
Marker	Gene region amplified by the primer set. It should be one of the registered markers in BOLD. For a multi-marker project, this field must be completed, if not, the traces will be uploaded to the primary marker.
Sequencing lab (<i>only for form-based trace for the single record</i>)	Name of the facility generating the sequence(s). If the institution is not listed, it must be registered.

coupled with associated metadata (Table 8) in a spreadsheet form, and creating a single zip file for upload. The batch method permits up to 10 trace files per record, with a total limit of 400 files per package. All trace files included in the upload must be listed in the associated spreadsheet form.

4.5 Record Validation

Validation is a critical phase in the life cycle of every record on BOLD. This stage employs a blend of automated and manual checks to ensure the completeness and integrity of each record.

There are two stages to validation. Initial validation occurs during data upload, but it is followed by a second stage during reconciliation with other records. During the first stage, BOLD largely automates the process, checking each record for completeness and comparing data against controlled vocabularies. This system swiftly flags discrepancies, allowing submitters to rectify errors. For instance, country names not recognized by the database are checked for spelling errors. BOLD performs similar automated checks on specimen data during upload, repairing or removing erroneous data elements after their recognition. However, automated detection can be difficult because of the complexity of certain data fields such as species names or detailed descriptors like the Exact Site. As a result, responsibility for comprehensive quality inspection rests with the data submitters and project managers.

Errors identified in specimen data usually involve typographic or transliteration mistakes, which can be easily corrected and are often addressed automatically. However, sequence data errors require a more in-depth investigation before resolution. For example, protein-coding sequences that yield stop codons upon translation could indicate sequencing errors or the presence of pseudogenes, necessitating the inspection of the sequence and its comparison with sequences from related organisms.

The second validation stage commences once specimen and sequence data have been uploaded to the record. This stage involves an evaluation of the record's validity by comparing it with existing records, either from the same project or from published data. BOLD performs some of these checks in an automated manner, such as scanning new sequences against a contaminant library. However, much of this validation work is carried out by users through data analysis. To assist with user-driven record validation, BOLD provides a suite of integrated analytical tools. Frequently used tools include the Taxon ID tree, Distribution Map, and Batch ID Engine.

The Taxon ID tree tool generates a Neighbor Joining tree [31, 32], encompassing all project records. The tree output, labeled with record taxonomy, enables easy comparison between tree topology and record taxonomy. Selecting the Matching Images option enhances this tool's utility as it generates an image array that is in the same order as the records in the tree (Fig. 9). This feature provides a morphological context to tree placement, offering an immediate indication of misplaced nodes. It also organizes images based on their sequence similarity, making instances of contamination or misidentification readily apparent due to their contrast with neighboring images.

The Distribution Map tool projects specimen coordinates onto a map, facilitating the swift identification of errors in coordinate values. This tool also supports direct correction of errors from points on the map, streamlining the removal or correction of erroneous coordinate data.

Lastly, the Batch ID Engine complements the function of the Taxon ID tree. It compares sequences in a project against all data on BOLD, detecting discrepancies based on the assumption that established records on BOLD have accurate taxonomy.

4.6 Dataset Assembly

BOLD supports the creation of customizable datasets that can, for example, assemble records from a particular taxonomic group or geographic region. Because a dataset can assemble both public and private (with permission) records, users have considerable flexibility in constructing datasets. Although these assembled datasets serve various functions, their primary utility lies in streamlining data publication.

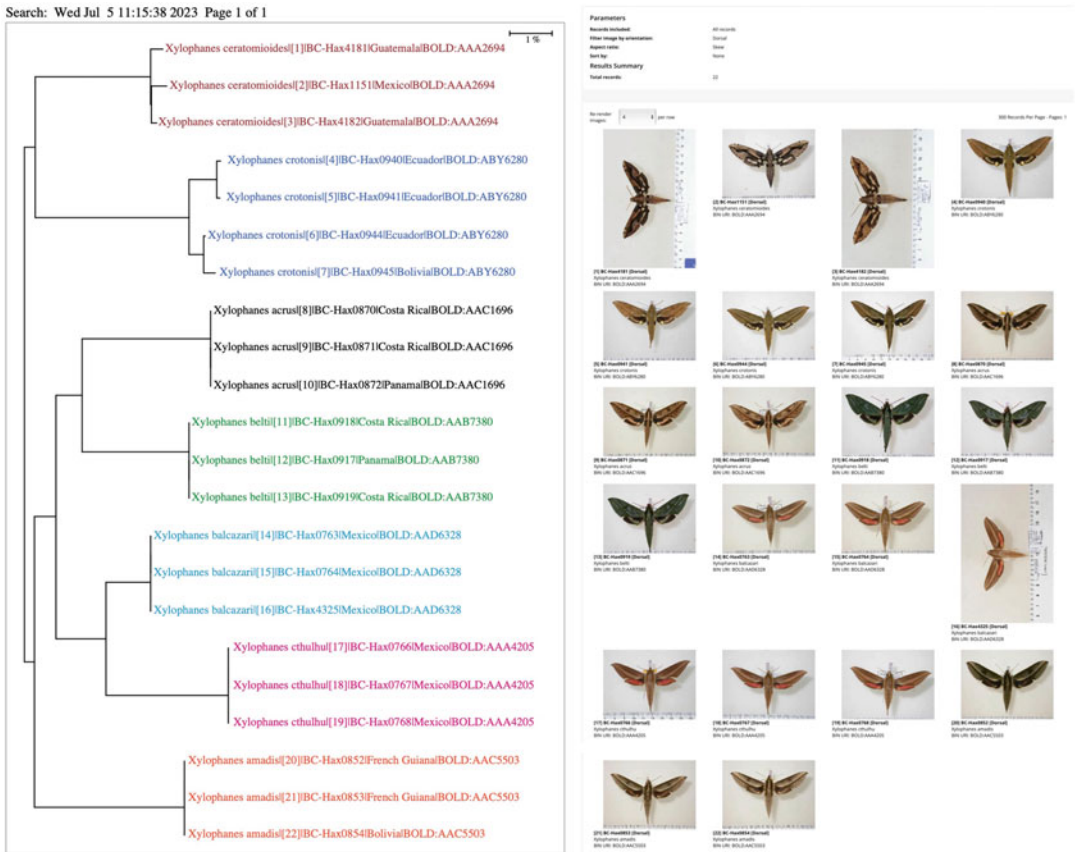


Fig. 9 Tree Analysis with Matching Images. The left panel displays the Neighbor Joining tree, resulting from the Taxon ID analysis of 50 records belonging to the genus of *Xylophanes* (Lepidoptera: Sphingidae) (project code: JHPUB). The tree is color-coded to represent the different BINs. The right panel shows the corresponding Image library produced using the Matching Images feature

The construction of a dataset for publication often necessitates the extraction of validated records from several projects. Because of variability in the reliability of records, it is best practice to only select records with high quality and accuracy for inclusion in a dataset.

To assemble a specific set of records, one must first establish a dataset within BOLD. This process parallels that of the creation of a project as it is executed from the Main Console. Each new dataset must be given both a unique dataset code as an identifier and a concise description.

Once a dataset has been established, users can add records to it. For private records, users must request access to each project with relevant data, navigate to the record list, and select the records of interest. Integrating the chosen records into the dataset is completed by simply selecting the Add To Dataset option in the main menu on the left of the Record Console.

The expansion of a dataset with new records can occur progressively as they meet the requirements for inclusion. BOLD also allows for the inclusion of previously published records through a comprehensive search function, supporting queries based on taxonomy, geography, and other parameters. The dataset function plays a key role in supporting BOLD's efforts to promote data reuse.

After a dataset has been assembled, records should be comprehensively validated before publication. This stage is critical as it helps to uncover and rectify any inconsistencies that emerge when records from diverse sources are compiled. Thus, this final validation step ensures the dataset's integrity and high quality.

4.7 Dataset Validation

The assembly and publication of DNA barcode datasets on BOLD demand rigorous validation to ensure data accuracy and reliability before their release. Numerous challenges can arise, including poor quality data stemming from errors in DNA extraction, PCR amplification, sequencing, or contamination resulting in non-target sequences. Species misidentification is another common problem, often reflecting limited access to taxonomic expertise. Furthermore, a lack of multiple independent sequences from a species limits the evaluation of intraspecific variation, a critical matter for the discrimination of closely allied species. Incomplete annotations, such as missing collection dates, locations, or collector's names, can also impede data interpretation.

BOLD offers four tools to aid dataset validation before publication: (i) BIN Discordance, (ii) Barcode Gap Analysis, (iii) Project Summary, and (iv) Batch ID Engine.

The BIN Discordance tool (Fig. 10) can only be applied to BINs with two or more records. It identifies discrepancies in taxonomic assignments among the records in a BIN, which may arise from misidentification or contamination during analysis.

Complementing the BIN Discordance tool is the Barcode Gap Analysis tool (Fig. 11), which compares intraspecific and interspecific divergences for each species. As intraspecific divergences are usually much lower than interspecific ones [33], deviations warrant investigation as potential errors or biological explanations such as recently diverged species [34].

The Project Summary tool (Fig. 12) checks record completeness. It generates a report detailing the distribution of collection locations and a species table, pinpointing gaps in coordinates, taxonomy, and sequences. This tool supports publications associated with reference libraries by providing supplementary materials.

The Batch ID Engine tool (Fig. 13) assesses the accuracy of taxonomic assignments within a dataset. As it performs cross-validation with other sequences on BOLD, it is advisable to run all datasets through this tool before publication. Cases where

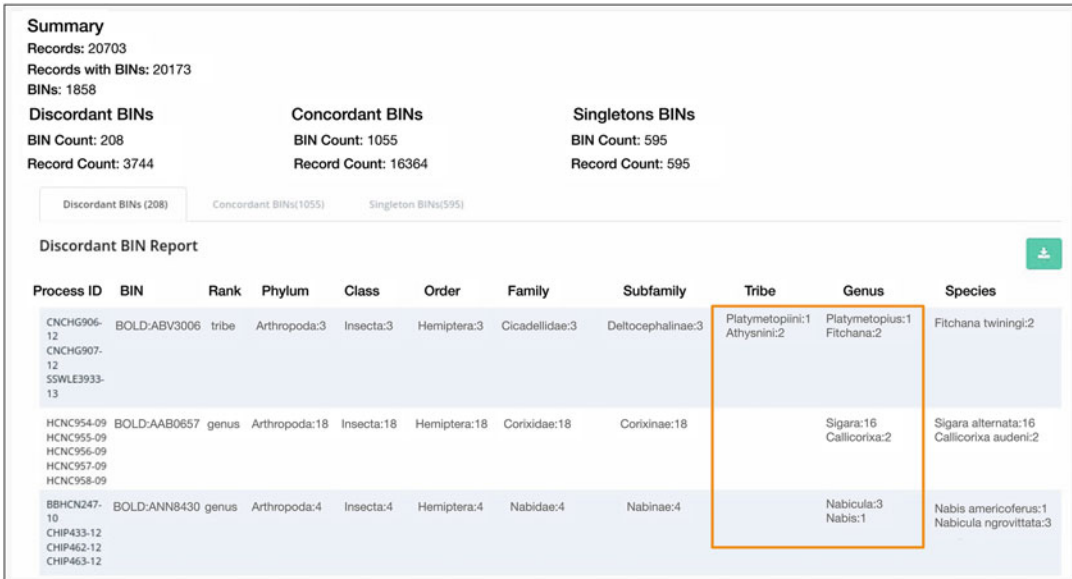


Fig. 10 BIN Discordance Report Summary: BIN discordance analysis for records of the order Hemiptera from Canada in the dataset DS-HECALIB. The orange rectangle highlights records that exhibit taxonomic discordance

queried records match different taxa often indicate analytical mishaps. Other mismatches may reflect errors in published datasets; such cases should be reported to the BOLD support team.

Compiling a DNA barcode library is a key task as it creates a reference for other researchers. Errors can lead to the replication of misidentifications and subsequent issues, emphasizing the importance of careful validation in creating a high-quality library.

4.8 Dataset Publication

The publication of BOLD records as accessible datasets is the final step in the lifecycle and an important step in making data widely available. Offering a path to enhance data visibility, this process changes data access through the BOLD workbench by shifting the access level associated with the dataset to public status in the Dataset Options in the main menu. This raises accessibility by making the data available to all users of BOLD, whether for direct access or retrieval via searches.

This publication of datasets is only the first step in achieving data visibility. The subsequent recommended measure involves the association of each dataset with a publication via a DOI. Offering unique identification, convenient data sharing, and seamless access, the DOI also ensures appropriate citation in associated research efforts, consolidating recognition of the dataset’s role in research. BOLD supports the minting of DOIs for each dataset at the time of publication.

Barcode Gap Analysis Result - Search: Sample IDs; Include public records (1005 records returned) (1005 records selected)

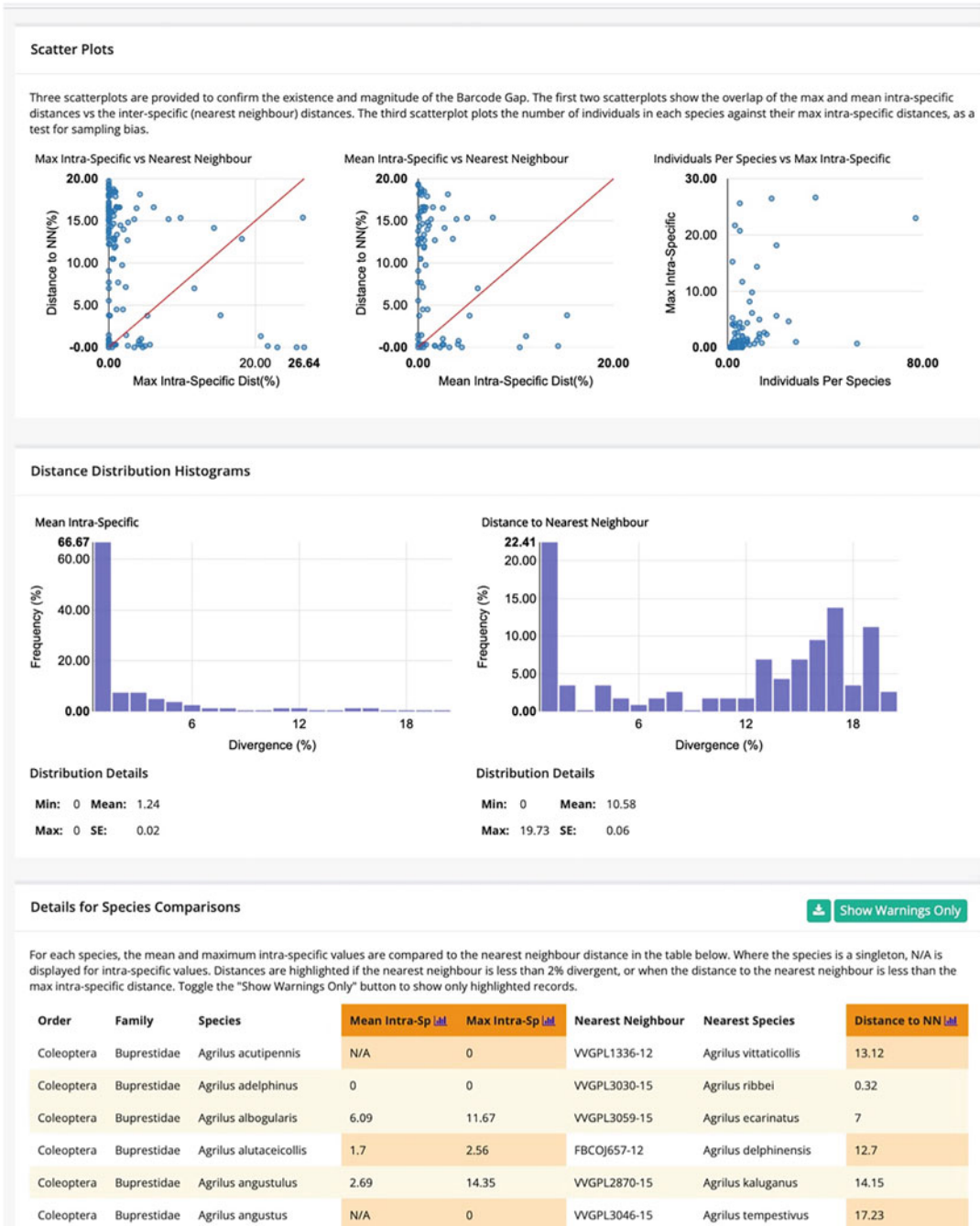


Fig. 11 Barcode gap analysis

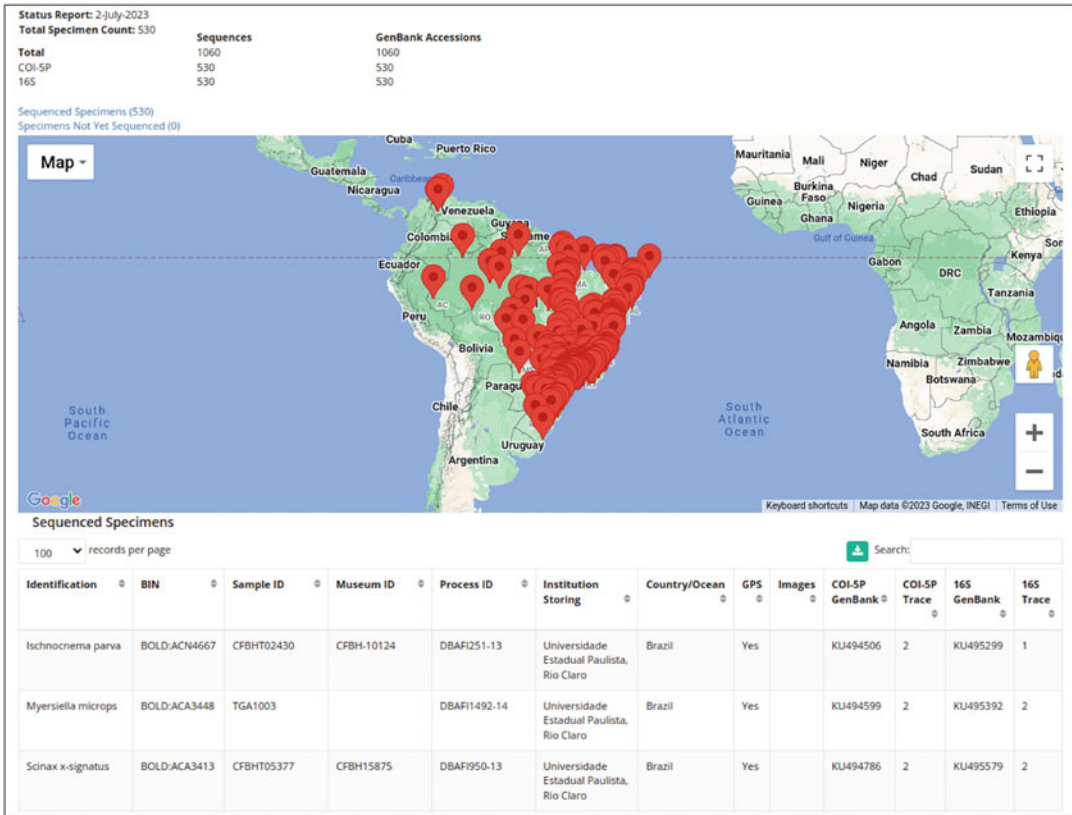


Fig. 12 Project Summary

Many scientific journals require the submission of sequence records to GenBank before publication. BOLD eases compliance by integrating a data exchange pipeline with the National Center for Biotechnology Information (NCBI), facilitating data transformation into GenBank formats, and performing the submission.

Following submission, GenBank accessions are shared with the submitters for incorporation in publications. An embargo period is established on GenBank to accommodate the time for peer review.

The final step following acceptance by a journal is to associate the citation with the dataset on BOLD and to boost visibility of both the data and the publication. This association is accomplished via the publication interface in the main menu of the Dataset Console.

By 2022, BOLD had published 794 datasets with attributed DOIs, demonstrating its dedication to making data available to the scientific community and acknowledging the contributions of researchers. This successful cycle of record compilation, validation, and publication emphasizes the BOLD’s integral role in the global DNA barcoding effort.

Identification Engine - DS-HECALIB (10 records selected)

Analysis Results

500 records per page

Search:

Processid	Sampleid	Current Order	Current ID	% Match	Overlap (bp)	Match Order	Match Family	Match Species	Match Subspecies	Match Processid	Match BIN
BBPEC186-09	09BBEHE-186	Hemiptera	Adelphocoris lineolatus	100	612	Hemiptera	Miridae	Adelphocoris lineolatus		RFM255-12	BOLD:AAA7631
BBPEC186-09	09BBEHE-186	Hemiptera	Adelphocoris lineolatus	100	612	Hemiptera	Miridae	Adelphocoris lineolatus		HMCN058-08	BOLD:AAA7631
BBPEC186-09	09BBEHE-186	Hemiptera	Adelphocoris lineolatus	100	612	Hemiptera	Miridae	Adelphocoris lineolatus		HMCN093-08	BOLD:AAA7631
BBPEC186-09	09BBEHE-186	Hemiptera	Adelphocoris lineolatus	100	612	Hemiptera	Miridae	Adelphocoris lineolatus		HMCN097-08	BOLD:AAA7631
BBPEC186-09	09BBEHE-186	Hemiptera	Adelphocoris lineolatus	100	612	Hemiptera	Miridae	Adelphocoris lineolatus		HMCN110-08	BOLD:AAA7631
BBPEC186-09	09BBEHE-186	Hemiptera	Adelphocoris lineolatus	100	612	Hemiptera	Miridae	Adelphocoris lineolatus		HMCN111-08	BOLD:AAA7631
BBPEC186-09	09BBEHE-186	Hemiptera	Adelphocoris lineolatus	100	612	Hemiptera	Miridae	Adelphocoris lineolatus		HMCN112-08	BOLD:AAA7631
BBPEC186-09	09BBEHE-186	Hemiptera	Adelphocoris lineolatus	100	612	Hemiptera	Miridae	Adelphocoris lineolatus		HMCN113-08	BOLD:AAA7631
BBPEC186-09	09BBEHE-186	Hemiptera	Adelphocoris lineolatus	100	612	Hemiptera	Miridae	Adelphocoris lineolatus		HMCN114-08	BOLD:AAA7631
BBPEC186-09	09BBEHE-186	Hemiptera	Adelphocoris lineolatus	100	612	Hemiptera	Miridae	Adelphocoris lineolatus		HMCN115-08	BOLD:AAA7631
BBPEC186-09	09BBEHE-186	Hemiptera	Adelphocoris lineolatus	100	612	Hemiptera	Miridae	Adelphocoris lineolatus		HMCN116-08	BOLD:AAA7631
BBPEC186-09	09BBEHE-186	Hemiptera	Adelphocoris lineolatus	100	612	Hemiptera	Miridae	Adelphocoris lineolatus		HMCN117-08	BOLD:AAA7631
BBPEC186-09	09BBEHE-186	Hemiptera	Adelphocoris lineolatus	100	612	Hemiptera	Miridae	Adelphocoris lineolatus		HMCN118-08	BOLD:AAA7631
BBPEC186-09	09BBEHE-186	Hemiptera	Adelphocoris lineolatus	100	612	Hemiptera	Miridae	Adelphocoris lineolatus		HMCN119-08	BOLD:AAA7631

Fig. 13 Results from the Batch ID Engine

5 Integrated Analytics

BOLD has a unique configuration that enables direct analytical operations on data in it (Table 9). This incorporation of analytical functionalities affords several advantages. Chief among them is the ability for frictionless and immediate analysis. Data can be efficiently scrutinized right after it is added or updated within the system.

Another advantage lies in the simplicity of data analysis. This approach avoids the need for data conversion steps during analysis, making the process seamless. BOLD’s analytical tools are designed to align with its data model, ensuring a fluid analytical process. BOLD also provides a functionality that promotes the sharing and distribution of analytical results. This reflects two complementary mechanisms: its capacity to store results and access them via URL and the inherent communal nature of BOLD—where users may run analyses on data they can access.

Table 9

The 11 tools on BOLD for analyzing specimen and sequence data. All can accessed from the project or dataset consoles

Tool	Functionality
Taxon ID tree	Interactive generation tool of phylogenetic hypothesis using the neighbor joining algorithm with various export options, including rooted or unrooted taxon ID trees, circular phylograms, and cladograms [35].
Distance summary	Reports divergence between sequences at the conspecific, congeneric, and confamilial levels.
Sequence composition	Frequency of DNA bases, observed with emphasis on GC-content.
Barcode gap summary	Nearest neighbor distances for each species in a list of specimens.
Diagnostic characters	Polymorphisms between sets of nucleotide or amino acid sequences to characterize consensus characters by their diagnostic potential.
Cluster sequences	Operational taxonomic unit (OTU) generation from identified or unidentified sequences using the refined single linkage algorithm (RESL) [5], providing an initial estimate of species diversity and specimen affinities.
Geo-distance correlation	Correlation investigation using two methods: Mantel Test [36] between the geographic distance matrix (in kilometers) and the genetic divergence matrix. Spread comparison with minimum spanning tree of collection sites and maximum intraspecific divergence [22].
Accumulation curve	Diversity estimation and sampling efficiency of areas of collections using standardized DNA barcodes.
Diversity measures	This tool provides insights into two key aspects of biodiversity: Alpha and Beta diversity. Alpha diversity focuses on the species richness in a specific set of data, helping to understand the diversity within a sample or at a location. Beta diversity examines the turnover or differentiation of species composition among multiple sites or samples. It assesses the degree of similarity or dissimilarity in species composition between different locations or samples [37].
Batch ID engine	Sequence matches based on similarity to entries in the BOLD database.
BIN discordance report	BIN records are shown by degree of conflict by comparing the taxonomy of selected records against others in the BIN records they are associated with.

While analysis is typically conducted on a project-by-project basis, BOLD allows the consolidation of records from multiple projects into “merged” projects, allowing more comprehensive examination. The system also supports data searches within BOLD or specific projects (Fig. 14) based on a wide range of criteria, including lists of identifiers (Process IDs, Sample IDs, and BINs), taxon names, geographic areas, tags, and collection dates. This search capability enables users to focus on specific subsets of the data to investigate localized patterns.

Record Search



For any multi-word searches, please use quotes to encapsulate each search term.

Taxonomy	Aves	
Geography	Country/Province	
Sequence Length	Marker	Minimum
Tags		
Depository		
Extra Info		
Collection Range	01-Jan-2000	01-Jan-2020

Region of Map
List of Identifiers

GPS Polygon: Clear Selection

Record Search

Keyboard shortcuts Map data ©2023 Google Terms of Use

Include public records

Fig. 14 Search functionality from the workbench. Search terms appear at the top of the panel while a geographic region can be selected at the bottom

Finally, it is important to note that size restrictions apply to certain analytical elements, like the Neighbor Joining (NJ) tree, which limits analysis to 25,000 records. However, other analytical tools, like the ID engine, scale linearly so they can handle all data records. These integrated analytical functions allow BOLD to aid the advance of DNA barcoding research.

6 Data Access

BOLD hosts public-facing search tools that allow users to query and extract data from four different databases. Two of these databases, the Public Data Portal and the BIN database, provide access to primary data on BOLD. The data in these databases is open to any user while private data is masked and only used for a statistical overview. These search tools and databases collectively allow researchers to access a wide range of sequence data and supplementary information.

6.1 *Public Data Portal*

The Public Data Portal is the central repository for all public data in BOLD. It grants users the capacity to explore and retrieve every public record in the database, making it a key resource for work on any taxonomic group. The portal returns all information associated with each record, providing users with a comprehensive view of the data. There are two ways to access the Public Data Portal: the Databases link and the Taxonomy Browser, when searching for a particular taxon. The Public Data Portal should be the first stop for users as they begin research on a specific taxon as it is the gateway to the reservoir of data on BOLD.

The Public Data Portal allows users to perform simple searches based on keywords, such as taxonomy, geography, specimen depository, project or dataset code, and specimen and sequence identifiers. More complex queries, involving the combination of two or more keywords, can also be constructed. For example, the following keywords Argentina[geo] Aves[tax] “Museo Argentino de Ciencias Naturales” (Fig. 15) would return all public records in BOLD for Argentinian birds that are held in the Argentinian Museum of Natural Sciences. Users can also search for BINs directly. The data obtained through these searches can be downloaded in varied formats (TSV, DWC, XML, and FASTA) along with related trace files. Additionally, maps can be generated from the data and downloaded at three resolutions. There is a limit of 100,000 records that can be downloaded at a time. Figure 15 provides an overview of the Data Portal and its capabilities.

PUBLIC DATA PORTAL - RECORD LIST

Argentina[geo] Aves[tax] "Museo Argentino de Ciencias Naturales" PUBLIC DATA

Unmatched Terms: Museo Argentino de Ciencias Naturales

Records 1 to 100 Page 1 2 3 4 5 6 7 8 9 10 next-> Records Per Page 100

Specimens:

Sequences:

Combined: 1

Map:

2

AWCB001-09 - Elaenia parvirostris [COI-SP-694]
Taxonomy: Chordata, Aves, Passeriformes, Tyrannidae, Elaenia
Identifiers: MACN-Or-ct 4253[sampleid], Or-ct 4253[fieldid]
Depository: Museo Argentino de Ciencias Naturales, Bernardino Rivadavia
Collected in: Argentina, Buenos Aires, Partido de Magdalena

AWCB002-09 - Elaenia parvirostris [COI-SP-694]
Taxonomy: Chordata, Aves, Passeriformes, Tyrannidae, Elaenia
Identifiers: MACN-Or-ct 4257[sampleid], Or-ct 4257[fieldid]
Depository: Museo Argentino de Ciencias Naturales, Bernardino Rivadavia
Collected in: Argentina, Buenos Aires, Partido de Magdalena

AWCB003-09 - Elaenia parvirostris [COI-SP-694]
Taxonomy: Chordata, Aves, Passeriformes, Tyrannidae, Elaenia
Identifiers: MACN-Or-ct 4260[sampleid], Or-ct 4260[fieldid]
Depository: Museo Argentino de Ciencias Naturales, Bernardino Rivadavia
Collected in: Argentina, Buenos Aires, Partido de Magdalena

AWCB004-09 - Sicalis flaveola [COI-SP-694]
Taxonomy: Chordata, Aves, Passeriformes, Thraupidae, Sicalis
Identifiers: MACN-Or-ct 4283[sampleid], Or-ct 4283[fieldid]
Depository: Museo Argentino de Ciencias Naturales, Bernardino Rivadavia
Collected in: Argentina, Entre Rios

AWCB005-09 - Zonotrichia capensis [COI-SP-694]
Taxonomy: Chordata, Aves, Passeriformes, Passerellidae, Zonotrichia
Identifiers: MACN-Or-ct 4284[sampleid], Or-ct 4284[fieldid]
Depository: Museo Argentino de Ciencias Naturales, Bernardino Rivadavia
Collected in: Argentina, Entre Rios

AWCB006-09 - Knipolegus cyanirostris [COI-SP-694]
Taxonomy: Chordata, Aves, Passeriformes, Tyrannidae, Knipolegus
Identifiers: MACN-Or-ct 4285[sampleid], Or-ct 4285[fieldid]
Depository: Museo Argentino de Ciencias Naturales, Bernardino Rivadavia
Collected in: Argentina, Entre Rios

AWCB007-09 - Zonotrichia capensis [COI-SP-694]
Taxonomy: Chordata, Aves, Passeriformes, Passerellidae, Zonotrichia
Identifiers: MACN-Or-ct 4286[sampleid], Or-ct 4286[fieldid]
Depository: Museo Argentino de Ciencias Naturales, Bernardino Rivadavia
Collected in: Argentina, Entre Rios

AWCB008-09 - Stephanophorus diadematus [COI-SP-694]
Taxonomy: Chordata, Aves, Passeriformes, Thraupidae, Stephanophorus
Identifiers: MACN-Or-ct 4287[sampleid], Or-ct 4287[fieldid]
Depository: Museo Argentino de Ciencias Naturales, Bernardino Rivadavia
Collected in: Argentina, Entre Rios

AWCB009-09 - Turdus amaurochalinus [COI-SP-694]
Taxonomy: Chordata, Aves, Passeriformes, Turdidae, Turdus
Identifiers: MACN-Or-ct 4288[sampleid], Or-ct 4288[fieldid]
Depository: Museo Argentino de Ciencias Naturales, Bernardino Rivadavia
Collected in: Argentina, Entre Rios

AWCB010-09 - Myiodynastes maculatus [COI-SP-694]
Taxonomy: Chordata, Aves, Passeriformes, Tyrannidae, Myiodynastes
Identifiers: MACN-Or-ct 4289[sampleid], Or-ct 4289[fieldid]
Depository: Museo Argentino de Ciencias Naturales, Bernardino Rivadavia
Collected in: Argentina, Entre Rios

AWCB011-09 - Myiodynastes maculatus [COI-SP-694]
Taxonomy: Chordata, Aves, Passeriformes, Tyrannidae, Myiodynastes
Identifiers: MACN-Or-ct 4290[sampleid], Or-ct 4290[fieldid]
Depository: Museo Argentino de Ciencias Naturales, Bernardino Rivadavia
Collected in: Argentina, Entre Rios

AWCB012-09 - Elaenia parvirostris [COI-SP-682]
Taxonomy: Chordata, Aves, Passeriformes, Tyrannidae, Elaenia
Identifiers: MACN-Or-ct 4291[sampleid], Or-ct 4291[fieldid]
Depository: Museo Argentino de Ciencias Naturales, Bernardino Rivadavia
Collected in: Argentina, Entre Rios

AWCB013-09 - Zonotrichia capensis [COI-SP-694]
Taxonomy: Chordata, Aves, Passeriformes, Passerellidae, Zonotrichia
Identifiers: MACN-Or-ct 4292[sampleid], Or-ct 4292[fieldid]
Depository: Museo Argentino de Ciencias Naturales, Bernardino Rivadavia
Collected in: Argentina, Entre Rios

AWCB014-09 - Thraupis sayaca [COI-SP-694]
Taxonomy: Chordata, Aves, Passeriformes, Thraupidae, Thraupis
Identifiers: MACN-Or-ct 4293[sampleid], Or-ct 4293[fieldid]
Depository: Museo Argentino de Ciencias Naturales, Bernardino Rivadavia

Results Summary

Found **3,286** published records, with **3,286** records with sequences, forming **639** BINs (clusters), with specimens from **1** country, deposited in **16** institutions.

Of these records, **2,534** have species names, and represent **595** species.

Specimen Distribution **3**

Data Summary **4**

Institutions

- Museo Argentino de Ciencias Naturales, Bernardino Rivadavia: 2669
- University of Kansas, Natural History Museum: 145
- Royal Ontario Museum: 89
- Cornell University, Museum of Vertebrates: 85
- Burke Museum of Natural History and Culture: 83
- Smithsonian Institution, National Museum of Natural History: 71
- American Museum of Natural History: 52
- University of Alaska Fairbanks: 27
- Louisiana Museum of Natural History: 23
- Mined from GenBank, NCBI: 21
- Others: 21

Fig. 15 Public Data Portal. In this example, a query was used to extract all records of birds from Argentina that are deposited in the Museo Argentino de Ciencias Naturales. By utilizing the keywords “Argentina[geo] Aves [tax]” and “Museo Argentino de Ciencias Naturales,” the results page displays several important features: (1) a list of available formats for downloading occurrences and sequences data; (2) detailed information about each individual record, including its hosting institution, location, and taxonomic classification; (3) distribution map that displays the locations of the records, and (4) data summary presenting comprehensive information about the taxonomic breakdown of the records (family, order, genera, species, and BINs). The summary also provides information about the institutions and countries where the records were collected. These features collectively offer users an overview of the data in the Public Data Portal, facilitating access and exploration

6.2 Data Packages

The rapid growth of the BOLD database and the rising adoption of DNA barcoding have led to a strong demand for immediate access to large datasets. For many users, accessing data through portals or Application Programming Interfaces (APIs) is inefficient for large-scale requirements. Bulk downloading, particularly within institutions with multiple users, is more effective, ensuring consistent access to datasets.

Additionally, methods like metabarcoding and environmental DNA (eDNA) analyses require copies of comprehensive datasets for accurate data analysis. These analyses need access to extensive and reliable datasets, whether run on local systems or cloud-based platforms.

To address these needs, the Data Packages module was introduced. This module serves packaged datasets in as single compressed files with supporting metadata, increasing the discoverability and accessibility of data for users. A key feature of the Data Packages module is its adherence to the Frictionless Data standard [38] created by the Open Knowledge Foundation. This standard offers a structured method for data description, packaging, and validation, using a container format to combine structured data with its associated metadata in JSON format. This format was chosen to simplify data sharing, validation, and interoperability.

The approach to package assembly is systematic. All public records are compiled into snapshots on a quarterly basis, with each package summarized on a dedicated page (Fig. 16) and

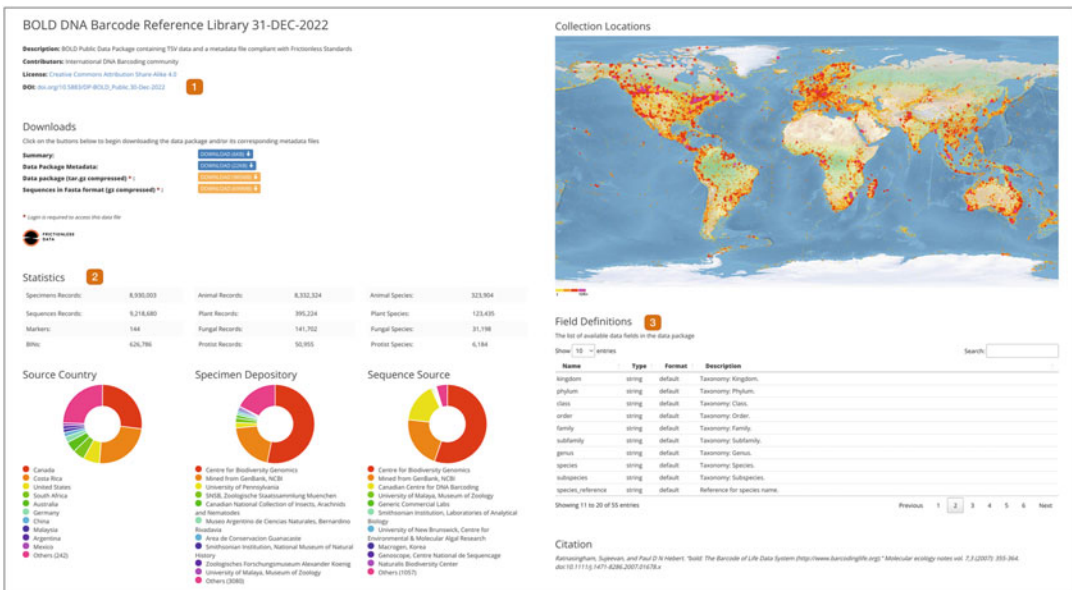


Fig. 16 The data package for the BOLD library for December 31, 2022, shows the following key features: (1) Digital Object Identifier; (2) Summary statistics, encompassing the number of records, sequences, markers, and BINs; and (3) Data dictionary

referenced by a DOI. For users needing the most up-to-date data, BOLD produces weekly snapshots. Project and program-focused data packages are also supported. One significant package in this category is from the iBOL's BARCODE 500 K program.

6.3 Taxonomy Browser

The BOLD Taxonomy Browser (Fig. 17) is indispensable for researchers engaged in constructing DNA barcode libraries. It provides a summary of barcode coverage for specific taxa and facilitates exploration of the taxonomic hierarchy from the kingdom down to the species. Accommodating entries for animals, plants, fungi, and protists, the browser is a valuable tool for diverse research areas.

Because the Taxonomy Browser is integrated with the BOLD and Barcode Index Number (BIN) databases, users can access associated specimen and sequence records for any taxon. In this way, the browser acts as a dynamic reflection of BOLD's current state, encapsulating the latest updates in taxonomy.

The browser allows access to varied data elements from summary statistics like the number of records and species for a particular taxonomic group to more detailed information such as associated BINs and individual records. Additionally, it provides information about the distribution of collection sites and the locations of specimens, enhancing the contextual understanding of the data.

The Taxonomy Browser also displays representative images for each taxon, enabling the visual verification of current assignments. Its most significant role lies in reducing the duplication of effort in library construction, enabling researchers to monitor progress on taxa of interest.

6.4 Application Programming Interface (APIs)

The BOLD API streamlines access to the DNA barcode sequences, specimen data, and associated metadata. Via the API endpoints (Table 10), users can programmatically retrieve data subsets based on various criteria, including taxonomy, geography, or specific gene markers, aiding more comprehensive, automated analyses. Supporting various data formats (i.e., XML and JSON), it ensures broad compatibility and easy integration with other bioinformatics tools/platforms. Notably, the API has been utilized by ROpenSci [39], an initiative that promotes open and reproducible research using shared data and reusable software. Their integration of the BOLD API facilitates easier access and analysis of barcoding data within the R environment, a widely used programming language and software environment for statistical computing. The API also offers an analytical endpoint that enables near real-time DNA barcode-based species identifications, bridging the gap between raw sequence data and practical applications. Developers can leverage this API to build custom applications, tools, or plugins, to enhance the utility of BOLD data in diverse contexts.



Taxon Description (Wikipedia)

The **emerald ash borer** (*Agrilus planipennis*) is a green beetle native to Asia and Eastern Russia. Outside its native region, the emerald ash borer (also referred to as EAB) is an invasive species, and emerald ash borer infestation is highly destructive to ash trees in its introduced range. The Emerald Ash Borer was first discovered in America in June 2002 in Michigan. It is believed to have been brought to America unintentionally in ash wood which was used to stabilize crates during shipping, [full article at Wikipedia](#)

© CC BY ▲ CBG Photography Group 📅 2015

1 Statistics

Specimen Records:	92
Specimens with Sequences:	71
Specimens with Barcodes:	57
Subspecies:	0
Subspecies with Barcodes:	0
Public Records:	61
Public Subspecies:	0
Public BINs:	1

[SUBSPECIES LIST](#)

[PUBLIC DATA](#)

2 Image of *Agrilus planipennis*

Specimen Depositories



- Mined from GenBank, NCBI [39]
- Canadian National Collection of Insects, Arachnids and Nem... [12]
- Centre for Biodiversity Genomics [6]
- Canadian Forest Service, Great Lakes Forestry Centre [6]
- Canadian Forest Service, Laurentian Forestry Centre [3]
- Canadian Food Inspection Agency [2]
- Sukachev Institute of Forest Siberian Branch of Russian
- Aca... [1]
- Research Collection of Tim Myles [1]
- 0 Others [0]

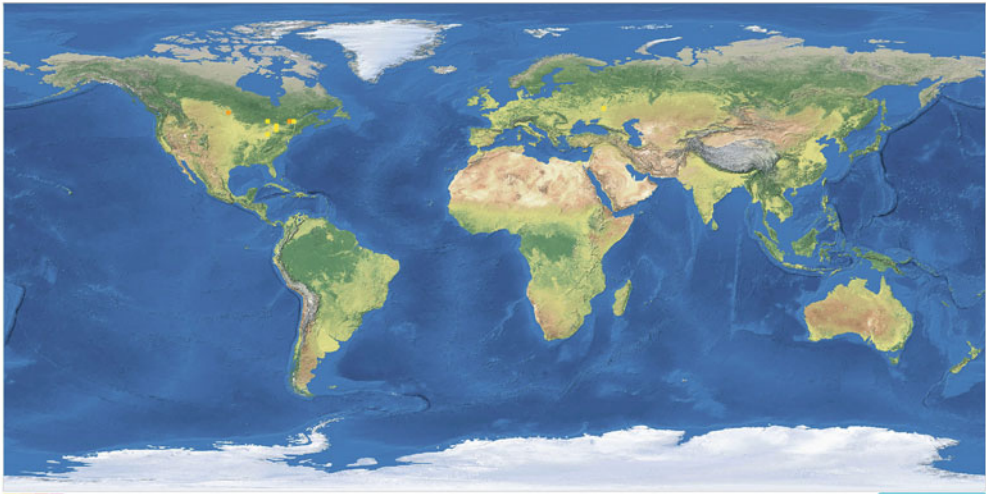
3

Sequencing Labs



- Mined from GenBank, NCBI [38]
- Biodiversity Institute of Ontario [25]
- Centre for Biodiversity Genomics [2]
- Canadian Forest Service, Great Lakes Forestry Centre [1]

4 Collection Sites



[DOWNLOAD MAP](#)

Countries

Collected from 3 countries. [Show all countries](#)

Canada: 30 Russia: 1

United States: 2

Imagery

Images representing specimens of *Agrilus planipennis*

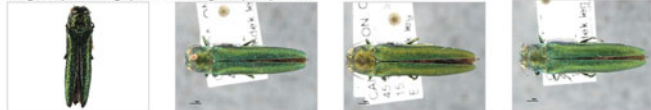


Fig. 17 The Taxonomy Browser page for *Agrilus planipennis* (Emerald Ash Borer Beetle) displays: (1) Coverage statistics; (2) Specimen repositories; (3) Laboratories responsible for sequence generation; and (4) Collection sites

Table 10
List of the nine endpoints/functions of the BOLD API

API endpoints	Description
Summary statistics	Retrieves summary information provided by BOLD public searches. e.g., http://www.boldsystems.org/index.php/API_Public/stats?taxon=Eacles&geo=Mexico&format=tsv
Specimen data	Retrieves data records matching any combination of the following 7 parameters: Taxon, IDs, BIN, container, institution, researcher, and geolocation. e.g., http://www.boldsystems.org/index.php/API_Public/specimen?taxon=Aves&geo=Costa%20Rica&format=tsv
Sequence data	Retrieves sequences matching any combination of the following 9 parameters: Taxon, IDs, BIN, container, institution, researcher, geolocation, gene marker, and format. e.g., http://www.boldsystems.org/index.php/API_Public/sequence?taxon=Chordata&geo=Florida&institutions=Smithsonian%20Institution
Specimen and sequence data	Retrieves specimen data and sequence records matching any combination of the following 9 parameters: Taxon, IDs, BIN, container, institution, researcher, geolocation, gene marker, and format. e.g., http://www.boldsystems.org/index.php/API_Public/combined?taxon=Mammalia&geo=Canada&format=tsv
Sanger trace files	Retrieves trace files matching any combination of the following 9 parameters: Taxon, IDs, BIN, container, institutions, researcher, geolocation, gene marker, and format. e.g., http://www.boldsystems.org/index.php/API_Public/trace?taxon=Bombus&institutions=York%20University
Taxonomy ID service	Retrieves taxonomic information generated by BOLD taxonomy ID. http://www.boldsystems.org/index.php/API_Tax/TaxonData?taxId=88899&dataTypes=basic,stats
Taxon name service	Retrieves taxonomy information by taxon name. e.g., http://www.boldsystems.org/index.php/API_Tax/TaxonSearch?taxName=Diplura
Rest API for ID engine	The BOLD ID engine web service can query the COI ID engine via URL. An XML file with the top public matches (up to 100) is retrieved by querying a COI sequence. e.g., http://www.boldsystems.org/index.php/Ids_xml?db=COX1_SPECIES_PUBLIC&sequence=TATAAATAATATAAGATTTTGACTTCTTCCACCTTCTTTAACTCTTC TCCTATCCAGAGGAATAGTTGAAAGAGGTGTTGGCACAGGATGAAC TGTTTATCCTCCTTTAGCTGCTGGAATCGCCATGCAGGCGCTTCT GTGGACTTAGGAATTTTTTCTCTTCATATAGCGGGAGCTTCTTCTAT TTTAGGGGCGGTAAATTTTATTACTACT

7 Conclusion

The BOLD platform has been a pivotal component of biodiversity research since its inception. As an essential part of the global initiative for species identification using DNA barcoding, BOLD has consistently played a vital role in consolidating and managing molecular and specimen-based data. Providing an interface for collecting, storing, analyzing, and publishing DNA barcode records, effectively bridging the gap between genetic data and taxonomic knowledge. Over nearly two decades of operation, the BOLD system has expanded in sophistication, scope, and utility, becoming an integral component of numerous research initiatives and programs.

Acknowledgments

We are profoundly grateful to the funders who have supported BOLD since its inception in 2004. They include the Canada Foundation for Innovation, Genome Canada, Gordon and Betty Moore Foundation, Natural Sciences and Engineering Research Council of Canada (NSERC), Ontario Ministry of Research and Innovation, Ontario Genomics, Canada First Research Excellence Fund, and the New Frontiers in Research Fund.

References

1. Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Sherry ST, Yankie L, Karsch-Mizrachi I (2023) GenBank 2023 update. *Nucleic Acids Res* 51: D141–D144
2. Edwards JL, Lane MA, Nielsen ES (2000) Interoperability of biodiversity databases: biodiversity information on every desktop. *Science* 289:2312–2314
3. Hobern D, Barik SK, Christidis L et al (2021) Towards a global list of accepted species VI: the catalogue of life checklist. *Org Divers Evol* 21: 677–690
4. Ratnasingham S, Hebert PDN (2007) The barcode of life data system. *Mol Ecol Notes* 7: 355–364
5. Ratnasingham S, Hebert PDN (2013) A DNA-based registry for all animal species: the barcode index number (BIN) system. *PLoS One* 8:e66213. <https://doi.org/10.1371/journal.pone.0066213>
6. Sonet G, Jordaens K, Braet Y, Bourguignon L, Dupont E, Backeljau T, de Meyer M, Desmyter S (2013) Utility of GenBank and the barcode of life data systems (BOLD) for the identification of forensically important Diptera from Belgium and France. *Zookeys* 365:307–328
7. DeSalle R, Goldstein P (2019) Review and interpretation of trends in DNA barcoding. *Front Ecol Evol* 7. <https://doi.org/10.3389/fevo.2019.00302>
8. Hebert PDN, Gregory TR (2005) The promise of DNA barcoding for taxonomy. *Syst Biol* 54:852–859
9. CBOL Plant Working Group, Hollingsworth PM, Forrest LL et al (2009) A DNA barcode for land plants. *Proc Natl Acad Sci U S A* 106: 12794–12797
10. Seifert KA (2009) Progress towards DNA barcoding of fungi. *Mol Ecol Resour* 9:83–89
11. Min XJ, Hickey DA (2007) Barcoding: assessing the effect of varying sequence length on DNA barcoding of fungi. *Mol Ecol Notes* 7: 365–373
12. Ewing B, Green P (n.d.) Base-calling of automated sequencer traces using Phred. II Error Probabilities

13. Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* 14:755–763
14. Somervuo P, Koskela S, Pennanen J, Henrik Nilsson R, Ovaskainen O (2016) Unbiased probabilistic taxonomic classification for DNA barcoding. *Bioinformatics* 32:2920–2927
15. Austerlitz F, David O, Schaeffer B, Bleakley K, Olteanu M, Leblois R, Veuille M, Laredo C (2009) DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. *BMC Bioinformatics* 10(Suppl 14): 14. <https://doi.org/10.1186/1471-2105-10-S14-S10>
16. Boratyn GM, Camacho C, Cooper PS et al (2013) BLAST: a more efficient report with usability improvements. *Nucleic Acids Res* 41: W29. <https://doi.org/10.1093/nar/gkt282>
17. Eddy SR (1998) Profile hidden Markov models. *Bioinformatic Rev* 14:755–763
18. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10:421
19. Page RDM (2016) DNA barcoding and taxonomy: dark taxa and dark texts. *Philos Trans R Soc Lond B Biol Sci* 371:20150334. <https://doi.org/10.1098/rstb.2015.0334>
20. Hausmann A, Godfray HCJ, Huemer P, Mutanen M, Rougerie R, van Nieukerken EJ, Ratnasingham S, Hebert PDN (2013) Genetic patterns in European geometrid moths revealed by the barcode index number (BIN) system. *PLoS One* 8:e84518
21. Pentinsaari M, Hebert PDN, Mutanen M (2014) Barcoding beetles: a regional survey of 1872 species reveals high identification success and unusually deep interspecific divergences. *PLoS One* 9:e108651
22. Blagoev GA, deWaard JR, Ratnasingham S, deWaard SL, Lu L, Robertson J, Telfer AC, Hebert PDN (2016) Untangling taxonomy: DNA barcode reference library for Canadian spiders. *Mol Ecol Resour* 16:325–341
23. Layton KKS, Martel AL, Hebert PDN (2014) Patterns of DNA barcode variation in Canadian marine Molluscs. *PLoS One* 9:e95003
24. Sharkey MJ, Janzen DH, Hallwachs W et al (2021) Minimalist revision and description of 403 new species in 11 subfamilies of costa Rican braconid parasitoid wasps, including host records for 219 species. *Zookeys* 1013: 1–665
25. Fernandez-Triana JL (2022) Turbo taxonomy approaches: lessons from the past and recommendations for the future based on the experience with Braconidae (hymenoptera) parasitoid wasps. *Zookeys* 1087:199–220
26. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM (2010) The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* 38:1767–1771
27. Salvi D, Berrilli E, D’Alessandro P, Biondi M (2020) Sharpening the DNA barcoding tool through a posteriori taxonomic validation: the case of Longitarsus flea beetles (Coleoptera: Chrysomelidae). *PLoS One* 15:e0233573
28. Riedel A, Sagata K, Suhardjono YR, Tänzler R, Balke M (2013) Integrative taxonomy on the fast track - towards more sustainability in biodiversity research. *Front Zool* 10:15
29. La SJ, Wheeler Q, Jackway P, Winterton S, Hobern D, Lovell D (2009) Zootaxa, accelerating taxonomic discovery through automated character extraction. *Zootaxa* 2217:43–55
30. Martineau M, Conte D, Raveaux R, Arnault I, Munier D, Venturini G (2017) A survey on image-based insect classification. *Pattern Recogn* 65:273–284
31. Howe K, Bateman A, Durbin R (2002) QuickTree: building huge neighbour-joining trees of protein sequences. *Bioinformatics* 18:1546–1547
32. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* <https://doi.org/10.1093/oxfordjournals.molbev.a040454>
33. Hebert PDN, Ratnasingham S, de Waard JR (2003) Barcoding animal life: cytochrome *c* oxidase subunit 1 divergences among closely related species. *Proc R Soc Lond B Biol Sci* 270:S96. <https://doi.org/10.1098/rsbl.2003.0025>
34. van Velzen R, Weitschek E, Felici G, Bakker FT (2012) DNA barcoding of recently diverged species: relative performance of matching methods. *PLoS One* 7:e30490
35. Paradis E, Claude J, Strimmer K (2004) APE: analyses of Phylogenetics and evolution in R language. *Bioinformatics* 20:289–290
36. Mantel N (1967) The detection of disease clustering and a generalized regression approach. *Cancer Res* 27(2):209–220
37. Dixon P (2003) VEGAN, a package of R functions for community ecology. *J Veg Sci* 14: 927–930
38. Fowler D, Barratt J, Walsh P (2018) Frictionless data: making research data quality visible. *Int J Digit Curation* 12:274–285
39. Boettiger C, Chamberlain S, Hart E, Ram K (2015) Building software, building community: lessons from the rOpenSci project. *J Open Res Softw* 3:8