BOLDSYSTEMS

Student Data Port@l

www.boldsystems.org/edu

BOLD-SDP: Features and Functionality
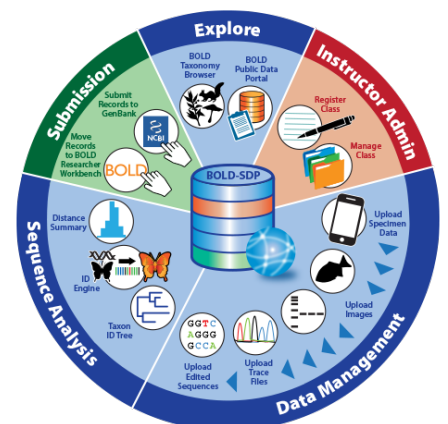
Table of Contents

System Overview

BOLD Student Data Portal (BOLD-SDP) is an educational module that supports the assembly and analysis of DNA barcode data by students in a classroom environment. It provides a web-based software framework for instructors and students to utilize the same tools used by professional researchers and make real contributions of complex data records to the scientific community. The module is built on top of the world class Barcode of Life Datasystem (BOLD), the primary data storage and analytics platform for the DNA barcode research community, and provides customized interfaces to BOLD analytical tools and visualizations.

BOLD-SDP promotes a partnership between educators and members of the scientific community, by providing a space where both communities can intersect. Students are exposed to valuable aspects of cutting edge biodiversity science and also given the opportunity to contribute valuable data. Researchers can utilize the data generated by students and can be reassured of high data quality. BOLD provides mechanisms to extensively validate student-generated data prior to deposition in the primary database ensuring that the scientific integrity of the project is maintained.

The system consists of three consoles: the Student Console, Instructor Console, and Administration Console, each providing custom tools to students, instructors, and scientists working together.

SDP Homepage

SDP Platform: Student Console

The student console offers an education-oriented suite of data management and analysis tools that enables students to perform activities similar to those performed by professional scientists. Interfaces contain integrated support documentation, written specifically for students, and provide step-by-step instructions. Assembly of single barcode records is broken down into four discrete steps allowing students to fully understand the data elements. Completed barcode records are vetted by instructors and passed to administrators for further review prior to publication in BOLD and the International Nucleotide Sequence Database (INSDC).

The student console on BOLD-SDP also encourages team based work while providing detailed tracking of the steps completed by each student. The entry of each data element is accompanied by student attribution information. This data is used on the published records which shows students as authors and also provides instructors with real-time feedback for marking or assessment of engagement.
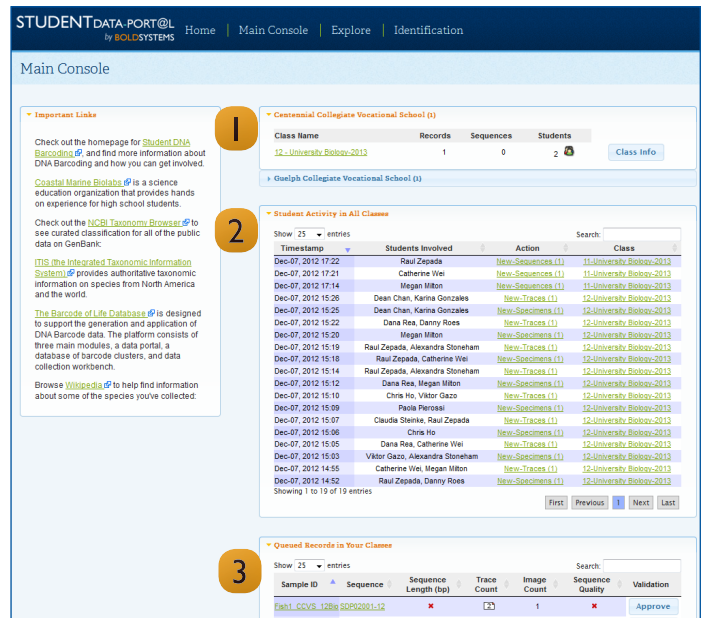

SDP Console for Students

SDP Platform: Instructor Console

The instructor platform is a powerful tool to monitor and assess student work. Instructors are also the first tier in the data validation process, and are required to review and approve all records submitted by the students. Real-time activity reports and classroom snapshot reports keep instructors updated on the work being done and gives them the opportunity to engage students falling behind.

This unique interface is designed to provide instructors with easy oversight of multiple classes/courses and allows a single instructor to span multiple schools. Following the collaborative philosophy of the BOLD platform, instructors can work together, sharing oversight of classes/courses. The instructor console provides three distinct views for data access and interaction.

SDP Console for Instructors

1. Course Overview	The course overview provides information for all courses that an instructor is involved in, including student names, student log in information, and the number of specimens, images, trace files, and sequences each student has uploaded
2. Student Activity	The student activity wall provides sortable real-time information about student contributions from all courses including student names, the specific contribution (including the record ID), and the course name
3. Records queued for approval	The record approval queue lists and summarizes records from all applicable courses, allowing instructors to quickly and easily approve completed high-quality records

SDP Platform: Administration Console

The Administration Console is designed for professional scientists supporting classes/courses. Through the console they are able to quickly review student generated data but only after it has been vetted by instructors. This crucial step in the validation process prevents contamination of the BOLD database by low quality records.

This console is similar to the instructor console but accesses a broader range of classes, allowing multiple instructors to be connected to a single administrator. Only through this console is publication of data into the BOLD and INSDC databases possible.

SDP Console for Administrators

Data Validation Workflow

To ensure that all student-generated records submitted to GenBank and the BOLD DNA barcode library are accurate and high quality, it is important to have data approved by both instructors and administrators. Using a three-tiered validation process, BOLD-SDP helps to maintain the integrity of the BOLD DNA barcode library by protecting the library from unverified data.

Instructors may be involved in several courses, so to simplify validation, records from all applicable courses are queued and summarized in one section on the Management Console. From here, instructors can evaluate complete barcode records as well as probe individual data elements. Further validation is supported using built-in analytical tools that look for data integrity issues (e.g. contamination, sample mix-ups, and specimen misidentification).

Following instructor approval, records are queued for approval by SDP administrators who perform additional validation considering the broader context of all the data on BOLD. The data management staff at BOLD perform a final analysis of the data records prior to publication of the data.

SDP Record Approval Form for Instructors and Administrators

Publication of Barcode Data

GenBank serves as an open access repository for sequence data that serves as the community standard for ensuring that data used in scientific research is available for others to analyze and re-use. Records submitted to GenBank are checked for originality and quality before they are issued a unique accession number which allows others to locate records for inclusion in their analyses.

Once a record has been approved by instructors, SDP administrators, and BOLD staff, it is automatically queued to be sent to GenBank following BOLD's standard GenBank submission workflow. Each student who has contributed to the record (by submitting specimen data, images, trace files, or sequences) will be listed as an author on the GenBank record, a field that is almost always occupied by professional scientists. This DNA barcode record becomes part of the DNA barcode library and may be used in future academic research by scientists worldwide.

1. GenBank header details	Identifications made by students that are verified are submitted to GenBank and receive accessions.
2. Author Names	Student names appear as authors in the public record
3. BOLD reference	GenBank and BOLD are linked together on records published through BOLD.

Sebastes chlorostictus voucher FT-69 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial

GenBank: JX487224.1

FASTA   Graphics

Go to: ☑

```
LOCUS       JX487224                657 bp    DNA     linear   VRT 16-SEP-2012
DEFINITION  Sebastes chlorostictus voucher FT-69 cytochrome oxidase subunit 1
            (COI) gene, partial cds; mitochondrial.
ACCESSION   JX487224
VERSION     JX487224.1  GI:404248246
KEYWORDS    BARCODE.
SOURCE      mitochondrion Sebastes chlorostictus (greenspotted rockfish)
  ORGANISM  Sebastes chlorostictus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Actinopterygii; Neopterygii; Teleostei; Euteleostei; Neoteleostei;
            Acanthomorpha; Acanthopterygii; Percomorpha; Scorpaeniformes;
            Scorpaenoidei; Sebastidae; Sebastinae; Sebastes.
REFERENCE   1  (bases 1 to 657)
  AUTHORS   Greenspan,P., Lewis,B., Tweed,S., Santschi,L. and Imondi,R.
  TITLE     Barcoding the Kelp Forests of California's Northern Channel Islands
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 657)
  AUTHORS   Greenspan,P., Lewis,B., Tweed,S., Santschi,L. and Imondi,R.
  TITLE     Direct Submission
  JOURNAL   Submitted (14-AUG-2012) Integrative Biosciences Program, Coastal
            Marine Biolabs, 1559 Spinnaker Drive, Suite 101, Ventura, CA
            93001-5302, USA
FEATURES             Location/Qualifiers
     source          1..657
                     /organism="Sebastes chlorostictus"
                     /organelle="mitochondrion"
                     /mol_type="genomic DNA"
                     /specimen_voucher="FT-69"
                     /db_xref="BOLD:FXIII055-12.COI-5P"
                     /db_xref="taxon:72064"
                     /country="USA: California, Ventura County, Santa Barbara
                     Channel"
                     /lat_lon="34.2081 N 119.512 W"
                     /collection_date="03-Feb-2012"
                     /identified_by="Milton Love"
                     /PCR_primers="fwd_seq:
                     tgtaaaacgacggccagtcaaccaaccacaaagacattggcac, rev_seq:
                     caggaaacagctatgacacttcagggtgaccgaagaatcagaa"
                     /PCR_primers="fwd_seq:
```

GenBank flatfile for a barcode record from BOLD

Exploring Scientific Data

By providing links to the BOLD Taxonomy Browser, BOLD Public Data Portal, and international taxonomy databases, BOLD-SDP allows students to learn the basic concepts of species classification and description. The tools on the Explore page will help students envision the end result of their personal contributions to the BOLD DNA barcode library. Students can explore research questions based on data previously collected by professional scientists, helping to emphasize the value of existing data and allowing them to explore analytical methodologies.

The BOLD Taxonomy Browser provides a single page for each species and for higher level taxonomic names (e.g. the class Mammalia for all mammals). This includes the number of specimens and sequences barcoded, where the specimens are stored (for example, Coastal Marine Biolabs), where they were collected, and images. Students can explore the taxonomic hierarchy for an organism and learn about the progress of DNA barcoding for a given group. Using this tool, students and instructors will be able to learn about the distribution of certain species and the progress of barcoding in different groups. Students are exposed to the taxonomic hierarchy and are given access to large volumes of biodiversity data.

1. Lineage	Displays the taxon name and the higher taxonomic levels.
2. Sub-Taxonomy	Links to all sub-taxa with the 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 downloaded for each rank that has sub-taxa. The published and released sequences for this taxon can be downloaded from this section.
5. Sample Sources	A graph of the top institutions that provided specimens with their specimen tallies.
6. Sequencing Labs	A graph of the top 10 labs that generated sequences.
7. Imagery	A single representative image is displayed for each sub-taxon. Mousing over an image selects it for higher-resolution display to the right.
8. Image Details	The taxonomic identifier, the sample identifier, license and attribution are displayed beneath the image that is selected.
9. Collection Sites	A map of the collection sites including a list of the top countries
10. 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 for the Genus *Semicossyphus*

Course Support

Documentation

The BOLD-SDP Support team has a variety of materials available to help system users get the most from BOLD-SDP. A Quick Start guide and FAQs are available on the website, and specific instructions are displayed alongside each page for both instructors and students.

Training videos, webinars, and lesson plans (including assignments and quizzes) are being added to the site as they are developed to help instructors plan lessons and projects for their students.

Detailed support for BOLD-SDP is available online the site. Users of the system can contact our support team via email through

edu@boldsystems.org

Staff Support

The BOLD-SDP support team currently consists of a main help desk contact, as well as several DNA Barcoding experts who work on validation and curation of the student generated data.

Alexandra Stoneham - Help Desk

Alexandra has been a central user support member for BOLD software for over 6 years. During her undergrad in biological science, she experienced first hand how DNA barcoding was taught in the classroom, which gives her a perspective of how students are engaging in this technology. She has been key in assisting educational users understand how to submit data and leverage the educational site at BOLD for several years.

Alexandra

Data Management: Overview

There are four data components that form a complete barcode record:
1. Specimen Data
2. Specimen Images
3. Trace Files
4. Sequence

Each of these components should be uploaded separately and in order, starting with the specimen data. Student attribution for the work should be submitted at each step.

Data Submission Console

New Specimen

Upload Images

Upload Traces

Add Sequence
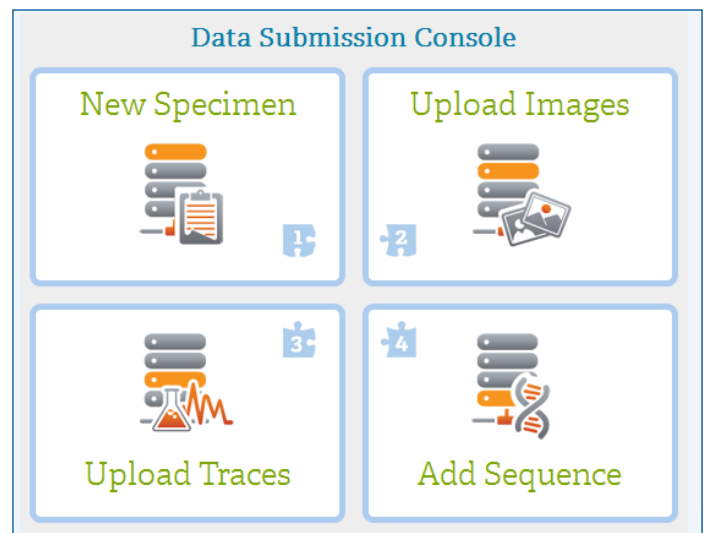
Student Data Submission Console

Data Management: Specimen Data

The Specimen page is accessed through the Main Student Console. There are 4 blocks of data that need to be completed: the Specimen Identifier, Specimen Details, Taxonomy, and Collection Details. This form is a simplified version of the data that is uploaded by researchers to BOLD to help students get their specimen data online quickly and efficiently. Instructions on the correct way to fill out the various fields are given in the menu on the left side of the screen.

| 1. Instructions	Detailed directions are provided to students on the specimen upload page explaining the requirements of and purpose for each section of the Specimen Data record.
2. Specimen Identifier	A unique Sample ID is required for all records uploaded to BOLD-SDP to maintain data integrity and help students develop good data management habits.
3. Taxonomy	The taxonomy drop-down menus use the existing taxonomic hierarchy on the BOLD database. Students sequentially select the taxonomy of their specimen from Phylum to Species, ensuring that no errors are made.
4. Collection Details	Students have the option of adding detailed collection information to ensure that a collection site can be revisited or included in analysis.

SDP Specimen Data Upload Form

Data Management: Images

The Upload Images page is accessed through the Main Student Console. By entering the correct sample ID, students will connect the uploaded image with the corresponding specimen. Instructions on the correct way to fill out the various fields are given in the menu on the left side of the screen. Specimen images will help students, instructors, and experts confirm the identification of samples even when the actual specimen isn't available. Students can upload one image at a time to their BOLD records, providing attribution information for each image. BOLD can support multiple images per specimen if different orientations are useful.

SDP Specimen Image Upload Form

BOLDSYSTEMS-SDP

Data Management: Trace Files

The Upload Traces page is accessed through the Main Student Console. By entering the correct sample ID, students will connect the trace files with the corresponding specimen. Trace files provide access to the raw data that can be used by instructors, administrators and other data users to verify sequences. Students have the ability to select PCR and sequencing primers and to upload a forward and a reverse trace file.
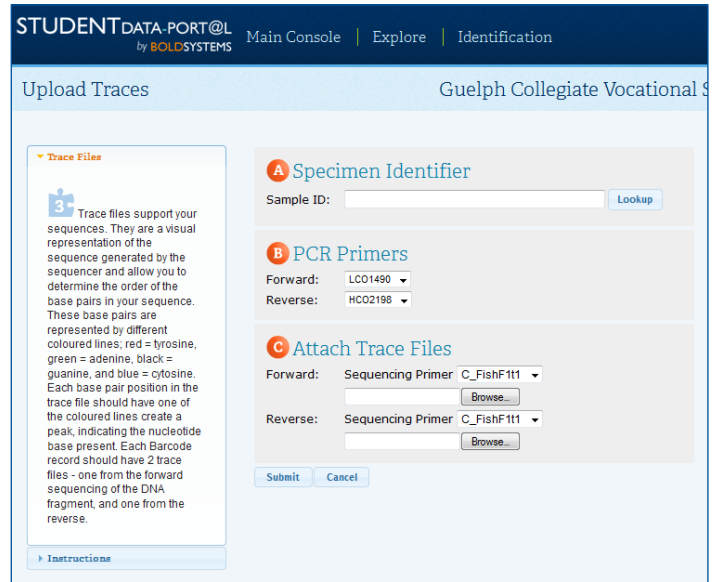
SDP Specimen Trace Upload Form

Data Management: Sequence

The Add Sequence page is accessed thought the Main Student Console. Each step in sequence uploading is sequential, and will only be available once the preceding step has been completed. There are two options for uploading a sequence: use the BOLD-SDP Sequence Editor to trim, assemble, and edit the uploaded trace files in order to generate a sequence, or upload a sequence created in third-party software.

### BOLD-SDP Online Sequence Editor

The BOLD Sequence Editor simplifies the generation of sequences from submitted trace files.

It uses an algorithm to assign base calls (nucleotides) to each peak in the trace file and to compute a confidence or quality (Q) score for each base call. The quality score represents the level of confidence that a base call was made correctly. A quality score of 40 indicates that the probability of an incorrect base call is 1 in 10,000.

Bringing two trace files into register and displaying them in the same window enables a researcher to identify regions of agreement or disagreement in base calls. In cases where a low quality base call is found in one trace file, the researcher can find the position of the base call in the other trace file and compare the differences in quality scores. If a higher quality score value (>20) is assigned to the base call in the second trace file, then that base call is regarded as the correct nucleotide and accepted.

The Sequence Editor compares the quality scores of base calls at every position of the trace files and accepts the base call with the higher quality score for inclusion in the contig.

The assembled sequence can then be saved into BOLD to be processed.
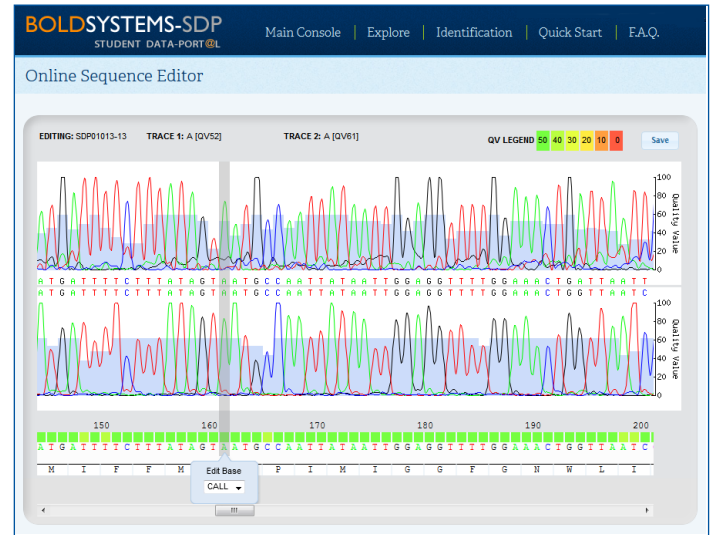
SDP Sequence Editor

Data Management: Sequence

Post-Processing of Sequences

Once the sequence has been pasted in from 3rd party software or saved from the sequence editor, it is ready to be processed.

This will involve three steps that will be performed automatically by BOLD:
1. Primer trimming
2. Contamination checking
3. Sequence alignment

If all steps are completed successfully the sequence is ready to be submitted to BOLD-SDP. Submitted sequences are sent to instructors and administrators to be validated to ensure they meet BOLD standards.

SDP Sequence Upload Form

Data Submission: Attribution

BOLD-SDP features a unique system of attributions that allows students to receive acknowledgement for the specific data elements they contribute – specimen data, images, trace files, or sequences. In addition to encouraging collaboration and data sharing among students, this feature also provides instructors with a quantifiable means to measure student involvement and participation on this project.

Section of record highlighted to show individual attribution

Data Analysis: Morphology & Genetic Variation

Being able to integrate genetic, morphologic, and geographic information is key to helping students understand some of the causes of genetic divergence and how these differences can manifest themselves in living species' and populations. BOLD-SDP tools can help students compare patterns of genetic and morphological diversity to visualize what genetic divergence can mean in real specimens. Using Taxon ID Trees and the Image Library in conjunction can help students understand how DNA barcodes can be used to discover new species. These features are also helpful in allowing students and instructors to confirm congruence between images and specimens.

Taxon ID Tree

The Taxon ID Tree allows students to visualize how the sequences are genetically related using a downloadable Neighbour-Joining tree. Students can add geographic location and higher level taxonomy to the tree labels to help them understand correlations between geographic divergence, physical traits, and genetic relatedness. This tool allows students to see that taxonomy is not always an exact science and that two specimens with different species names may be genetically closely related.

Image Library

The Specimen Image Library allows students to view all of the images uploaded to a course in a single page. Students can filter the images by specimen orientation (e.g. Dorsal, Lateral) to view all of the images of a particular orientation at once. This tool helps assist in record validation by allowing students and instructors to quickly verify that the images are uploaded to the correct specimen. An image library can also be generated when a Taxon ID Tree is made, listing the images in the order of the records on the tree. This makes it easy for students to visualize how related specimens group together and how similar morphologies don't always mean specimens are of the same species.

Using the Taxon ID Tree and Image Library together may help students understand how and why taxonomy can change. For example, if a student notices two specimens with different species names clustered together on a Taxon ID Tree, he or she can look at the corresponding images in the Image Library. They may notice that these two specimens are almost identical morphologically which could indicate that these two species names are in fact synonyms, starting them on an investigation of the name, or they may notice that the specimens are misidentified, leading them to revisit their field guide or consult an expert.
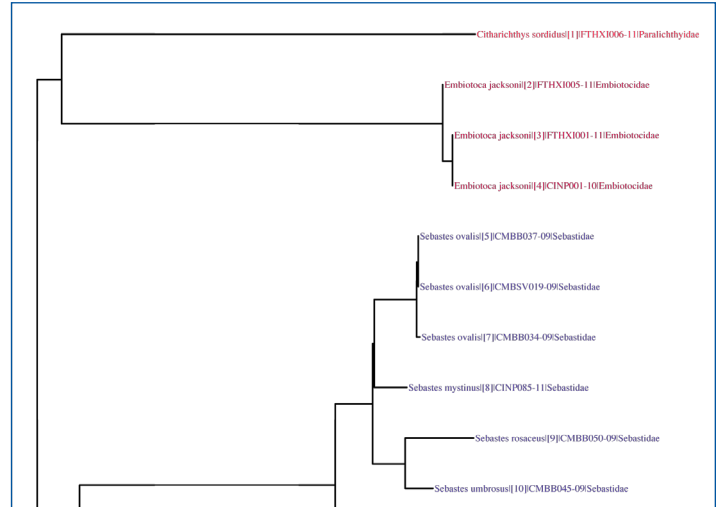
Standard Taxon ID Tree

Image Library

Data Analysis: Ecological Relationships

BOLD-SDP allows students to infer ecological relationships from their DNA barcoding data by comparing sampling efforts and progress from different collection sites. BOLD tools like the Distribution Map and Taxon Accumulation Curves allow students and instructors to evaluate their own sampling efforts during their personal field collections and also to see how DNA barcoding can be affected by different collection sites and protocols. These tools can highlight areas that may need more sampling or are simply hard to sample – areas that may have lower biodiversity, be home to animals that are difficult to collect, or just be difficult to get to like a deep region in a coastal bay or steep rocky terrain. They also provide the opportunity to answer a wide range of ecological and evolutionary questions, including comparing local and global species distributions, investigating species migration patterns, and measuring habitat-dependent diversity.

Distribution Maps

The BOLD-SDP Distribution Map feature plots collection points for all specimens based on GPS coordinates submitted with the record. This tool helps highlight over-sampled and under-sampled areas, patterns in sampling distributions, and correlations between sampling efforts and habitat. Using the built-in map feature as well as Google Earth, students can zoom to their collection points to see accurate details of the collection sites. Students can use this tool to learn about the geographic ranges of species, including areas of high concentrations of specimens.

Taxon Accumulation Curve

The Accumulation Curve tool on BOLD-SDP compares the number of specimens collected to the number of taxa collected (species, genera, families, etc.) in order to determine the diversity and sampling efficiency of an area or collection. These curves can be grouped by taxa or geographic region. This tool can be used by students and instructors to evaluate the completeness of sampling effort when taxonomic information is limited. Steeper curves indicate that newly collected specimens are still generating new species, while flatter curves may indicate that most of the diversity has already been explored.

When used together, the Distribution Map and the Taxon Accumulation Curve can be useful tools for helping students understand how species ranges and population densities can affect sampling. When sampling fish species near two islands, one with greater biodiversity than the other, they will see two very different curves. One curve will level off quickly, indicating that collecting more samples did not increase the number of species collected. The other curve will take longer to level off, meaning that continuing to collect samples increased the number of species that were collected. By comparing these curves to the Distribution Map of the samples they've collected, students will be able to compare the two sampling habitats and think about the differences that might make one habitat more suited to higher biodiversity than the other.
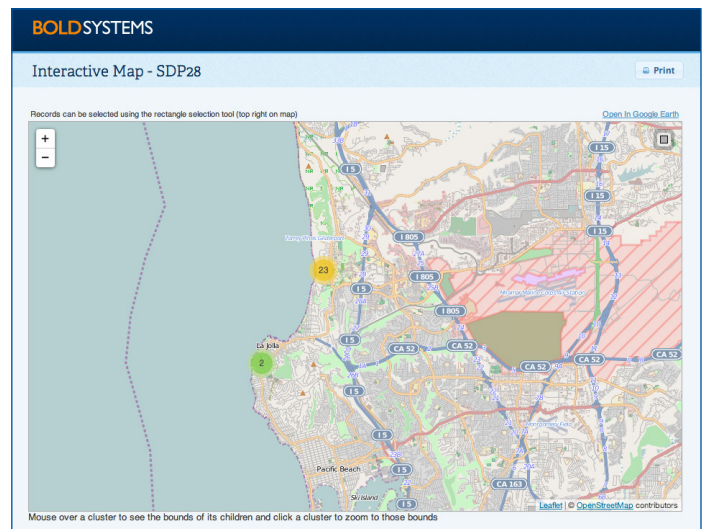
Distribution Map

Accumulation Curves Page
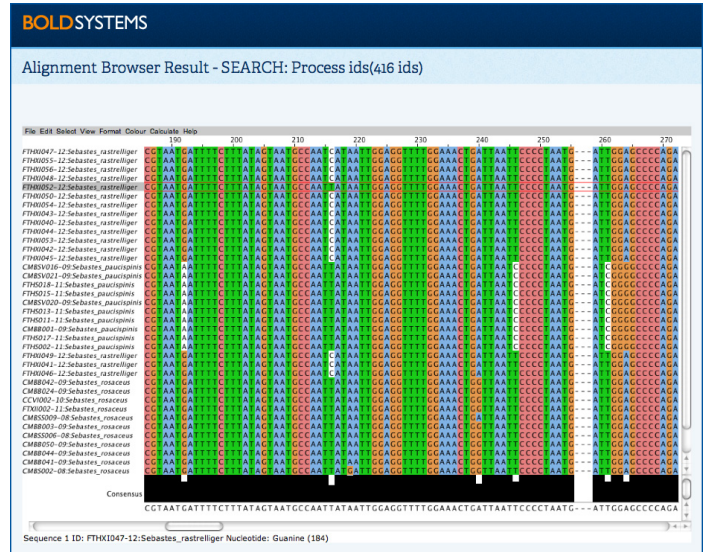
Data Analysis: Validation of Sequences

The Alignment Browser makes it easy to visualize and compare the sequences of multiple specimens at once. Students can compare and contrast sequences of all the specimens in their class/course to see the genetic difference between closely and distantly related specimens. The Trace Viewer displays raw data (the trace file chromatogram) allowing students to see a visual representation of a barcode sequence. Used together, these tools both help to connect the visual sequence to the series of bases that make up a DNA barcode. Students can easily compare the coloured peaks in the trace files to the corresponding bases listed in the Alignment Browser, helping them to understand how a trace file is read to generate a sequence.

Alignment Browser

The Alignment Browser feature on BOLD-SDP makes it easy to visualize and compare the sequences of multiple specimens at once. It allows students to colourize sequences by individual bases so they can be easily identified and distinguished from one another, or by amino acids for easy detection of stop codons and shifts to the reading frame. By combining it with the Trace Viewer, students will be able to recognize how trace files complement each other to form a sequence. It can handle multiple sequences and will soon support direct editing to the database.

Alignment Browser Page

Trace Viewer

The Trace Viewer on BOLD-SDP displays the graphical representation of a trace file generated during sequencing that is used to determine base calls for the sequence. It allows students and instructors to examine the raw files used in sequence assembling and to verify that the arrangement of the nucleotides is correct in order to ensure that a credible DNA barcode is being created.
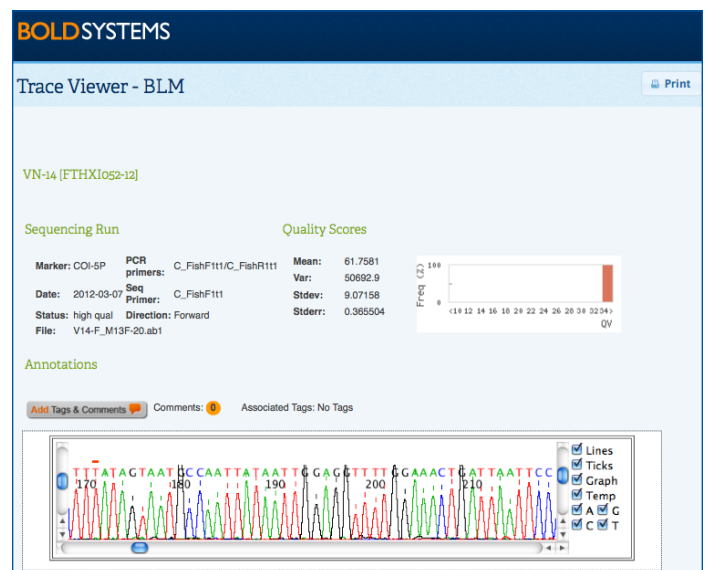
Trace Viewer Page

Data Analysis: Barcode Exploration

The Barcode Gap Analysis tool and Identification Engine can help students understand some of the underlying principles behind DNA barcoding. Both tools help highlight the quantitative differences between sequences – the Barcode Gap tool shows the distance to the most similar sequences of a different species, while the Identification Engine compares a sequence to the entire BOLD barcode library and lists the closest sequence matches, including the percent similarity. These tools can help students see that closely related species and specimens have sequences that are more similar than distantly related species and specimens using actual data and real numbers. Exceptions of this pattern are exposed for further investigation.

Barcode Gap Analysis

The Barcode Gap Analysis tool on BOLD-SDP provides an evaluation of the range in sequence divergences within a species and between species for all the records in a course. The ratio of these values allows students to highlight problematic cases where the divergence between species is higher than the divergence within. The Barcode Gap Analysis tool describes the distribution of sequence divergences using frequentist statistics, creating an opportunity for the application of statistical theory with real data.
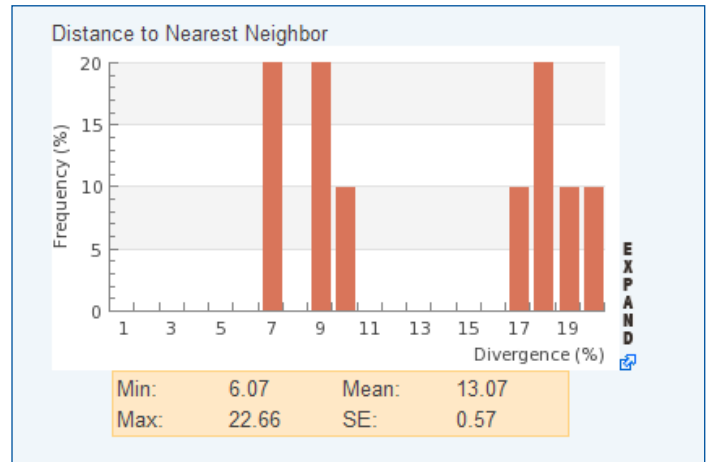
Highlighted graph from Barcode Gap Report Page

ID Engine

BOLD-SDP offers students access to the complete DNA barcode reference library through the Identification Engine tool. Student can compare their sequences against all sequences uploaded to BOLD from public and private projects to locate the closest match. Students are able to combine the analysis provided by the barcode gap and the closest species match to statistically verify their species identifications.

Identification Engine Results Page, showing closest matches and occurrence map

BOLDSYSTEMS STUDENT DATA PORT@L
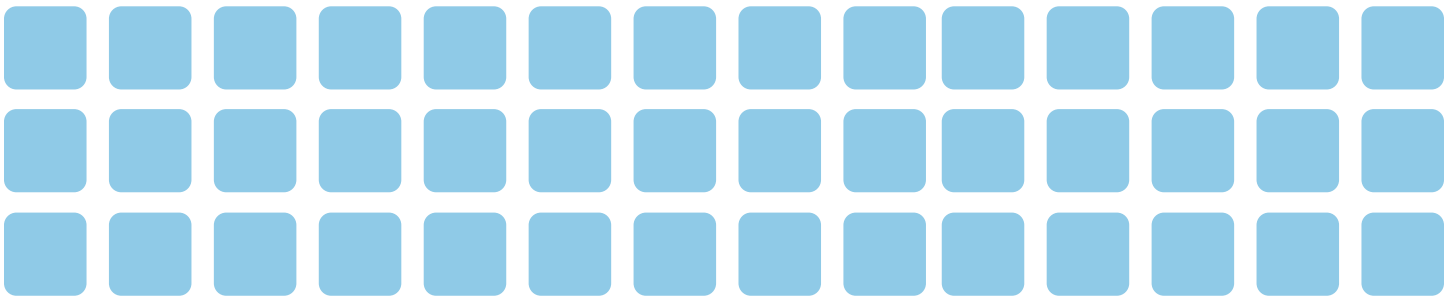
Ontario
MINISTRY OF
RESEARCH AND INNOVATION

NSF

international
BARCODE
OF LIFE

BOLD Systems is hosted by the
Biodiversity Institute of Ontario
at the University of Guelph