BOLDSYSTEMS

Barcode Of Life Data Systems Handbook

September 2012

www.boldsystems.org

Version 3

Table of Contents

This handbook provides details on functionality, data structures and best practices for BOLD version 3. It explains how to use this system to collect, manage and publish Barcode and ancillary data. It also provides details on the integrated analytical tools. At any time while using BOLD, you can access the online documentation by clicking on the "Get Help" link in the footer of every page, or by selecting "Documentation" from the page header.

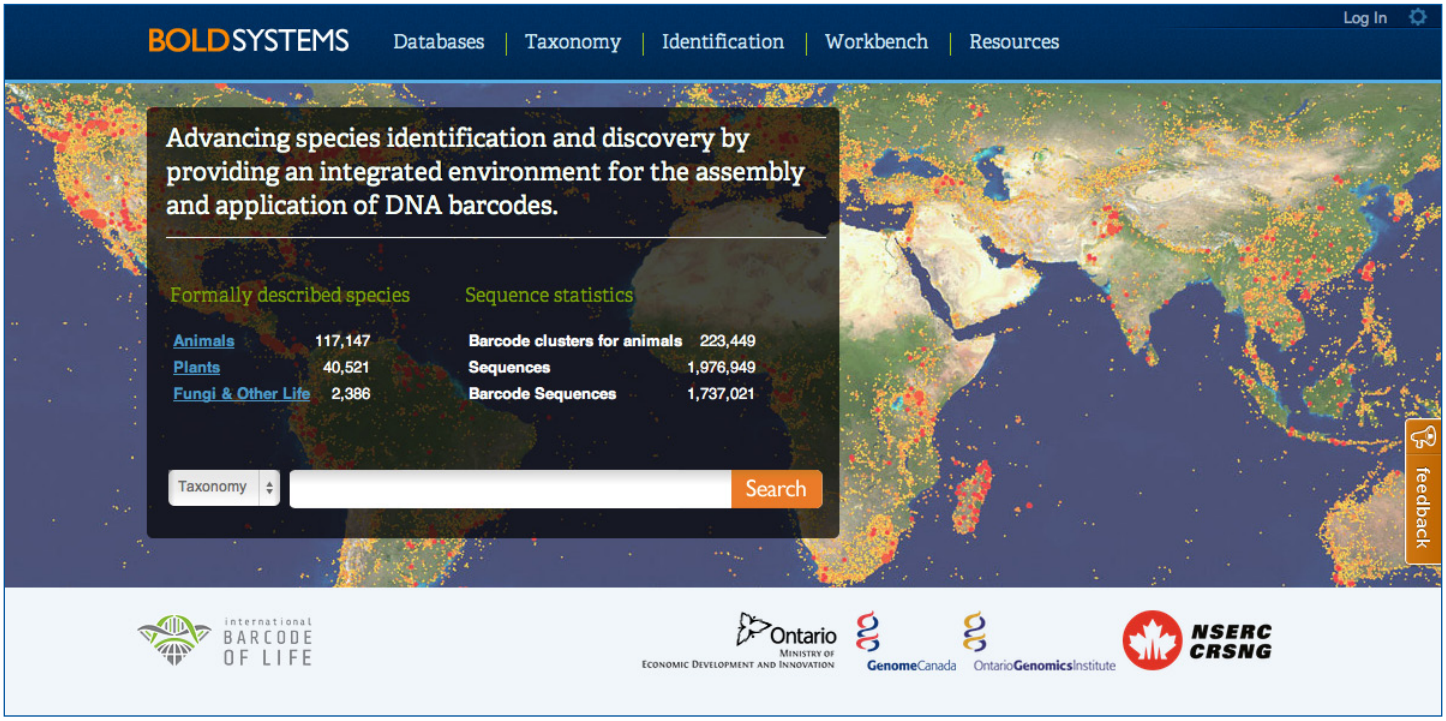
For assistance with any feature of BOLD, please email the BOLD Support Team: support@boldsystems.org

Introduction

The Barcode of Life Data Systems (BOLD), established in 2005, is a web platform that provides an integrated environment for the assembly and use of DNA barcode data. It delivers an online database for the collection and management of specimen, distributional, and molecular data as well as analytical tools to support their validation. Over the past few years, BOLD has grown to become a powerful online workbench and the central informatics hub of the DNA barcoding community.

BOLD is freely available to any researcher with interests in DNA Barcoding. By providing specialized services, it aids in the publication of records that meet the standards needed to gain BARCODE designation in the international nucleotide sequence databases. Due to its web-based delivery and flexible data security model, it is also well positioned to support projects that involve broad research alliances.

The release of version 3.0 of BOLD in April 2012 represents an evolutionary update to the system. Significant revisions have been made to support an increasing diversity of workflows and an increasing volume of data. A major advance is the activation of Barcode Index Numbers (BINs), an interim taxonomic system for animals, and an annotation framework that supports rapid community-based validation of barcode data.



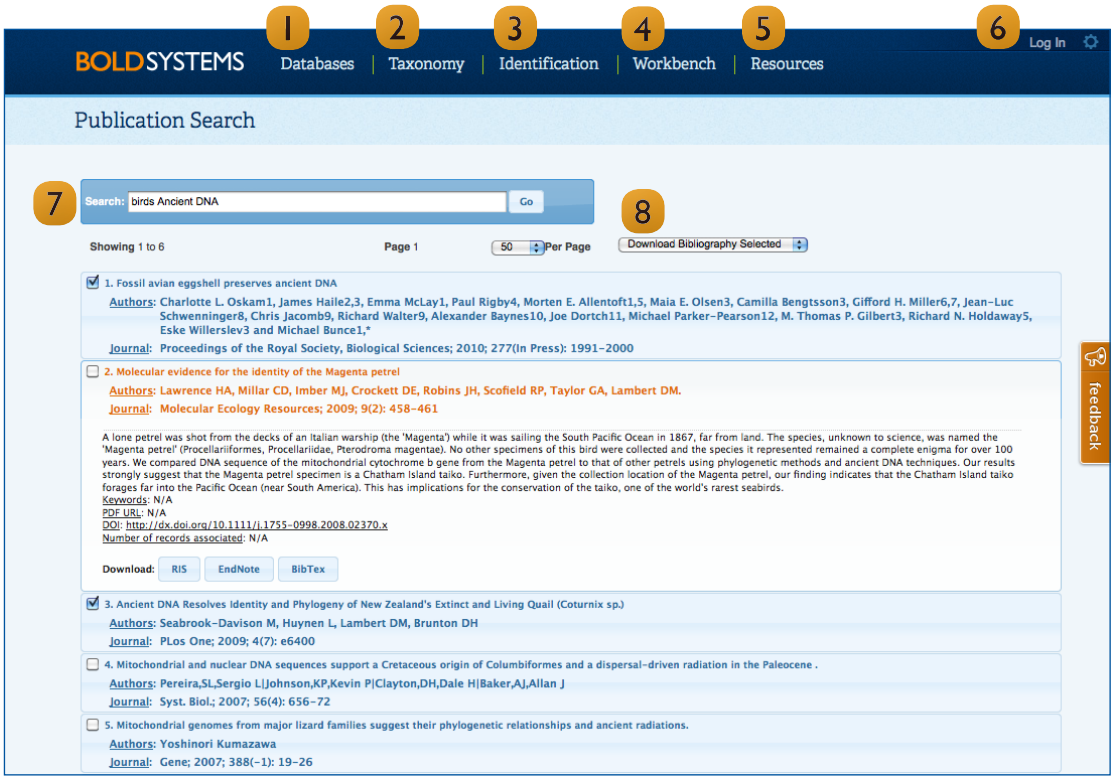
www.boldsystems.org

Getting Started on BOLD

The interface in the newest version of BOLD improves access to commonly used features and databases.

From the top menu, users can now access the public databases, the taxonomy browser, the identification engines, the user workbench and general resources. In addition, the footer provides access to details on the BOLD organization, community and partner sites.

Please see the diagram and table below for a description of the navigation features.



A publication search page illustrating navigation features

Description of features numbered in the image above

1	Databases	The Databases link provides access to the following resources that are accessible without signing in. **Public Data Portal**: A database of all of the public sequences on BOLD, including those in the early data release phase of the iBOL project. This database can be used to access and download the associated specimen data and sequences. Search by taxonomic, geographic, institution or identifier keywords. **BIN Database**: Barcode Index Numbers (BINs) are an interim taxonomic system for animals. Barcodes are clustered algorithmically, generating a web page for each cluster which is deposited in this database. Clusters show high concordance with species, which provides a fast-track for documenting diversity where taxonomic resources are limited. Search BINs by taxonomic, geographic, institution or identifier keywords. **Primer Database**: A searchable database of barcode primers, which includes primer statistics. Search by primer code, submitter name or reference keywords. **Publication Database**: A searchable, community maintained database of barcode papers linked to published datasets. Search by title, abstract or author keywords.
2	Taxonomy	The taxonomy link provides access to the taxonomy browser, a public resource which contains a page that displays the images, distribution map and other details for each taxon on BOLD. Each image uploaded to BOLD has a license applied to it. Images may be used from the taxonomy browser if the image licensed as Creative Commons or No Rights Reserved, following the rules of the license.
3	Identification	The identification link provides access to the animal, plant and fungal identification engines based on the COI, matK, rbcL, and ITS genes. This resource is available without need for a user account.
4	Workbench	The workbench link provides access to the BOLD data analysis and management workbench. After logging in, the initial page is the User Console.
5	Resources	The resources link provides access to Site Documentation (including an online version of this handbook), Barcoding resources and access to data releases from Barcoding initiatives like iBOL.
6	Log in/out	In the top right corner of any page, users can log in or log out. When logged in, the user's name appears in this section.

Table continued on next page...

Table continued from previous page...

7	Search Bars	In the public databases listed on the previous page, a search bar is present at the top of the page. Users can enter any combination of keywords to search within these databases. For example, searching "Lepidoptera Canada" in the Public Data Portal will return all of the Lepidoptera records collected in Canada. For further details and examples an using the search functionality, see the search help section that is available by clicking on the help button to the right of the search bar in each database.
8	Downloads	In each database, there is also an option to download the public data available based on your search terms. This includes barcode sequences, traces, specimen data, bibliographies, and primer sequences. Taxon ID Trees can also be downloaded from within BIN pages, and Distribution Maps are available for BINS and Public Data searches.

Registering on BOLD

To register on BOLD, click on the Workbench link and then select "Create Account" to fill in the registration form. After you have submitted your registration, a welcome e-mail will be sent to you with the information you need to log in and begin using your BOLD account. Upon signing in, you will see the User Console. After you start to contribute data and join your collaborators projects, the User Console will give you progress statistics and activity feeds (see page 22 for a depiction of the User Console).

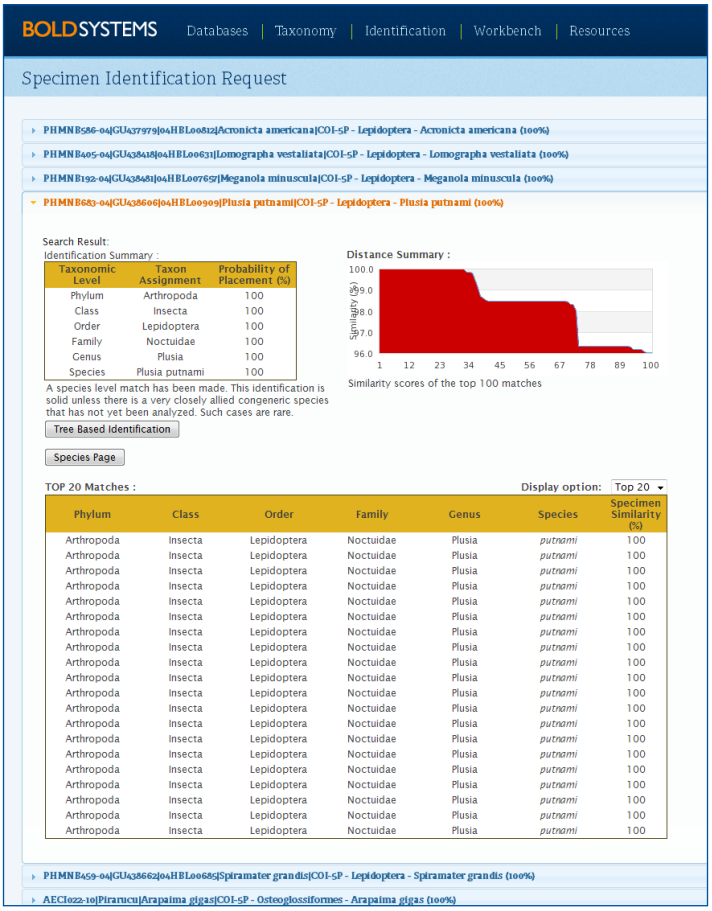
The User Console provides access for searching the data on BOLD by project code, title or tag, or by using the record search (see page 25 for more details on the record search within the Workbench). After gaining access to projects (either by creating them or being added to colleague's projects), users can upload specimen data, images, and sequence data.

Why register for a User Account?

Getting an account on BOLD expands the list of options available to a user beyond access to public data and use of the identification engine. Users can annotate published data, to help curate and cleanse the barcode library. Moreover, users will be able to submit data to BOLD and gain access to other in-progress, private, projects with the permission of the data owners. Once records are on BOLD, a large set of analytical tools are available for validation and generation of reports for publication. The system will automatically check sequences submitted for Barcode Compliance and provide reports on records lacking pertinent details. BOLD also provides a simple mechanism to release sequences publicly, as well as into partner nucleotide databases such as GenBank.

Identification

The library of sequences collected in BOLD is available for facilitating identification of unknown sequences. The ID engines use all sequences uploaded to BOLD from private, as well as public projects, to locate the closest match. To protect BOLD users, no sequence information from private records is exposed.



Identification Engine Results Page for batch identification

Animal Identification (COI)

The BOLD Identification System for animals accepts sequences from the 5' region of the mitochondrial gene COI and returns a species-level identification when possible. Further validation with independent genetic markers is desirable in some forensic applications. BOLD uses the BLAST algorithm to identify single base indels before aligning the protein translation through profile to a Hidden Markov Model of the COI protein. There are four databases within BOLD for use in identification of COI sequences:

1. All Barcode Records Database includes:
Every COI barcode record on BOLD with a minimum sequence length of 500bp (Warning: This is an un-validated database and includes records without species level identification). This includes many species represented by only one or two specimens, as well as all species with interim taxonomy. This search only returns a list of the nearest matches and does not provide a probability of placement to a taxon.

2. Species Level Barcode Database includes:
Every COI barcode record with a species level identification and a minimum sequence length of 500bp (Warning: This is an un-validated dataset). This includes many species represented by only one or two specimens, and all species with interim taxonomy.

3. Public Record Barcode Database includes:
All published COI records from BOLD and GenBank with a minimum sequence length of 500bp. This library is a collection of records from the published projects section of BOLD.

4. Full Length Barcode Database includes:
A subset of the Species library with a minimum sequence length of 640bp and containing both public and private records. This library is intended for short sequence identification as it provides maximum overlap with short reads from the barcode region of COI.

Fungal (ITS) and Plant (rbcL & matK) Identification

In the BOLD Identification System, ITS is the default identification tool for fungal barcodes and rbcL and matK are the defaults for plant barcodes. Both return a species-level identification when possible. Further validation with independent genetic markers will be desirable in some forensic applications. The BLAST algorithm is employed in place of BOLD's internal identification engine for these sequences. There are relatively few fungal and plant records on BOLD so most queries will likely not return a successful match. This will improve as sampling efforts continue in these kingdoms. These databases include many species represented by only one or two specimens, as well as all species with interim taxonomy. Both searches only return a list of the nearest matches and do not provide a probability of placement to a taxon.

Fungal Database includes:
Every ITS barcode record on BOLD with a minimum sequence length of 100bp (Warning: This is an un-validated database that includes records without species level identification).

Plant Database includes:
Every rbcL and matK barcode record on BOLD with a minimum sequence length of 500bp (Warning: This is an un-validated database that includes records without species level identification).

Batch Identifications

The newest version of BOLD provides the ability to submit a batch of query sequences for identification for up to 100 samples at a time.

Taxonomy Browser

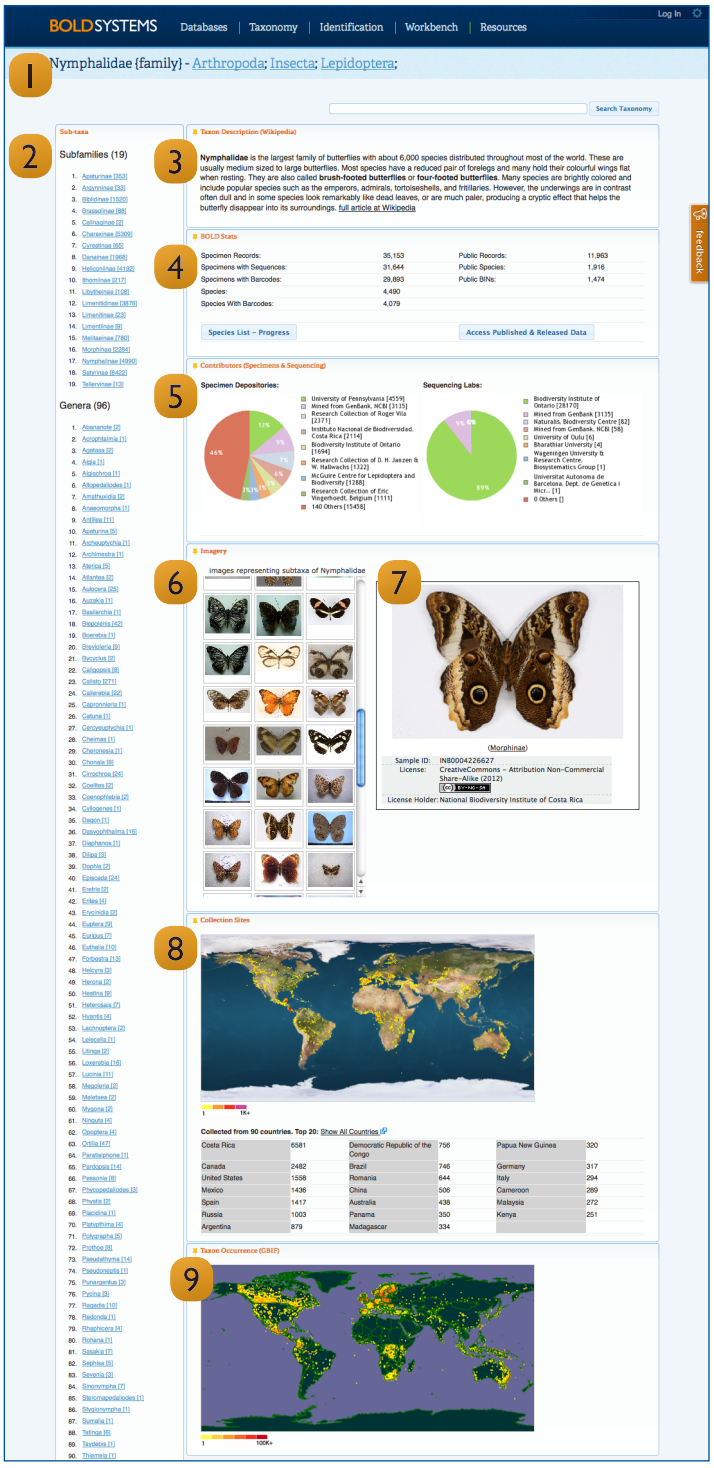
The Taxonomy Browser is a synthetic database that allows users to examine the progress of DNA barcoding by browsing through the different levels of the taxonomic hierarchy available on BOLD.

Within the Taxonomy Browser users are able to select between the animal, plant, fungus, and protist kingdoms and navigate from phylum to species level. Statistics on the progress of DNA barcoding at each taxon are generated from both public and private data while protecting private user-owned data.

To look up a specific taxon directly, use the search function by entering a taxonomic name into the search bar at the top of the Taxonomy Browser or on the BOLD main page.

1. Lineage	Displays the taxon name and the higher taxonomic levels.
2. Sub-Taxonomy	Links to all sub-taxa with number of specimen records for each.
3. Taxon Description	Displays the description of this taxon from the Wikipedia website.
4. Statistics	These statistics are compiled by BOLD for this taxon. A species progress list can be download for each rank that has sub-taxa. The published and released sequences for this taxon in the Public Data Portal can be accessed from this section.
5. Contributors	Graphs depicting the institutions that provided the samples and sequencing for the samples.
6. Imagery	A random selection of the images available for the subtaxa of this taxon. Mousing over an image selects it for higher-resolution display to the right.
7. Image Details	The taxonomic identifier, the sample identifier and image licensing details are displayed beneath the image that is selected.
8. Collection Sites	A map of the collection sites for records in BOLD, including a list of the top countries
9. Taxon Occurrence	A map of the occurrence data for this taxon worldwide, streaming from the GBIF website.

Information available at each taxonomic level in the BOLD taxonomy browser.



BOLD taxonomy browser page for the Nymphalid butterflies

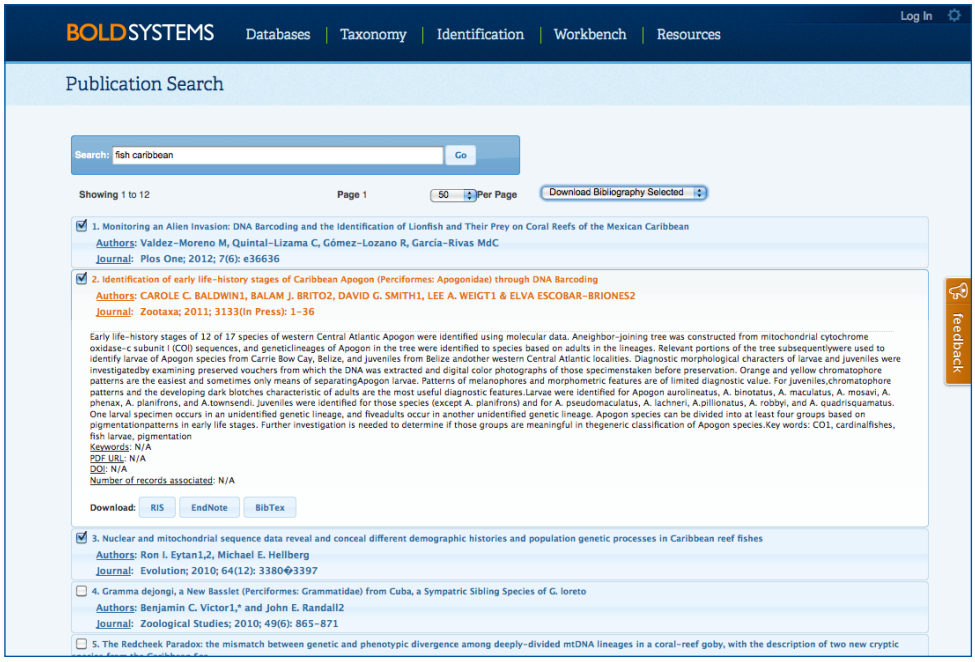
Databases: Publication

The Publication Database is accessible from anywhere in the application. This database indexes title, abstract, year, and authors, allowing for broad searches.

Selecting a publication from the database will provide further details, including a link to the article on the journal's site and access to the records if they are in BOLD.

A citation or set of citations can be downloaded from BOLD using the button to the right of the search bar.

Bibliographies can be submitted to this database by following the directions on page 21.



Publication Database

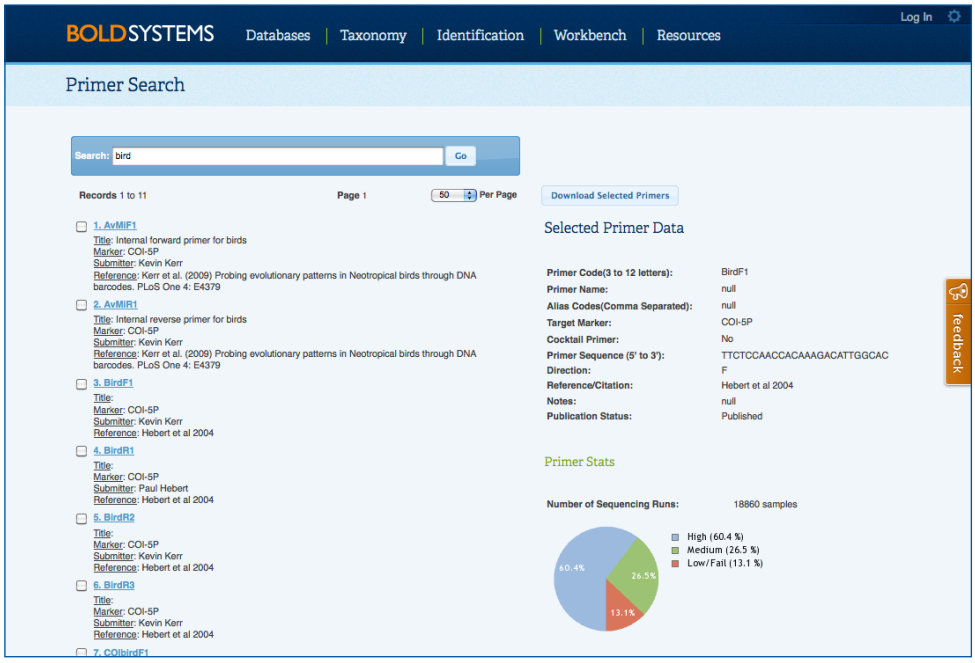
Databases: Primer

The Primer Database is accessible anywhere in the application. Using the search bar, users can search for terms that appear in the primer code, submitter or reference fields.

Selecting a primer will provide details on the primer, including primer performance statistics derived from data submitted to BOLD.

A primer or set of selected primers can be downloaded in FASTA format using the button to the right of the search bar.

New primers must be registered with BOLD before trace files generated using them are submitted. For details on registering a new primer, see page 21.



Primer Database

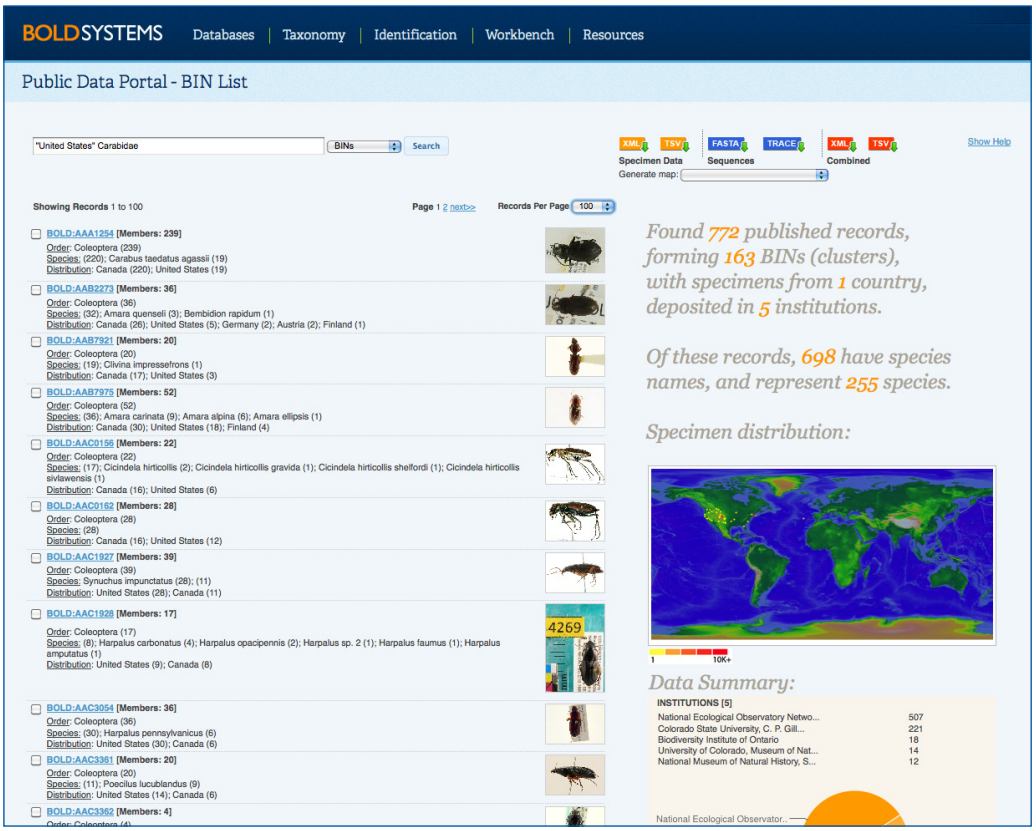
Databases: Public Data Portal

The BOLD Public Data Portal is a publicly accessible database of all of the public sequences on BOLD, including those in the early data release phase of the iBOL project. This database can be used to access and download the associated specimen data and sequences.

Searching the Data Portal

By accessing the Public Data Portal search from the Databases link in the header of the BOLD home page, users can search the public database using taxonomy, geography (country or state/province), and institution keywords, or by using Sample ID or BOLD Process ID to find an individual record.

Users can enter any combination of keywords into the search bar. For example, searching "Lepidoptera Canada" will return all of the Lepidoptera records collected in Canada. Searching "Lepidoptera Canada -Ontario" will return the same results with the specimens collected in Ontario removed. For further details and examples for using the search functionality, see the search help section that is available by clicking on the help button to the right of the search bar.



Public Data Portal - BIN List

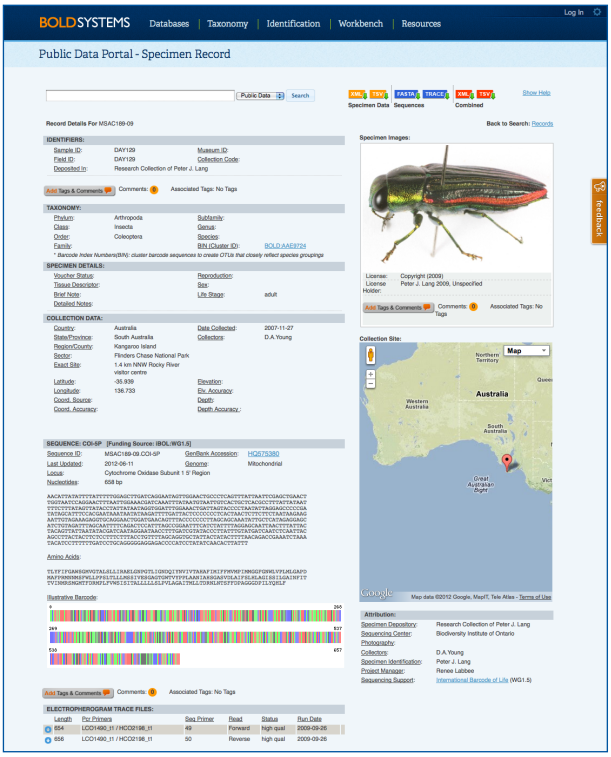
Results

The search results will display a list of BINs or records based on the options selected. For more information on BIN pages, please see the next page. Clicking on the "Record List" will convert the result list to public records matched only and clicking on "BIN List" will convert the list to all BINs available.

Specimen Record

The record page gives information on the specimen identifier, taxonomy, specimen details, collection data (including collection site), sequence information, specimen image details, and attribution details. The image to the right shows the details page for a particular record.

A record page will reference a BIN when one is available and provides links to GenBank records.



Data Portal - Specimen Record

Databases: Barcode Index Numbers (BINs)

The Barcode Index Number System is an online framework that clusters barcode sequences algorithmically, generating a web page for each cluster. Since clusters show high concordance with species, this system can be used to verify species identifications as well as document diversity when taxonomic information is lacking.

This system consists of three parts:

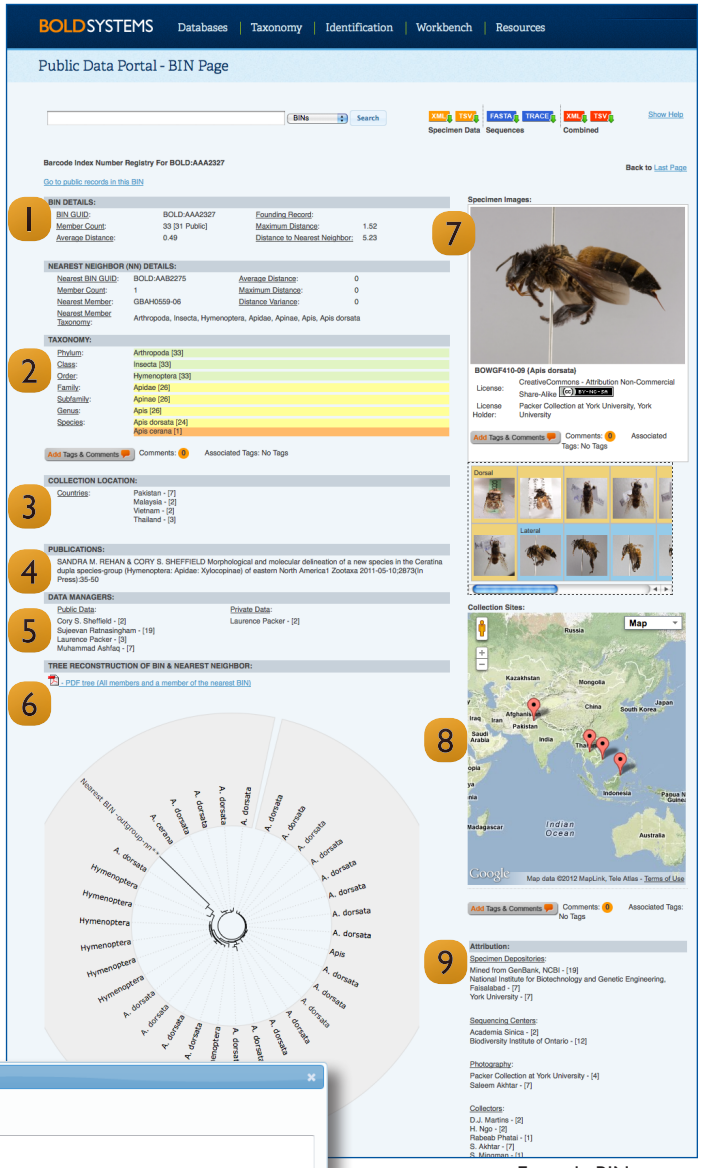
- A clustering algorithm employing graph theoretic methods to generate operational taxonomic units (OTUs) and putative species from sequence data without prior taxonomic information.
- A curated registry of barcode clusters integrated with an online database of specimen and taxonomic data with support for community annotations.
- An annotation framework that allows researchers to review and critique the taxonomic identifications associated with each BIN and notify data owners of errors.

BIN pages display aggregated data in several sections:

1. Distance Statistics	View BIN details including the member count, average distance between members, and BIN criteria. Also, details of the nearest neighbour BIN, including BIN GUID, taxonomy, and member count are provided
2. Taxonomy	The taxonomy of the public data is visible for the BIN, with highlighting to indicate taxonomy concordance and discordance.
3. Collection Locations	List of the collection countries and number of specimens collected per country.
4. Associated Publications	Details of publications that used sequences contained in the BIN.
5. Users Responsible for Data	Lists the owners of the public and private sequences contained within a BIN.
6. Dendrogram of Sequences	For BINS with a smaller number of sequences, a circle tree is displayed, and including the nearest neighbour. A PDF version of the tree is available for all BINs.
7. Specimen Images	View images for associated records.
8. Sampling Sites	Displays a map of the collection sites based on GPS values.
9. Attribution	Lists institutions where specimens are deposited, sequencing centres, photographers, collectors, taxonomists and funding sources.
10. Annotation	Pop-up window allows for annotation via tagging or comments on several aspects of the BIN page.

Descriptions of elements in BIN pages.

The BIN framework can greatly expedite the evaluation and annotation of described species and putative new ones while reducing the need to generate interim names, a non-trivial issue in barcoding datasets. The BIN algorithm has been effectively tested on a broad set of taxonomic groups and shows potential for applications in species abundance studies and environmental barcoding. The registry employs modern GUID and web service functionality enabling integration with other databases.



Example BIN page

Annotation

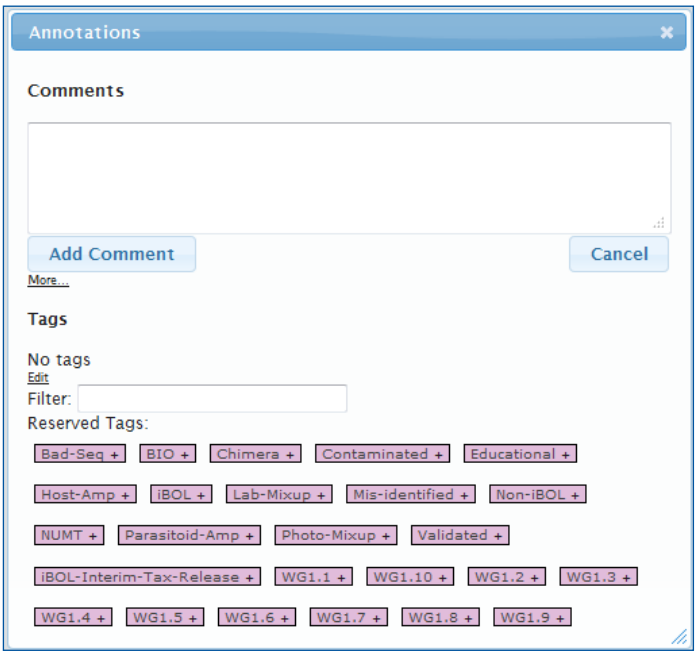
As the volume of barcode data being generated increases rapidly, the need for routine curation has become apparent. BOLD's annotation and notification system supports rapid community based validation of barcode data.

Annotation can occur at the project level, record level and also on specific data elements including taxonomy, images and sequences on BIN pages and public records.

Comments leverage the large user-base and expert knowledge for curation of both private data within collaborative projects and public data through the taxonomy browser and BOLD Public Data Portal.

Tagging allows for categorization using custom and controlled tags. Both custom and controlled tags can be used for filters, searches, and workflow management.

Comments and tags applied to your data by other users will appear in your activity feed on your User Console and the activity feed on the appropriate Project Console.



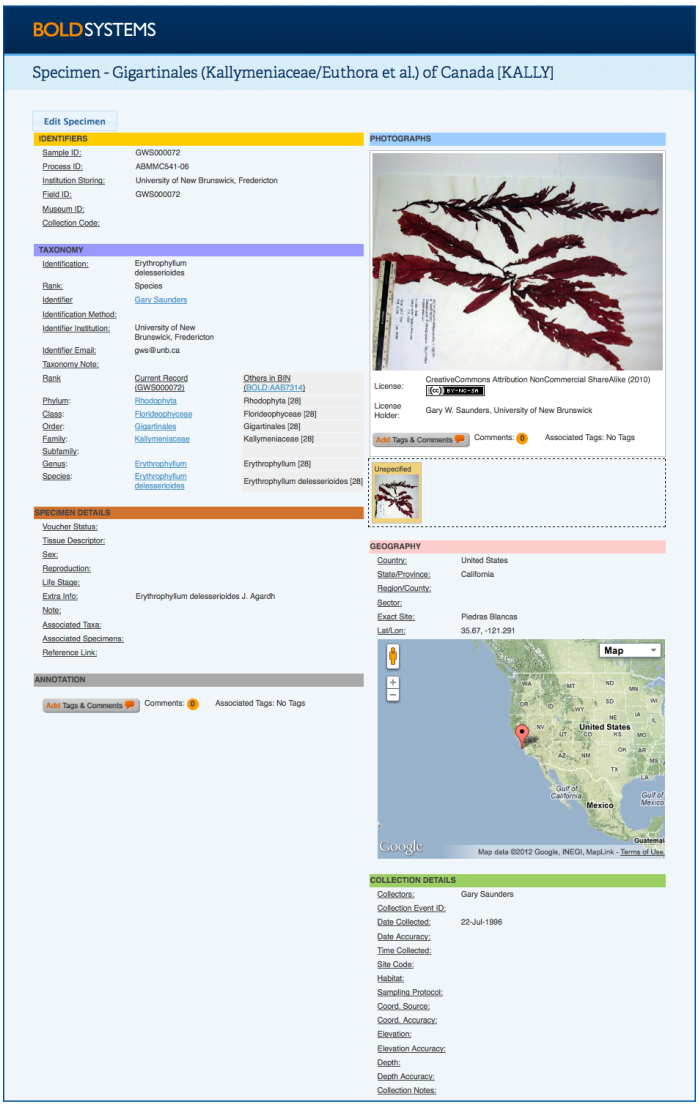
Annotation pop-up window

Specimen Data and Sequence Data Pages

BOLD connects specimen data with sequence data in a biphasic record. Please see below for what each part consists of, as well as how to navigate through the pages.

Specimen Data

The Specimen page stores voucher details, taxonomy, specimen details and collection data for a specimen. Any user with specimen editing permissions can edit the records by selecting "Edit Specimen" from the upper left corner. There is a world map marked with the location where the specimen was collected if GPS data is provided. The images for the specimen are located at the top of the window. By selecting an image, users will access a zoomable version with further details.



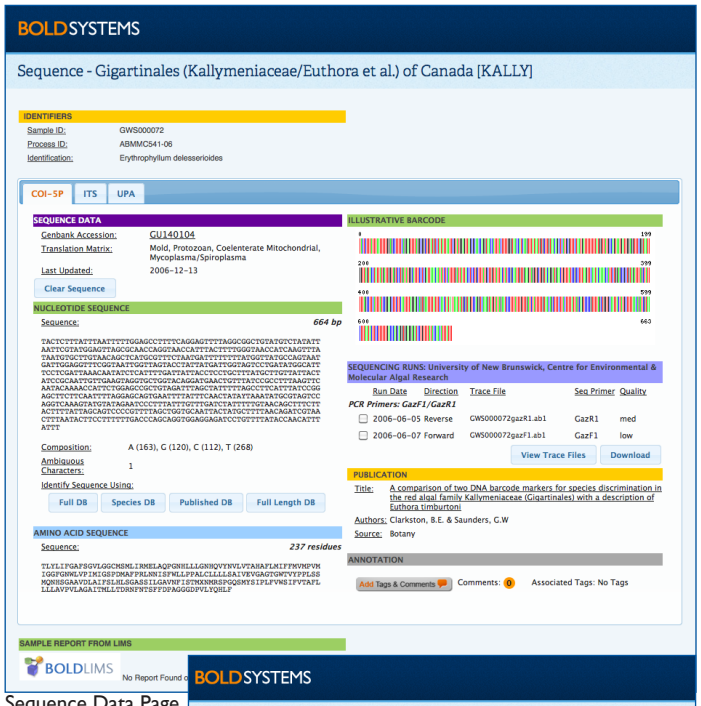
Specimen Data Page

Users can annotate specimen data, sequence data and individual images or traces using the "Add Tags and Comments" buttons on these pages.

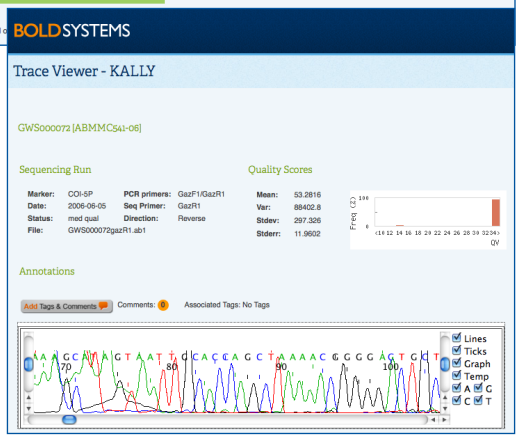
Sequence Data

The Sequence page stores details about the sequence data for a specimen. Different markers can be accessed by clicking on the tabs in the blue bar. Trace files can be viewed or downloaded from this window. Sequences can be deleted by users with full sequence access. If desired, the ID Engine can be used to identify the sequence from this page.

An illustrative barcode sequence of the species is provided by BOLD, along with a link to the Laboratory Information Management System (LIMS) for the Canadian Centre for DNA Barcoding when available. Finally, publication details will be available on the record once it is published.



Sequence Data Page



Trace Viewer

Workbench: Specimen Data Submission

The first step to creating records on BOLD, is specimen data submission. Each record is assigned a BOLD Process ID when uploaded. After specimen data is uploaded, images, traces, and sequences can then be uploaded. There are two ways to enter records onto BOLD: manually through the online interface or with bulk spreadsheet submissions through the BOLD Data Managers.
This protocol assists in the submission of bulk data to BOLD through the BOLD Data Managers. This is the easiest way to populate your project with records, as well as the only way to enter new taxonomy into the BOLD library. Described below are the specimen data fields available on BOLD. See the next page for further details.

*** Required Fields**

Sample ID *	ID associated with the sample being sequenced (often identical to or an extension of Field or Museum ID).
Field ID *	Field number from the collection event or the specimen identifier from a private collection.
Museum ID *	Identifier for specimen assigned by formal collection upon accessioning; also referred to as Voucher ID
Collection Code (only to be filled in if Museum ID is)	Code associated with given collection. Used in conjunction with Museum ID to disambiguate a specimen ID number that might be used in different collections within the same institution.
Institution Storing *	Full name of the institution where specimen or tissue is stored.

Table 1: Field definitions for Voucher info page on accompanying spreadsheet.

Sex	Male/female/hermaphrodite **only**.
Reproduction	Sexual/asexual/cyclic parthenogen **only**.
Life Stage	The age class or life stage of the specimen(s) at the time of sampling. The field supports free text but is a controlled vocabulary for validation. Example: "Adult", "Immature", "pupa", etc
Extra Info	A brief note or project term associated with the specimen for rapid analysis.
Notes	General notes regarding the specimen
Voucher Status	Status of the specimen in an accessioning process. Controlled vocabulary: • "Museum Vouchered:Type" • "Museum Vouchered:Type Series" • "Vouchered:Registered Collection" • "To Be Vouchered:Holdup/Private" • "E-Vouchered:DNA/Tissue+Photo" • "Dna/Tissue Vouchered Only" • "No Specimen"
Tissue Descriptor	A brief description of the type of tissue or material analyzed. Example: "muscle", "leg", "thorax", "liver", "blood", "feces", etc.
Associated Taxa	A list (concatenated and separated) of taxa associated with the taxon at the time of its collection. References to other taxa should be preceded by the relationship. Use of this field implies knowledge of an associated specimen where the identification is an inference. Use of the associated Specimen(below) field is preferred when the associated specimen is databased. Examples: "host: Quercus alba", "prey: caterpillar"
Associated Specimens	A list (concatenated and comma separated) of other specimens associated with the subject specimen at the time of its collection. References to other specimen identifiers should be preceded by the relationship. Examples: "host: PLANT23452, prey: USNM45677" when both prey and host specimens have been captured.
External URLs	Web accessible links that provide additional information about the specimen preceded by a descriptor. Multiple links should be pipe separated ("|").

Table 3: Field definitions for Specimen Details page on accompanying spreadsheet.

Full Taxonomy	Full taxonomy consisting of phylum*, class, order, family, subfamily (optional), genus, species in binomial format.
Identifier	Full name of primary individual responsible for providing taxonomic identification of the specimen.
Identifier E-mail	E-mail address of the primary identifier.
Identifier Institution	The full name of the identifier's institutional or organizational affiliation if one exists.
Identification Method	The method(s) used to identify the specimen.
Taxonomy Notes	Additional notes relating to the identification of the organism.

Table 2: Field definitions for Taxonomy page on accompanying spreadsheet.

Collectors	Comma delimited list of collectors.
Collection Date	Date of collection, must be in DD-MMM-YYYY
Continent	ISO Continents
Country/Ocean *	The full, unabbreviated name of the country, major political unit, or ocean.
State/Province	The full, unabbreviated name of the state, province, territory, or prefecture within the given country.
Region	Park, county, district, lake or river.
Sector	Sector of park or county/city.
Exact Site	Exact location of collection site
GPS Coordinates	Latitude & Longitude in "degrees.decimal degrees" format (e.g. 45.837).
Elevation	Elevation of sampling site. Measured in meters relative to sea level. Negative values indicate a position below sea level.
Depth	For organisms collected beneath the surface of a water body. Measured in meters below surface of water.
Elevation Precision	A numerical representation of the precision of the elevation given in meters and is represented as +/- the elevation value.
Depth Precision	A numerical representation of the precision of the depth given in meters and is represented as +/- the depth value.
GPS Source	The source of the latitude and longitude measurements.
Coordinate Accuracy	A decimal representation of the precision of the coordinates given in the decimalLatitude and decimalLongitude.
Event Time	The time or time of day during which the sample was collected.
Collection Date Accuracy	A numerical representation of the precision of the eventDate given in days and is represented as +/- the eventDate value. Default is 0 days.
Habitat	A category or description of the habitat.
Sampling Protocol	The name of, reference to, or description of the method or protocol used during a collection event.
Collection Notes	Comments or notes about the collection event.
Site Code	The name of the sampling location.
Coll. Event ID	A optional event ID.

Table 4: Field definitions for Collection Data page on accompanying spreadsheet.

As outlined below, data can be entered on the Data Submission Template spreadsheet and sent to BOLD. Data managers will review and validate the data, ensure that it meets the minimum requirements, and upload it into BOLD.

1. Create Excel file submission

- New submissions are project specific, so that their data can be associated with a project on BOLD. If records need to be entered into different projects on BOLD, a separate excel file for each project needs to be created. BOLD supports the upload of multiple specimen records in a spreadsheet format.
- The data spreadsheet consists of 4 worksheets; a main specimen identifier worksheet (voucher info) that is linked to three other worksheets: taxonomy, specimen details, and collection data. Tables 1-4 describes the information accepted in the batch specimen data submission. Minimal information can be submitted to start and records can be updated at a later date. Figures 1-4 below illustrates example data filled into the accepted fields for Template 3.0 (This spreadsheet template is available from the online version of this protocol in the Resources tab, or at: http://www.boldsystems.org/submissionTemplates/SpecimenData_v3Transitional.xls.)
- The minimal requirements for a new specimen record on BOLD are:
  - Voucher Info Page - Sample ID
  - Voucher Info Page - Field ID and/or Museum ID
  - Voucher Info Page - Institution Storing
  - Taxonomy Page - Phylum
  - Collection Page - Country

2. Submit file to BOLD for processing

- Open the destination project in BOLD
- Click on "Specimen Data" under the Uploads menu and choose "Initiate Batch Submission", and select "New" for the submission type. This option is available to project managers and project users with edit specimen access.
- In the form, select the Excel file to submit to this project, along with email addresses for collaborators that should be cc'd on further communications regarding the submission, a priority level and note if needed. Then click "submit" to submit the spreadsheet for the first pass of validation.
- If there are any errors detected with the first pass of a validation, please resolve these in the submission and re-submit.
- The Data Management team will contact you if there are any issues during validation, and once the records have been uploaded to your project.

Specimen Info				
Sample ID	Field ID	Museum ID	Collection Code	Institution Storing
demo01	Sample-demo01	15466-JUC-ISC	ISC	Burke Museum

	Taxonomy											
Sample ID	Phylum	Class	Order	Family	Subfamily	Genus	Species	Identifier	Identifier Email	Identifier Institution	Identification Method	Taxonomy Notes
demo01	Arthropoda	Insecta	Diptera	Asilidae	Hydro-psychinae	Efferia	Efferia aestuans	Joe Smith	jsmith@BIO.org	Oxford	Morphology	

	Specimen Details									
Sample ID	Sex	Repro-duction	Life Stage	Extra Info	Notes	Voucher Status	Tissue Descriptor	Associated Taxa	Associated Specimens	External URLs
demo01	Female	Sexual	Adult	Region 1	Collected with predator	vouchered: registered collection	leg	Predator: Hornet	Predator: BITK002-12	www.burke.edu/mus/spec15466

	Collection Info										
Sample ID	Collectors	Collection Date	Continent/ Ocean	Country	State/ Province	Region	Sector	Exact Site	Latitude	Longitude	Elevation	...
demo01	Joe Smith	2-Jul-2009	North America	Canada	Ontario	Wellington	Guelph	Riverside Park	43.563	-80.270	325m	

Depth	Elevation Precision	Depth Precision	GPS Source	Coordinate Accuracy	Event Time	Collection Date Accuracy	Habitat	Sampling Protocol	Collection Notes	Site Code	Collection Event ID
	2m		Garmin	1m	morning	2	dry forest	Malaise	park entrance	#14	#M872a

Figures 1-4: Example Specimen Data Submission.

Here are some important notes on fields for new or update submissions:

Sample IDs (Voucher Page):

- It is important to use a unique and original format for the Sample IDs. If the Sample IDs provided are not original on BOLD, they will need to be changed before the data can go online.
- Only the following characters may be used in the Sample ID, Field ID, and Museum ID: Numbers, letters, and ^ . : - _ ( ) # All other characters will be removed.

Collection Code (Voucher Page):

- The Collection Code must be used in conjunction with Museum ID in order to disambiguate a ID number that might be used in different collections within the same institution (i.e., a specimen number in a large museum may appear in the bird, mammal and reptile collections). This field is only to be used if Museum ID field is used.

Interim Species Names (Taxonomy Page):

- Interim names should contain non-Linnean characters such as numbers, punctuation and/or extra capitalization. Taxonomists are encouraged to append interim names with initials. (Example: Morpho sp. 1KHR)

Extra Info (Specimen Details Page):

- The "Extra Info" field can be displayed on a Taxon ID Tree on BOLD and thus it is possible to include information that may aid in analysis when illustrated on a tree.

Collection Date and Accuracy (Collection Page):

- If there is a date range, the Collection Date should be the mid-point, and the +/- can be entered into the Accuracy field.

What are Projects and Process IDs?

All of the data in BOLD are organized by projects. Related projects can be grouped into containers or temporarily merged with related projects for analysis, etc.
An individual entry in the database represents a barcode of a given specimen. The Process ID (assigned by BOLD upon specimen data record upload) uniquely represents a sample in BOLD. This is the identifier that is used to track a sample through the barcoding process: collection, taxonomic identification, sequencing, analysis and final publication of data.

Updating Specimen Data

An update means to modify records that already exist in a project. To only update one or two records, please manually select the specimen from the species record listing in your project and click on the "edit" button in the upper right corner. Any details can be edited in this way, except for adding new taxonomy to BOLD.

The quickest way to update a large number of records is to download and revise the Data Spreadsheet from BOLD. To do so:

1. Click on "Data Spreadsheets" from the Downloads menu on the left side of your project, merged projects or record search. (Please note: Records from any number of projects can be updated in one submission spreadsheet, and the number of records are (in theory) infinite for this type of update.)
2. Only download the worksheets that will be affected by the update (e.g. if the taxonomy needs to be updated, only download the Taxonomy worksheet; if specimen details and collection data need to be updated, only download the Specimen Details and Collection Data worksheets, etc. *Please do not download and submit updates on the progress report.*)
3. Modify the data on the downloaded worksheets. The submitted update must reflect what the data should be on BOLD.
4. Please send this to the Data Management Team through submissions@boldsystems.org noting the scope of the update, or submit through BOLD if it only affects one project.

If Sample IDs need to be changed after uploaded to BOLD, please contact the support team through support@boldsystems.org.

Workbench: Image Submission

Images should be uploaded to BOLD to complete a specimen record. An image provides support for identifications and makes comparisons easier between species.

This protocol outlines the image submission process for BOLD. It describes the necessary format of the images and the ancillary data and the steps required to build the uploadable package required for a successful submission.

1. Collect Images:
Group high-quality images of specimens in .jpg format for your records. BOLD accepts high resolution images (up to 20 megapixels), but only displays a greatly reduced thumbnail. The high resolution image is archived but will not be distributed without the submitter's explicit consent. Refer to page 18 for a guide on picture orientation and quality.

2. Assemble Package:
The image submission package should consist of all .jpg format images and a spreadsheet with the file names and ancillary data. Make sure that all images in the package are accounted for in the spreadsheet. When submitting more than one image per specimen simply copy the 'Sample ID' to the next line with the file name of the consecutive image.
You can upload up to 10 images per specimen, depending on organism characteristics. Please photograph several different orientations if needed.

Image File *	Complete (incl. extension) and identical file name (case sensitive) of images.	
Original Specimen *	Enter "Yes" if the image shows the actual specimen for this record. Otherwise enter "No".	
View Metadata *	Controlled vocabulary term to group media depicting a specific set of features of the organism or related environment. Dorsal, Lateral, Ventral, etc.	
Caption	Free text description of the subject. Short descriptions are recommended, such as: part of organism photographed, life stage, sex, etc. (400 Characters)	
Measurement	Any single relevant measurement that was taken in metric units.	
Measurement Type	Item or feature that was measured.	
Sample ID *	Sample ID for record, which must match Sample ID in BOLD.	
Process ID	Process ID for record (not necessary).	
License Holder*	The primary individual holder of the license. This is less critical when using creative commons licenses.	
License*	Pick one of the following license types or short-forms: • 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	(or) c nrr by by-sa by-nd by-nc by-nc-sa by-nc-nd
License Year*	The year of license declaration (not the year of submission to BOLD).	
License Institution*	The primary license holder's institutional or organizational affiliation. Decisions regarding use of material falls to the institution when the individual is unreachable or unresponsive.	
License Contact*	Contact information for the license holder. Can be an email address, mailing address, phone number, or all of the above.	
Photographer	The individual or team responsible for photographing and editing the media prior to submission.	

Field definitions for accompanying image submission spreadsheet. *** Required Fields**

The submission spreadsheet must be named ImageData.xls and contain the columns described in the table above. This spreadsheet template is available from the online version of this protocol in the Resources tab on BOLD, or at: http://www.boldsystems.org/submissionTemplates/ImageData.xls.

Steps:
A. Fill in the ImageData.xls spreadsheet with all the data related to the images in the submission package. To easily create the list of image files in a folder, open a terminal window (Start > Run > cmd in Windows), navigate to the folder containing the image files, and run one of the following commands:

- Windows — dir /b *.jpg>list.txt
- MacOS — ls *.jpg*.JPG>list.txt
- Linux/Unix — ls *.jpg*.JPG>list.txt

These commands will generate a list of all the files in the current folder and save it in a document called 'list.txt' that will appear in the current folder. You can then open list.txt and move the data into the Image File column. Please see the next page for an example of the completed ImageData.xls spreadsheet and further steps to upload.

Image File	Original Specimen	View Metadata	Caption	Measure-ment	Measure-ment Type	Sample Id	Process Id	License Holder	License	License Year	License Institution	License Contact	Photo-grapher
ROM912D.jpg	yes	Dorsal	skull	15 mm	skull length	ROM 10912	BMI272-03	Jane Beck	Copyright	2010	Biodiversity Institute	photo@bio.org	Joe Smith
ROM912L.jpg	yes	Lateral	lower jaw	7 mm	length	ROM 10912	BMI272-03	Jane Beck	Copyright	2010	Biodiversity Institute	photo@bio.org	Joe Smith
ROM913L.jpg	yes	Lateral	skull	15 mm	skull length	ROM 10913	BMI273-03	Jane Beck	Copyright	2010	Biodiversity Institute	photo@bio.org	Joe Smith

Image Submission Spreadsheet (ImageData.xls) completed with sample data.

B. These two components (Image files and Spreadsheet) need to be placed in a single folder. Compress them all into a single file before submitting. The following free tools are available to provide this functionality, however, most modern operating systems have built-in functionality for zipping:

- » WinZip - http://www.winzip.com
- » WinRar - http://www.rarsoft.com
- » MacZipIt - http://www.maczipit.com

C. BOLD will accept a maximum zipped file size of 190 MB. Upload the images to BOLD by clicking on the "Images" link in the Uploads menu of the User Console. Select the zipped folder of images to submit to BOLD. The images will appear immediately on the records once the upload is complete.

Tips and Troubleshooting for Image Uploads

- Zipped files must be under 190MB in size. If the upload fails to initialize, the zipped file may be too large. Break it into more than one upload, each with its own spreadsheet.
- Select the images and spreadsheet and zip those directly. Zipping the containing folder only may not work.
- The spreadsheet cannot contain any formulas.
- If the upload program cannot find the image files, it is possibly because it can not read the names. Make sure that the spreadsheet contains text values only.
- Full filenames must be used in the spreadsheet. The extension (.jpg or .JPG) must be included in the image file name, which is case sensitive.
- Spreadsheet must be named ImageData.xls. If the upload program can not find the spreadsheet, confirm that it is named correctly (case sensitive).
- Data must start on the second line of the spreadsheet. There is only one line for the column headers.
- Adding extra columns to the sheet will cause errors.
- Image names can not contain the characters "&" or "-". Please rename your images so that they do not have these characters.

You can upload more images in separate batches to any record at any time. If you wish to delete images for a record, please contact the BOLD Support Team through support@boldsystems.org.

Image Licensing and Use

BOLD assumes no license for images uploaded to the database. The image owner maintains the license and may change the license on the images at any time. Revisions to the given license should move towards making the license more liberal over time as it is very difficult to retract an open license and make it more restrictive.

If no license is chosen for an image, by default BOLD will forward all requests for that image to the owner for response. Adding a license reduces that burden and makes access easier. BOLD encourages the use of CreativeCommons - Attribution Non-Commercial Share-Alike, as this license has a good balance of protection and access.

There are three reasons for having images uploaded for each specimen on the BOLD database.

1. Quality Assurance - images can be used to confirm the taxonomic identification of organisms during sequence analysis.
2. Peer Review/Quality Assurance - once records are made public, peers can utilize your images and sequences to assist in their own validation of related specimens.
3. Taxonomy Browser Taxon Profiles - a random selection of the images on BOLD for each taxon are displayed on the public Taxonomy Browser at a highly reduced size (320 x 240) to create an online profile for each taxon that is stored in BOLD.

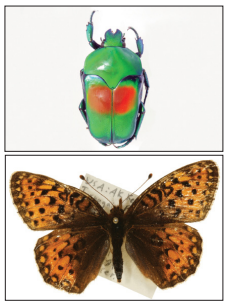
Photography Guide

This guide has been developed with assistance from the Canadian Centre for DNA Barcoding in an effort to provide best practices for taking voucher photographs. The BOLD database can accept up to ten images per specimen, so besides photos of a mounted or live organism, photos of distinguishing features or habitat are also acceptable.

To provide the best specimen image for viewing on the web, the following guidelines should be adhered to when appropriate:

- Please take pictures using the high quality mode on your camera (please note that BOLD can accept up to 20MP photos).
- The specimen should be centered in the image frame.
- Photos should be taken as close-up to the specimen as possible, leaving very little gap around the edges.
- It is often beneficial to include a measurement scale in the image to provide a size reference or a colour scale to provide colour reference.
- Background should be a contrasting colour where possible
- Standardizing the aspect ratio during specimen photography for your project(s) will make your images easier to compare
- Standardizing the position/orientation of each specimen will make it much easier to compare specimens across a project or BOLD.

See below for some common standardized orientations for animals, plants and fungi.

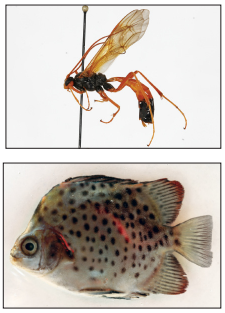


Dorsal

- The anterior of the specimen should be facing the top of the image frame.
- The specimen should be face-down, with the dorsal aspect of the head visible.

Herbarium Sheets:

- The full sheet should be included in the frame
- The sheet should be oriented so that text is legible in photo
- Label as "Herbarium Specimen"



Lateral

- The anterior of the specimen should be facing the left side of the image frame.
- The specimen should be oriented with the feet towards the bottom of the image.

Specific Features:

- These are often live photographs with focus on distinguishing characteristics
- For plants, these may include opened fruit, adaxial veins, male and female components
- Macro focus is recommended if available
- Label as "Leaf", "Fruit", "Buds", "Bark", etc

Ventral

- The anterior of the specimen should be facing the top of the image frame.
- The specimen should be face-up, with the ventral aspect of the head visible.

Whole Specimens:

- These are often live photographs of the organism in it's natural location
- The specimen should be centered in the frame and provide information such as shape of plant, etc
- If shooting outside, ideal environment has no wind and is slightly overcast (to avoid overexposure)
- Label as "in situ", "Plant", etc

Suggested orientations for photographs.

When entering a new orientation, please capitalize only the first letter and do not add any words (such as "Ventral view"). If your specimen does not fit into any of these categories, then please feel free to create a new category of view. (Displayed Specimen Images: All Rights Reserved)

Workbench: Trace Submission

Trace files provide support for sequences and should be uploaded for every specimen record. They can be uploaded once the data submission step is completed and BOLD has assigned a Process ID to each record. This protocol assists in the submission of trace files to BOLD. It describes the necessary format of the files and the ancillary data that is required for the correct submission.

1. Confirm primers are registered on BOLD.
See page 8 for details on how to search the primer database to confirm your primer is registered with BOLD.

2. Assemble Package:
The submission package consists of trace files (.ab1 or .scf), corresponding Phred (score) files if available (.phd.1) and a spreadsheet with the file names and ancillary data. The submission spreadsheet must be named data.xls and contain the columns described in the table to the right. This spreadsheet template is available from the online version of this protocol in the Resources tab, or at: http://www.boldsystems.org/submissionTemplates/data.xls

Trace File *	Complete (including extension) and identical file name (case sensitive).
Score File	Complete (including extension) and identical file name (case sensitive).
PCR Primers Fwd/Rev *	Primer codes are case sensitive. Both must be filled in.
Sequence Primer	Primer codes are case sensitive.
Read Direction *	Forward or Reverse.
Process ID *	Process ID of record, which must match Process ID in BOLD.
Marker (2 blank columns must be left after the Process ID column)	If sequencing multiple genes, the marker needs to be filled in to match the short form marker code in your project, such as one of the following: • COI-5P • ITS • rbcLa • matK

Field definitions for accompanying trace submission spreadsheet.
*** Required Fields**

Steps:
A. Fill in the data.xls sheet with all the data about your files. To easily create the list of the files in a folder, you need to open a terminal window (Start > Run > cmd in Windows), navigate to the folder where the trace and score files have been placed and run one set of the following commands:

- Windows dir /b *.ab1>ab1.txt and dir /b *.phd.1 >phd.txt
- MacOS ls *.ab1>ab1.txt and ls *.phd.1 > phd.txt
- Linux/Unix ls *.ab1>ab1.txt and ls *.phd.1 > phd.txt

These commands will generate lists of all the files in the current folder. They will be saved as ab1.txt and phd.txt text files. You can then open the text files and move the data into the appropriate columns, as illustrated below.

To obtain the Process IDs for a set of records on BOLD, click on "Data Spreadsheets" under the Downloads menu on the left side of a project console. Download the Progress Report (Downloads>Specimen Data) to get the Process IDs that are assigned to each Sample ID submitted to BOLD.

Trace File	Score File	PCR Fwd	PCR Rev	Seq Primer	Read Direction	Process ID	blank	blank	Marker
KKBNA001-04.ab1	KKBNA001-04.phd.1	BirdF1	BirdR1	BirdF1	Forward	KKBNA001-04			COI-5P
KKBNA001-04r.ab1	KKBNA001-04r.phd.1	BirdF1	BirdR1	BirdR1	Reverse	KKBNA001-04			COI-5P
KKBNA002-04.ab1	KKBNA002-04.phd.1	BirdF1	BirdR1	BirdF1	Forward	KKBNA002-04			COI-5P

Trace File Submission Spreadsheet (data.xls) completed with sample data.

B. These components (Trace files, Score files and Spreadsheet) need to be placed in a single folder. Compress them all into a single zipped file before submitting.

C. BOLD will accept a maximum file size of 190MB. Upload the traces to BOLD by clicking on the link "Trace Files" in the Uploads menu on the User Console. Select the zipped folder of files, and choose the institution that performed the sequencing to submit to BOLD.

Tips and Troubleshooting For Trace Uploads

- Primers must be registered before upload. If the primers are not registered, there will be an error. Please refer to the next page for details on how to register new primers.
- Zipped file must be under 190MB in size. If the upload fails to initialize, it is probably because the zipped file is too large. Try breaking it into more than one upload, each with its own spreadsheet.
- Select the traces and spreadsheet and zip those directly. Zipping the containing folder only may not work.
- The spreadsheet cannot contain any formulas.
- If the upload program can not find the files, it is possibly because it can not read the names. Make sure that you have text values only in the spreadsheet.
- Full filenames must be used in spreadsheet. The extension (.ab1, .scf, .phd.1) must be included in the file name. These extensions are case sensitive.
- The spreadsheet must be named data.xls. If the upload program can not find the spreadsheet, confirm that it is named correctly (case sensitive).
- Data must start on the second line of the spreadsheet. There is only one line for the column headers.
- Do not add extra columns to the spreadsheet.
- Trace files may not be downloaded or viewed BOLD until processed (up to 24 hours after they have been submitted).

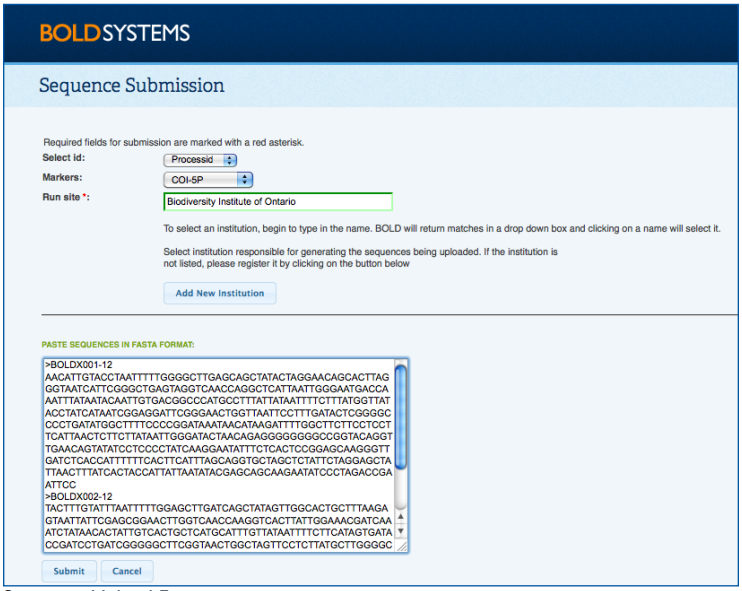
You can upload more traces in separate batches to any record at any time. If you wish to delete any traces for a record, please contact the BOLD Support Team through support@boldsystems.org.

Workbench: Sequence Submission

This protocol outlines the DNA sequence submission process on BOLD, describing the sequence format and steps required for a successful submission. Only users with sequence editing access on a project may upload sequences.

1. Assemble Package:
The sequence submission package should consist of aligned sequences in FASTA format referenced by BOLD Process IDs or Sample IDs.
To upload with **Process IDs**, the FASTA header line must conform to the following format: it should begin with a '>' followed by the Process ID, with any additional information separated by either a bar ('|'), an underscore ('_') or a space (' '). There can be no spaces before the end of the Process ID.

To upload with **Sample IDs,** the FASTA header line must conform to the following format: it should begin with a '>', followed by the Sample ID, with any additional information separated by a bar ('|'). Do not use a space or an underscore to separate information from the Sample ID.



Sequence Upload Form

2. Upload Package:
You can include up to 1000 sequences into one upload. Upload the sequences to BOLD by clicking on "Sequences" in the Uploads menu of the User Console. Select the marker and the institution that assembled the sequences. Paste the sequences into the text box. When confirmed, "submit" to upload the sequences. These will appear immediately on the records.

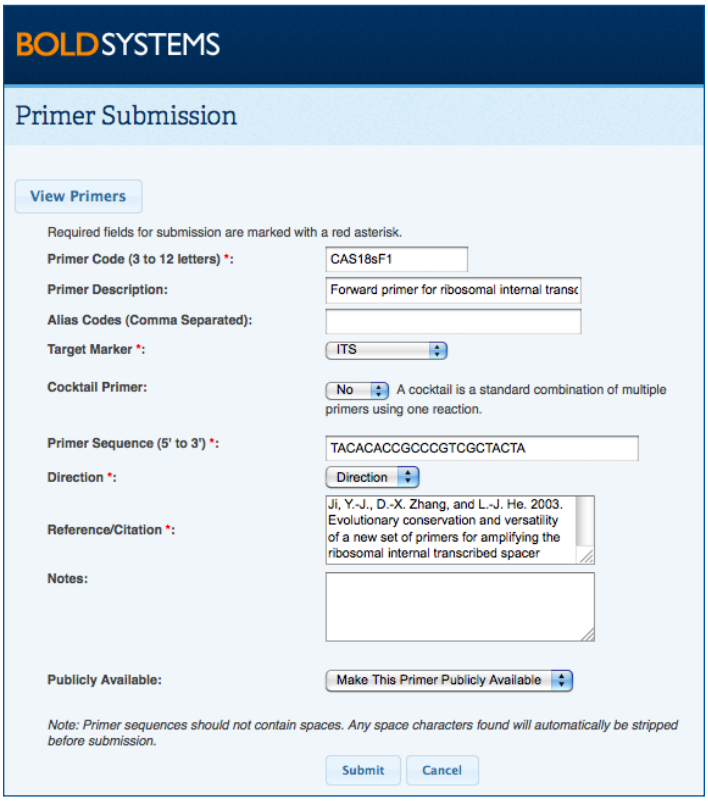
- If you wish to replace a sequence on BOLD, simply upload the new one with the same Process ID or Sample ID.
- To delete an individual sequence, you can do so by using the Delete button within a record's sequence data page (for more info on Sequence pages, please see page 12). Contact the BOLD support team through support@boldsystems.org for batch deletions.

Workbench: Primer Submission

New primers need to be registered on BOLD prior to submitting a trace file package. To register new primers, select "Register Primers" from the User Console. Please note: If a primer sequence has already been registered under a different code, you will be provided with the registered code to be used in your submission. Primers you register on BOLD can be edited at any time after they are created (e.g. to make them public).

Primer Code	Create a code for the primer. If the primer is already published in a manuscript, please use the code that is in press.
Primer Description	A description of what the primer is used for.
Alias Codes	Other known code names for the primer, separated by commas.
Target Marker	Select the target marker from the controlled list (e.g. ITS, COI 5', matK, etc.).
Cocktail Primer	Select "Yes" if it is a cocktail primer. This will create extra fields to add multiple sequences.
Primer Sequence	Fill in the sequence(s), 5' to 3'.
Direction	Select the direction of the sequence.
Reference/Citation	List references and/or citations.
Notes	Any notes about the primer.
Publicly Available	If the primer has already been published, or if you wish to make it publicly available, this should be left public. The other option is to keep the primer private until publication.

Field definitions for Primer Submission.



BOLD Primer Submission Form

Workbench: Bibliography Submission

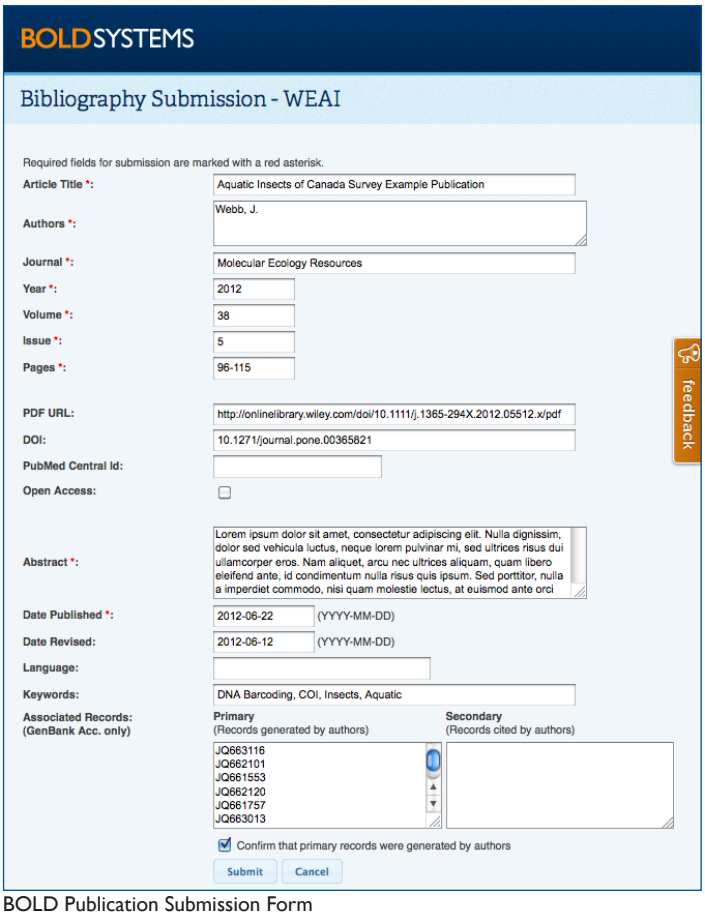
Users can submit bibliographies to BOLD using the Bibliography Submission Form available in three locations:

- From the User Console under Data Uploads
- From a Project Console or Dataset Console
- From a list of searched records (Record List page)

Any user with edit sequence or edit specimen permissions to records will have the ability to submit a bibliography connected to those records as primary or secondary associations.

The primary and secondary GenBank accessions can be filled in here, separated by a line. These accessions will auto-populate in the form for the records selected if the submission is from the Project Console or Record List. (See page 26 for details on submitting records for GenBank Accessions via BOLD)

The publication details will then appear in the Publication Database, and the citation will appear on each record.



BOLD Publication Submission Form

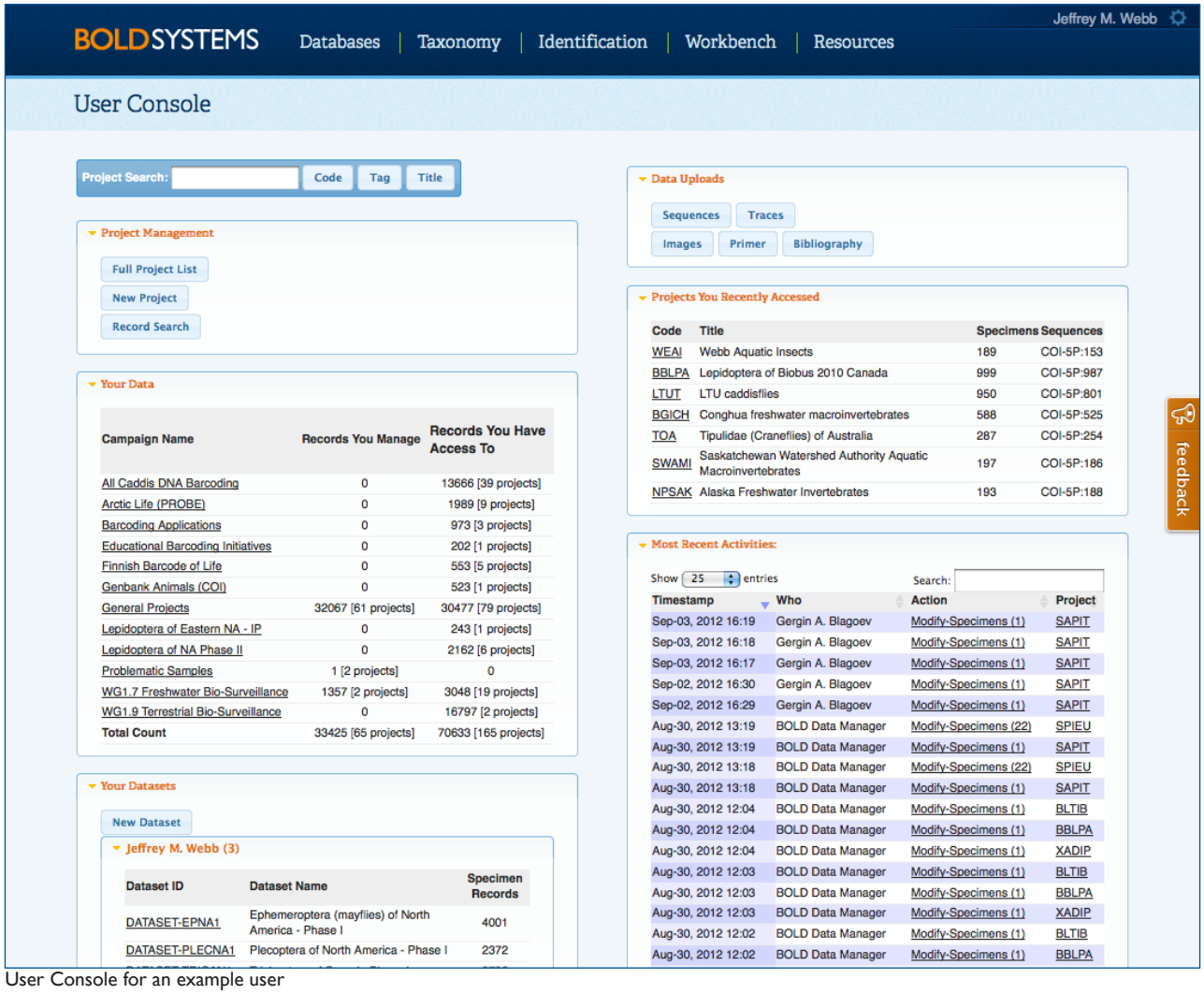
Workbench: User Console

The newest version of BOLD introduces a new management console as the landing page when users log in. This new console allows for rapid access to frequently accessed projects, near real-time reports on project activity as well as powerful search tools.

Data Uploads
Upload sequences, traces, images, primers, and bibliographies directly from user console.

Real-time Activity Reporting
BOLD maintains a detailed log of all the actions taken by users in the system. This log is transformed into a report where events pertinent to each user are extracted and displayed on this console. As users work together on projects, submitting and refining data, the Activity Report allows them to stay informed on the steps taken by their colleagues. Logs can be downloaded, allowing users to keep personal records and perform additional analysis.

Searching
This console provides two ways to perform searches. The first is a project search, where a user looking to open a project can jump directly to it by entering the code in the project search bar. If the code is not known, the user can generate a short list by entering a project tag or part of the project title.
The second search functionality, "Search Records", generates a list of records based on search terms consisting of geography, taxonomy, tags, sequence length, and pasted lists of identifiers. Records retrieved from a search can be downloaded or analyzed right on the system.



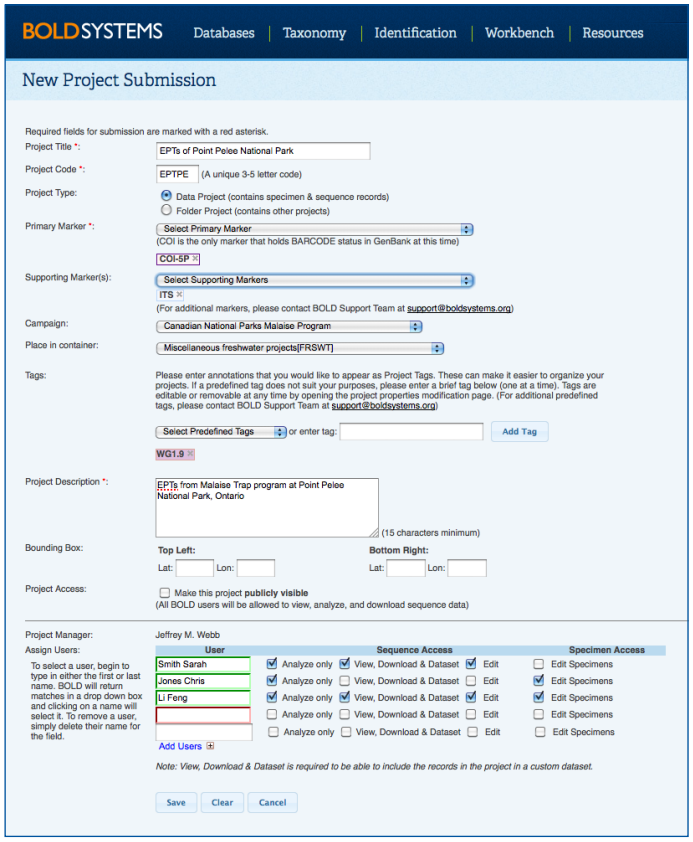
User Console for an example user

Workbench: Creating a project

To upload Barcode records to BOLD, a project must first be created in which to house the records. From the User Console, select the 'New Project' button from the Project Management menu.
All project details can be edited at any time (with the exception of Project Code, Project Type and Project Manager) by simply clicking on 'Modify Project Properties' in the Project Options menu of the Project Console. Only the Project Manger is able to modify these details.

Project Title*	Please create a descriptive name.
Project Code*	A 3-5 letter code that needs to be unique across BOLD. A good approach is to use your initials and 2 or 3 other letters as an acronym for the title.
Project Type*	Choose between the following options: • Data Project (contains records) • Folder Project (contains other projects)
Primary Marker*	Select your primary marker. CO1 is the default. Primary marker options are: • Cytochrome Oxidase Subunit 1 5' (COI-5P) • Interspacer Region (ITS) • Ribulose-bisphosphate carboxylase (rbcL) • Maturase K (matK)
Supporting Markers	Select as many secondary markers as needed from the list of registered markers.
Campaign	Select the name of the campaign the project is part of if desired.
Place in Container	Select the name of the Folder Project if desired.
Tags	Please enter annotations that you would like to appear as Project Tags on the Project List and Project Console pages. These Project Tags can make it easier to organize or define relationships between your projects.
Project Description*	Enter a summary of the use and intention of the project. 15 - 500 characters.
Bounding Box	Define the bounding box of the collection area covered by the project using GPS coordinates.
Project Access	Check to make project publicly visible on BOLD and submit to the BOLD Public Data Portal.
Project Manager	The person who creates a project is automatically the Project Manager.
Assign Users	Other BOLD users can be added to a project. Different levels of access are possible, and are described below. **Sequence Access:** • Analyze Only - user can perform analysis on the data, but cannot view more than a summary of the data (sequence and related information remain hidden). • View & Download - user can view or download the sequence data, as well as analyze. • Edit Sequences - user can upload trace files, upload, edit and delete sequences, as well as view and analyze. **Specimen Access:** • Edit Specimens - user has control over sample identifiers, taxonomy, collection data, and images of the specimens: this edit permission level is intended for project managers, collectors, and taxonomists.

Field definitions for BOLD project creation form. *** Required Fields**



New Project Form

Please note that the person who creates a project is automatically assigned as the **Project Manager**. To change the project manager, the current Project Manager must send a request to the BOLD support staff through support@boldsystems.org.

Supporting markers are added upon request. If a marker you require is not on the list, please contact BOLD support staff to register one through support@boldsystems.org.

Workbench: Managing Data

Once a project has been populated with the specimen data, images, traces and sequences that have been uploaded to BOLD, it will resemble the figure on the right.

Project Console

The project console presents an overview of the status of records within the projects as well as an audit trail of the activities in the project. This includes a report of the number of specimens, along with tallies of any missing components of the records. Also included are graphs to provide a quick visual overview of the project, as well as a list of all the users with access to the project. The links to the left provide access to uploads, downloads and various analysis tools.

Project Managers will see the "Modify Project Properties" button with which they can change the project title and description, add or remove markers, and add, remove or modify permissions of users at any time. The Project Manager also has access to publish the records in the project to GenBank. (See page 26 for more details on GenBank submissions.)

To access a list of the records within each project, click on "View All Records" in the project options menu.

Record List

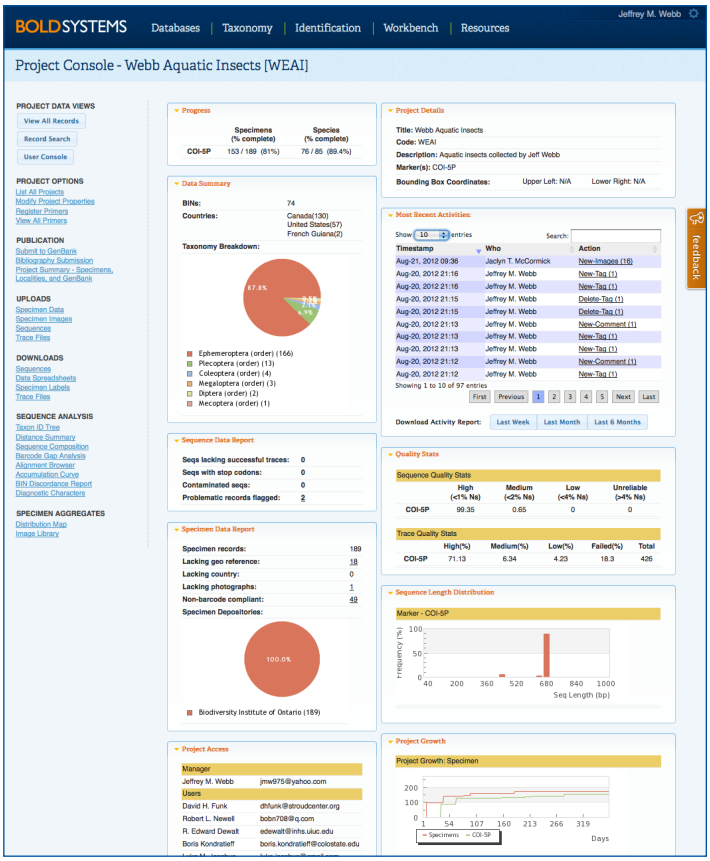
A Project Record List is the full list of all records within a project, along with the actions and tools open to a project member.
The record list gives access to individual specimen and sequence data for each record. The red arrows along the column headers can be used to sort the records by header.
Users can select specific records for download, analysis, or annotation using the checkboxes.

Flags

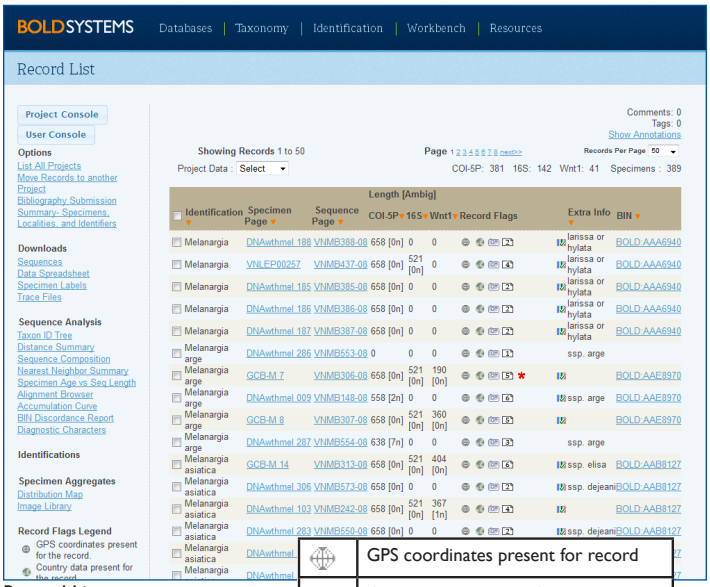
- Icons appearing next to a record indicate the presence of certain characteristics of a record; see legend to the right for more details.
- A red-highlighted, bolded sequence length is a warning that the sequence contains more than 1% ambiguity and won't meet the Barcode Standard.

The Project Manager or a user with full edit permissions can move records from one project to another by selecting the records needed and then clicking on "Move records to Another Project" in the Options menu. The destination projects that will appear in the list will be ones in which the user has full edit permissions.

Click on the Sample ID or the Process ID to access the Specimen Data and Sequence Data respectively, for each record. These are illustrated on page 12.



Project Console



Record List

	GPS coordinates present for record
	Country data present for record
	Images present for specimen
	The number of traces present
	Stop codons present in sequence
	Contamination present in sequence
	Flagged record, not in ID engine
	Sequence is Barcode Complaint

BOLD Record List icon legend

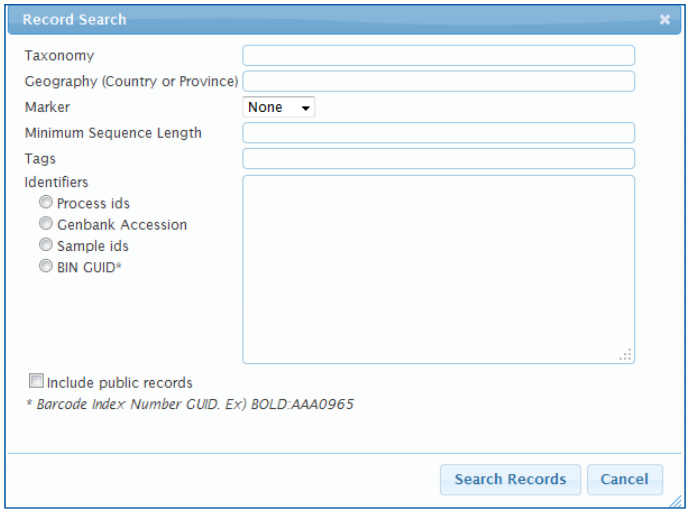
Workbench: Record Search

The improved search interface in the BOLD workbench allows for rapid access to large numbers of records.

To search across all records that a user has access to, select the "Search Records" button in the User Console or Project List. The form gives the option to add in all public records on BOLD.

To search for a sub-set of records from within a merged set of projects or a single project, select the "Search Records" button from the Project Console.

The results from the search will appear in a Record List like the one illustrated on the previous page.



Record Search Engine pop-up window

Taxonomy	Searches specific taxonomic names. Use quotes to surround Species name and separate multiple taxa by spaces. Example: "Bos taurus" "Bos indicus"
Geography	Searches the country and province names. Use quotation marks to surround multi-word names and separate multiple terms by spaces. Example: "Costa Rica" Mexico "United States"
Marker	Filter results by the presence of a sequence for a specific marker
Min. Seq Length	Define minimum number of base pairs desired.
Tags	Enter terms found in tags on records
Identifiers	Search by Sample IDs, Process IDs, GenBank accessions and BIN GUIDs. Can be used in conjunction with the above fields to return the intersection of records.
Include Public Records	Checking this box will allow the search to go across all public data on BOLD, as well as the records in projects the user has direct access to.

Search Parameters Available

Workbench: Publication of Records on BOLD and GenBank

Publication through BOLD

To submit records to the BOLD Public Data Portal, the Project Manager can make a project public. This option is available in the Project Properties form which the Project Manager can access by clicking on "Modify Project Properties" within a project.
Once GenBank accessions have been obtained, submit a bibliography to the BOLD database following the directions on page 21. This will associate your publication with the records using the GenBank Accession numbers. This citation will appear on the records in the workbench and the Public Data Portal.

Publication to GenBank, NCBI

BOLD shares a tightly integrated data exchange pipeline with NCBI (GenBank) that allows for the automatic submission of data to GenBank. Users are only required to fill in the author and publication information and which is sent to GenBank along with the specimen, sequence, and trace data which has been transformed to the required formats. GenBank responds directly to the user with the accessions for their records to be included in publications. Accessions are also sent to BOLD to ensure bidirectional linkage.

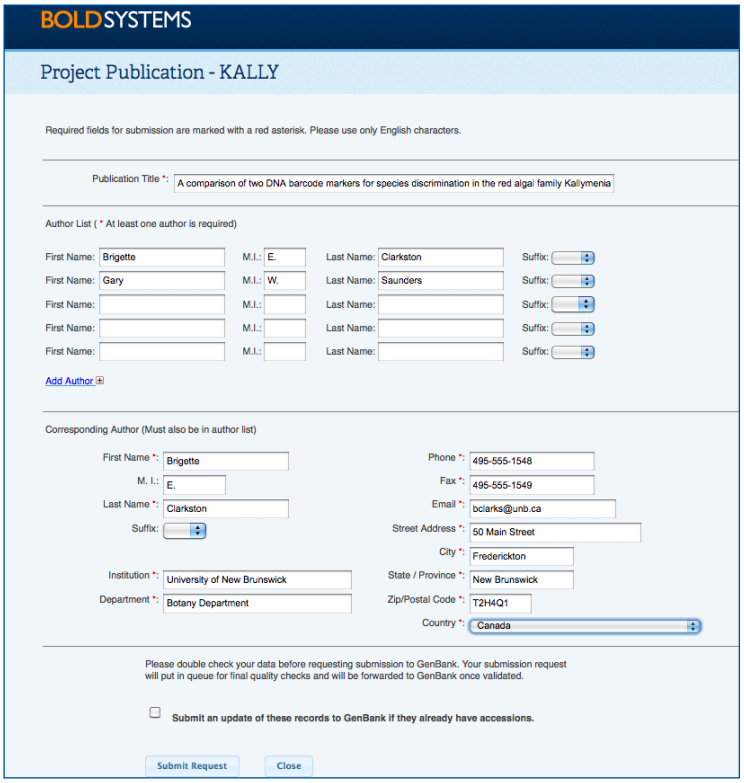
The data exchange pipeline is further utilized to send GenBank updates to records. Identifications of records submitted through BOLD to GenBank can still be refined and updated as new information is obtained. Changes to the taxonomy of BOLD records are automatically sent to GenBank on a weekly basis so that GenBank has the most current and up to date information.

To submit a set of sequences to GenBank, the Project Manager can access the form shown to the right via the "Submit to GenBank" button within a Project Console.
All records within the project will be submitted to GenBank. If only a subset of records need to go to GenBank, then the records should be moved to a new project or copied to a BOLD Dataset which can be submitted.

GenBank accession numbers are generally returned by email to the project manager within five business days. The accession numbers will be associated on BOLD with your records for quick reference.

Submissions to GenBank from BOLD are automatically locked for 1 year to allow time for publication. If the publication is released sooner than one year, the corresponding author should contact GenBank directly to request public release at the time of publication.

Contact the BOLD support team through support@boldsystems.org for assistance with any aspect of publication.



GenBank Submission Form

*** Required Fields**

To access the list of GenBank accessions for a set of records on BOLD, select the "Summary - Specimens, Localities, and GenBank" within a project or record list.

Workbench: Analytical Tools

BOLD includes core and extended tools to analyze specimen and sequence data:

Core Analysis Tools

- Image Library: Compare morphological characteristics
- Distribution Maps: Interact with geographical data
- Taxon ID Tree: Visualize a neighbour joining tree with matching images
- Barcode Index Numbers (BINs): Barcode clusters (see page 10)
- Identification Engine: Locate closest matches to an unknown sequence (see page 6)

Extended Analysis Tools

- Distance Summary: Browse sequence divergence at multiple taxonomic levels
- Sequence Composition: Explore compositional variation at all codon positions
- Barcode Gap Analysis: Evaluate the Barcode gap
- Accumulation Curve: Review sampling efficiency
- Alignment Browser: Diagnose unaligned sequences
- Diagnostic Characters: Examine polymorphism
- BIN Discordance Report: Utilize BINs to highlight possible issues

Tip: Run Multiple Analytic Tools in Parallel

Use the newly available option to run multiple analyses and have the results emailed to you when the analysis is finished Results can be stored for up to 4 weeks, saved for future comparison, and links to the results can be shared between collaborators.
Find this option on the parameters page for most analysis tools.

Workbench: Image Library

Once images have been uploaded to a project, they can be viewed in two ways. The first is by opening an individual record where corresponding images will be displayed. The second is the Image Library for viewing a group of specimens, shown in the figure to the right.

The Image Library allows sorting by orientation so users can compare morphological differences between specimens. This tool is useful for diagnosing contamination or misidentifications as taxonomy is displayed below each image.

Image Licensing is viewable upon mouse rollover. To view the attribution and further details on specific images, open the specimen data for that record from the record listing page.

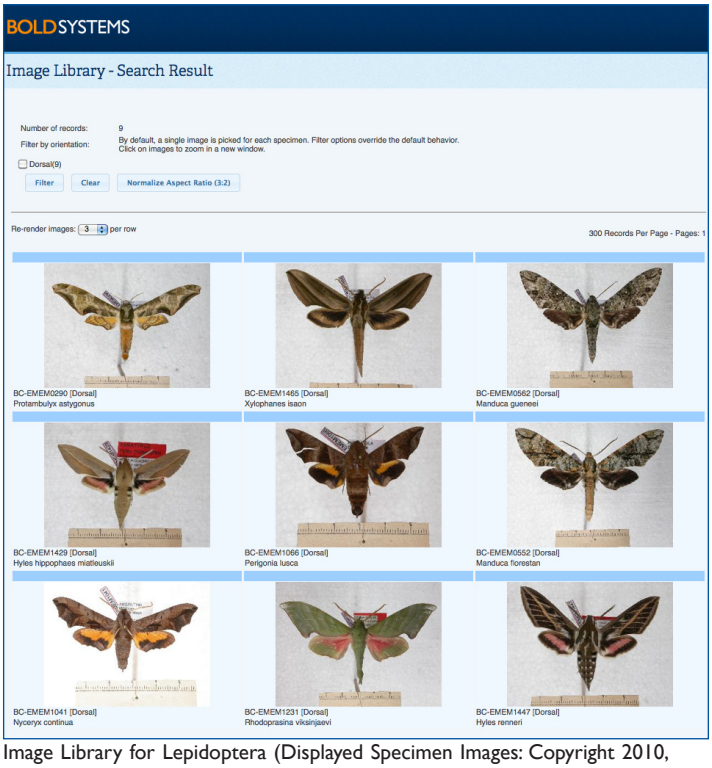


Image Library for Lepidoptera (Displayed Specimen Images: Copyright 2010, Research Collection of Ulf Eitschberger).

Workbench: Distribution Map Analysis

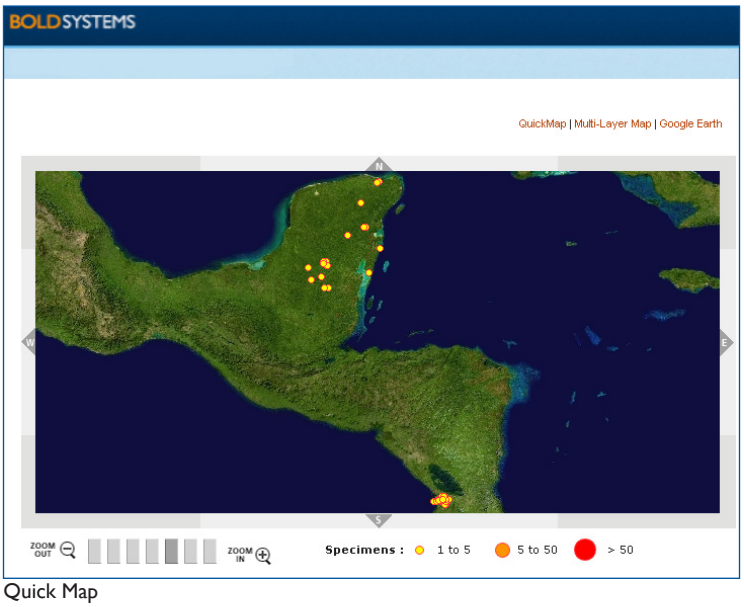
BOLD's Distribution Maps plot the collection points for a selected set of specimens when geographic reference data is available. The multiple mapping tools available on BOLD are described below.

Quick Map

The Quick Map is built using the NASA Blue Marble Project. They are open access and therefore can be re-used and modified at the user's discretion.

Users can zoom in to a max of 1 km per pixel by using the scale at the bottom or clicking on a region in the map and can pan the map in 4 directions by clicking on the N/E/S/W directions in the frame.

The collection points are shown on the map using markers specifying the density of sampling at each point. Larger markers are placed beneath smaller ones so all points can be visible.



Quick Map

Multi-Layer Map

The Multi-Layer Map is based on Google maps. (Google gives permission for use in publications as long as the Google logo remains on the image.) The layers include political boundaries with regional names, as well as a satellite map of the world. These can be viewed individually or combined, which is shown to the right.

The markers on this map are active and can be clicked to retrieve a list of the specimens from BOLD.

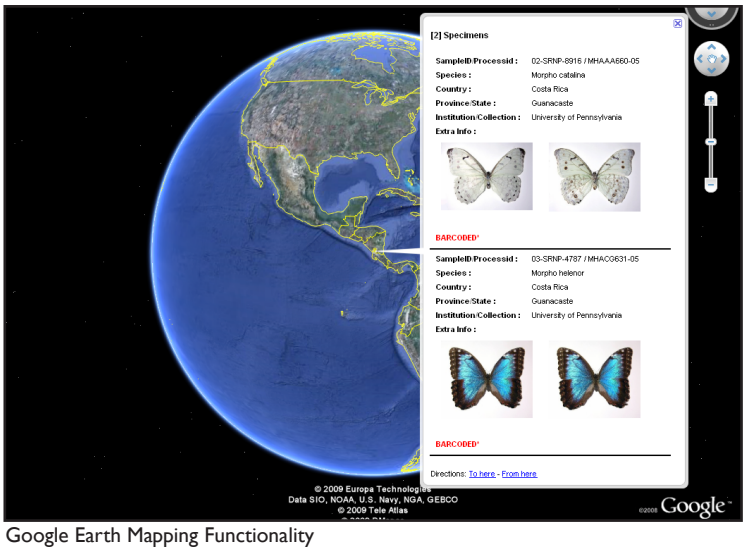
This record list is also active, meaning users can open specimen records, which can be edited directly. This is a great way to validate and correct GPS data.

Multi-Layer map with hybrid view

Google Earth

The Google Earth map is a display of the specimen collection points in the program Google Earth. This is free software that can be downloaded from the web. (Google gives permission for use in publications as long as the Google logo remains on the image.) The benefit of this type of map is that it provides a portable KML file download which can be shared among colleagues.

This map is embedded with specimen images, along with specimen identifiers, country, province/state, institution/collection, and extra info.



Google Earth Mapping Functionality

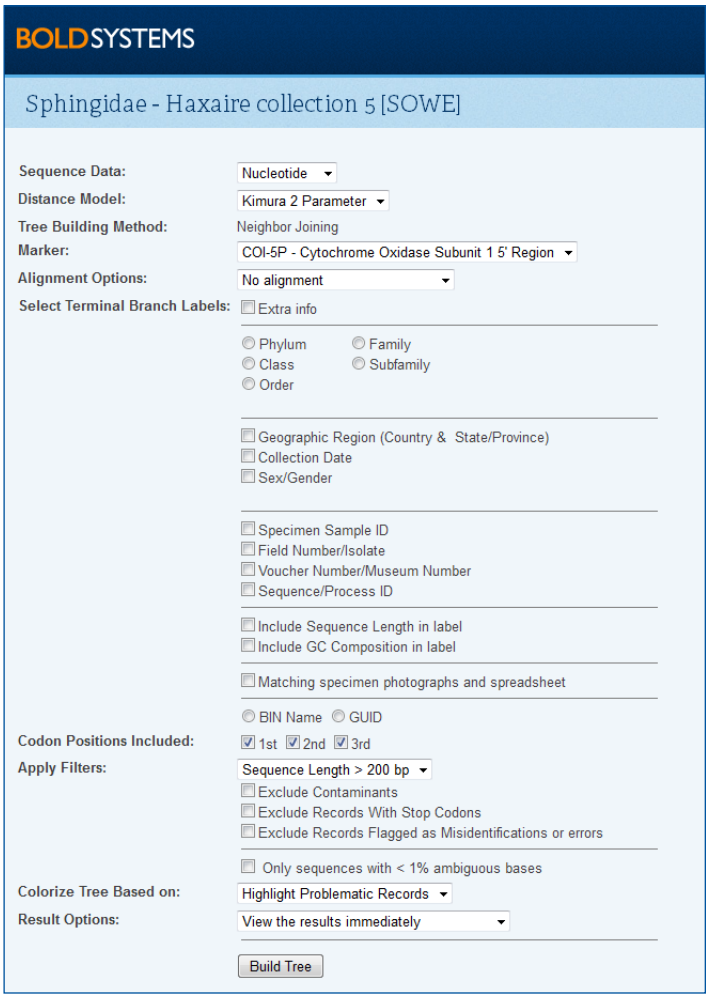
Workbench: Taxon ID Tree

The Taxon ID Tree functionality allows for the generation of dendrograms from sequencing using the Neighbour Joining algorithm. Sequence alignment and multiple labelling options are available. This tool can be accessed by clicking on "Taxon ID Tree" under the sequence analysis panel on the Project Console and Record List pages.

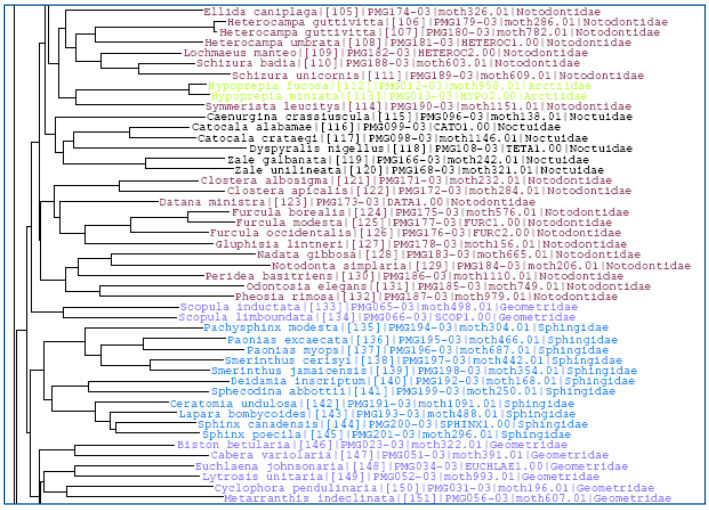
Result of the analysis include a tree in PDF format for easy dissemination and publication, Newick format for rendering in other formats and layouts (i.e. circular), and a taxonomy report that provides a breakdown of the data used. When the option to include matching images and spreadsheet is selected, the system returns a spreadsheet with specimen identifiers and full taxonomy in the same order the records appear in the tree as well as the images of the specimens in the same order as the tree. Utilizing these options help catch data entry and lab errors where samples may be mixed up. They are also useful for detecting misidentified specimens.

Sequence Data	• Nucleotide • Amino Acid
Distance Model	• Kimura 2 Parameter • Jukes Cantor • Pairwise Distance
Tree Building Method	Neighbour Joining is the only method at this time.
Marker	Select the marker from a multi-gene dataset.
Alignment Options	• No Alignment • Allow BOLD to align sequences • Kalign • Muscle
Select Terminal Branch Labels	Many options for labels to add to the end of each branch including taxonomy, geography, identifiers, sequence details, BIN identifiers
Photographs	Option to include matching specimen photographs and spreadsheet for comparison.
Codon Positions Included	1st, 2nd and 3rd Codon Positions are included as default but may be excluded
Apply Filters	Can be applied to disregard sequences below a given length (since very short sequences can skew the results) or more than 1% ambiguous bases. Exclude problematic sequences.
Colourize Tree Based on	• Problematic Sequences • Taxonomy: Class • Taxonomy: Order • Taxonomy: Family • Taxonomy: Subfamily • Location: Country • Extra Info • Sequence Age • BINs
Result Options	Choose to view the results immediately or to have the results emailed to your account when available.

Parameters available for Taxon ID Tree



Taxon ID Tree Parameter Page

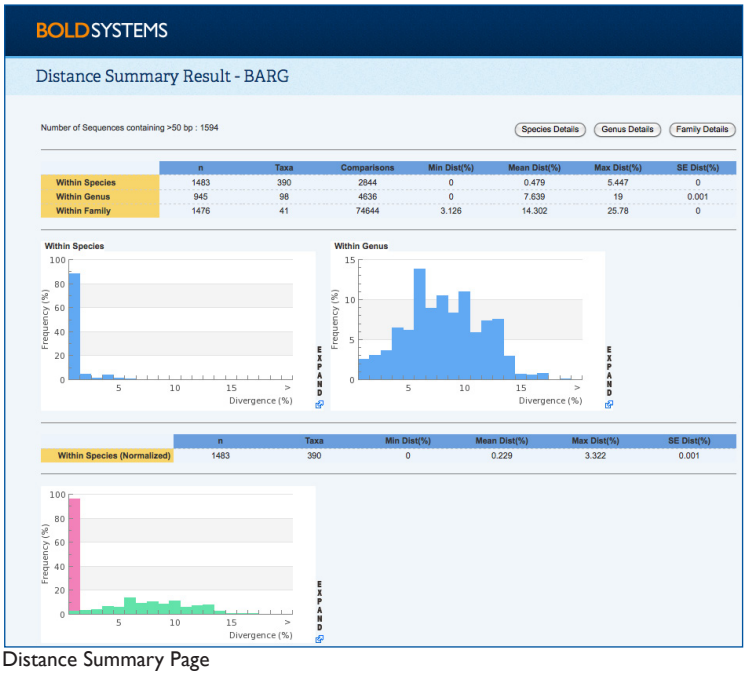


Standard Taxon ID Tree (Lepidoptera)

Workbench: Distance Summary

It is desirable for barcodes to show very low sequence divergence within a species, with significantly higher sequence divergence at higher taxonomic levels. The Distance Summary tool gives a report of sequence divergence between barcode sequences at the conspecific and congeneric levels.

Distance values are calculated using a user selected distance metric. Comparisons are performed between the given taxonomic levels with the frequency plotted as shown in the figure to the right. Details for the comparisons done at the level of species and genus are available by clicking on the links in the top right corner of the Distance Summary browser.

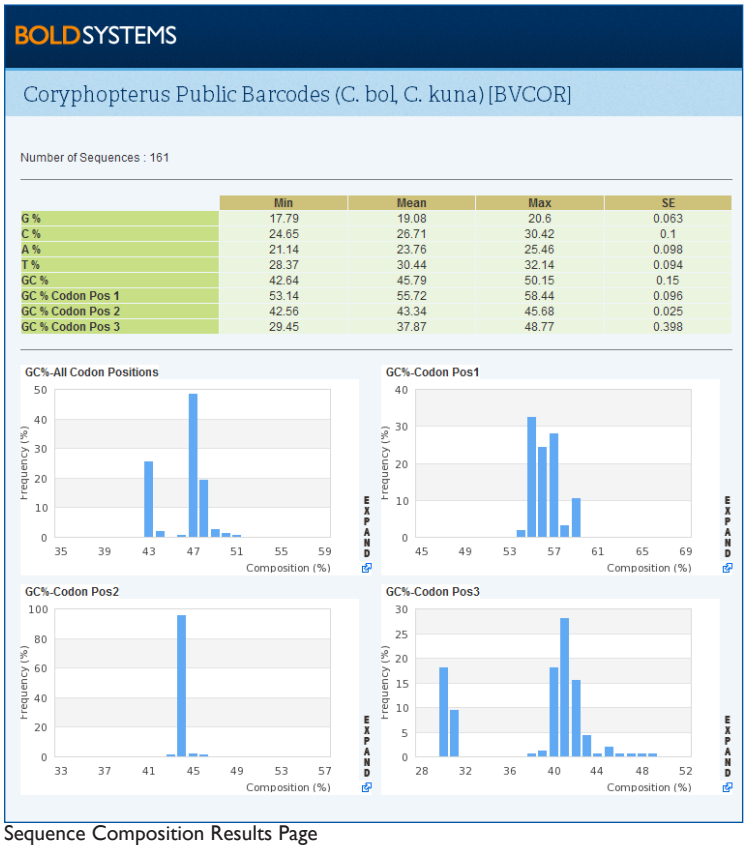


Distance Summary Page

Workbench: Sequence Composition

The frequency of DNA bases, observed with emphasis on GC-content, can be a useful metric for evolutionary biologists. GC-content within the barcoding region of CO1 has been correlated with GC-content of the entire mitochondrial genome for many species.

Using the Sequence Composition tool allows the user to view the frequency of each base, G, C, A and T, as well as graphics for GC content on all codon positions. This information includes overall sequence composition, as well as for codon positions 1, 2, and 3. "Detailed View" tabulates the results for each specimen.



Sequence Composition Results Page

Workbench: Barcode Gap Analysis

The Barcode Gap analysis presents users with an examination of the distance to the nearest neighbour for each of the species in the list of specimens. Distances are highlighted if the nearest neighbour is less than 2% divergent, or when the distance is less than the intra-specific distance.

Barcode Gap Analysis Page

Tip: Graphs for Publication

When the "Expand" icon (shown to the right) appears next to a graph, the graph is expandable for a better quality version that can be used in publications.

EXPAND

Workbench: Accumulation Curve

An accumulation curve of standardized DNA barcodes and related features provides a clear, transparent and reproducible estimate of the diversity and sampling efficiency of areas or collections.

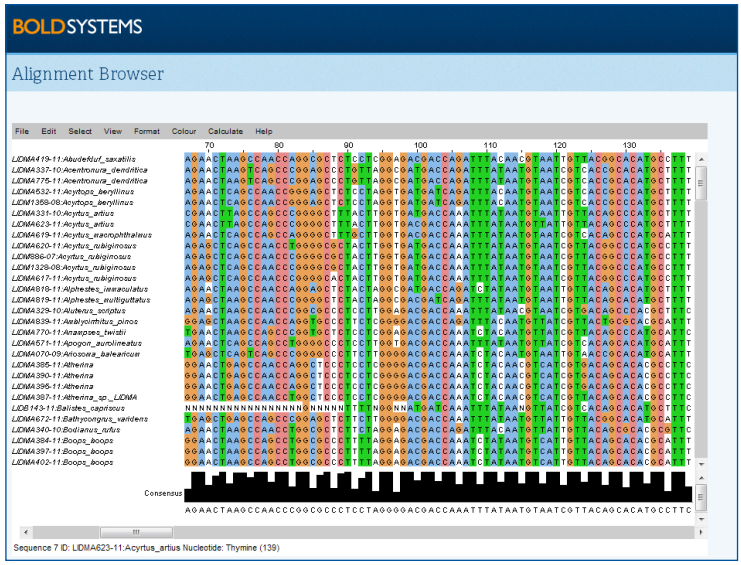
This tool also allows users to quickly compare sampling efficiency at multiple regions by multiple taxonomic levels.

Accumulation Curves Page

Workbench: Alignment Browser

Managing sequence alignments and base calls are a critical step in any barcode analysis. To prevent the inconvenience of importing sequences into 3rd party software to analyze and often edit, BOLD provides an integrated alignment browser that will include many features popular in other packages. In BOLD 3.0, the updated alignment browser can handle thousands of sequences and will soon support direct editing to the database.

Multiple alignment options such as Muscle and Kalign algorithms as well as colourization options are now available.

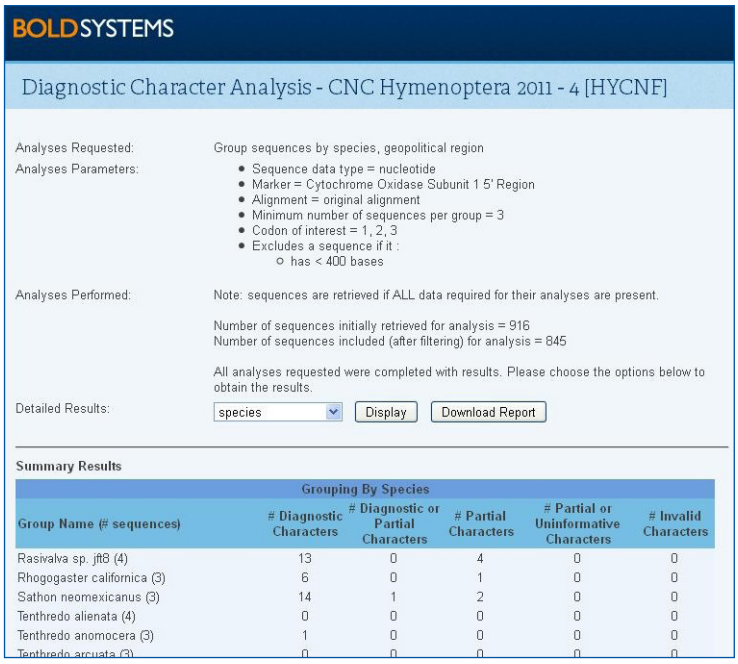


Alignment Browser Page

Workbench: Diagnostic Characters

The Diagnostic Character analysis provides a means to examine nucleotide or amino acid polymorphism between sets of sequences that are grouped by taxonomic or geographic labels. More specifically, this tool identifies consensus bases from each group, compares them to those from the remaining sequences in other groups, then characterizes how unique each consensus base is. The tool categorizes consensus bases by their diagnostic potential.

Since this tool only performs the analysis on the set of sequences selected by the user, the result is greatly affected by the initial data and the analysis parameters. Even the smallest change in the initial sequences, filtering options or the analysis parameters can affect the consensus sequences in each group and hence the diagnostic potential will be different between analyses. As a result, the interpretation of each analysis is absolutely dependent on all the factors combined. In general, having more sequences per group will provide a more accurate diagnosis of each group, as it reduces the problem caused by small sample size.

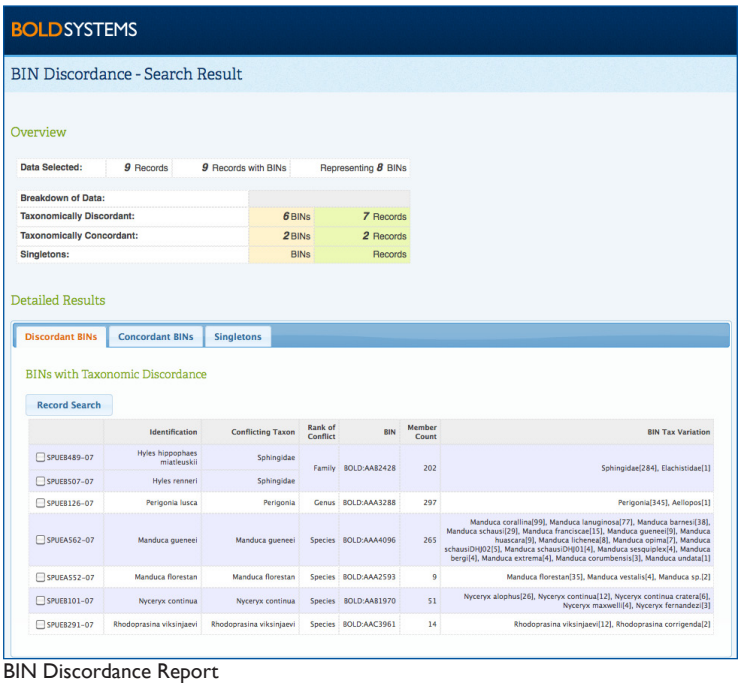


Diagnostic Characters Page

Workbench: BIN Discordance Report

The Barcode Index Number module analyzes new COI sequences and assigns them to an existing or a new BIN. Besides generating BIN pages, this system acts as a rapid check of the validity of taxonomic designation on specimen records. The BIN Discordance report facilitates this check by comparing the taxonomy on selected records against all others in the BINs they are associated with.

The results are sorted by the degree of conflict, displaying those records in BINs where there is a phylum level conflict first (likely the result of cross-contamination) down to species level conflicts. Users can pull up records from this page to examine ancillary data or to edit the taxonomy where there is an error.



BIN Discordance Report

Notes

Last modified: September 2012

BOLDSYSTEMS

For online version, please visit: www.boldsystems.org/docs/
For support with any feature of BOLD, please email: support@boldsystems.org

Barcode of Life Data Systems
Biodiversity Institute of Ontario
University of Guelph
50 Stone Road East
Guelph, Ontario, Canada
N1G 2W1

Copyright ©2012 Biodiversity Institute of Ontario

Authored by: Megan Milton (mmilton@boldsystems.org)
Mallory Van Wyngaarden (mvanwyng@boldsystems.org)
Sujeevan Ratnasingham (sratnasi@boldsystems.org)

Designed by: Suzanne Bateson, Biodiversity Institute of Ontario