

DNA Barcoding Kit

Catalog #12016300EDU, 12016353EDU, 17007432EDU, and 17007366EDU

Bioinformatics Guide

Note: This document is for planning purposes only and may vary from the final product specifications.

Note: Duplication of any part of this document is permitted for noncommercial, educational use only. Commercial use of copyrighted content, including use in instructional materials for which payment is received, requires express permission from Bio-Rad Laboratories, Inc. Contact us at explorer@bio-rad.com for more information

For technical support, call your local Bio-Rad office, or in the U.S. call **1-800-4BIORAD** (1-800-424-6723) option 2.

Table of Contents

Instructor's Advance Preparation	1
Timeline for Required Advance Preparation	3
Instructor Required Advance Preparation Steps.....	4
Step 1: Registering an account.....	4
Step 2: Creating specimen folders into which student groups will upload data.....	10
Step 3: Uploading forward and reverse trace files into your account on BOLD-SDP.....	17
Step 4: Assessing quality values for student sequencing traces to determine bioinformatics workflow.....	21
Student Instruction Manual	22
DNA Sequencing of Amplicons Using Dye Terminator Cycle Sequencing.....	22
Sequence Assembly and Editing in BOLD-SDP	25
Step 1: Assessing the quality of DNA sequencing data	26
Step 2: Querying either the BOLD database or GenBank for matches to forward and reverse sequencing data.....	34
Step 3: Assembling barcode contigs from trace files.....	43
Step 4: Inspecting contigs for the presence of stop codons, trimming primers, and checking for contaminants	49
Step 5: Reviewing the sample record.....	52
Quick Guide	58

Instructor's Advance Preparation

Before the Barcode of Life Data Systems-Student Data Portal (BOLD-SDP) can be used to analyze student data, you must register for an account. The steps required to do so are outlined below. There is also a video tutorial on the BOLD-SDP website that demonstrates these steps. Please be aware that you need to register with the BOLD-SDP website 2–3 days before students run through the bioinformatics lessons.

The procedure outlined below contains steps the instructor must complete before the students can enter the BOLD-SDP database. Other steps, such as creating specimen folders and uploading sequencing trace files, can be done by either the students or the instructor. Where there are wait times for calculations the BOLD-SDP software needs to perform, the amount of time is listed in the instructions to help with classroom time planning. Ultimately, the instructor can choose which steps are performed by whom and cater the course to teaching goals.

Setting Expectations



The BOLD-SDP portal was designed to serve as a workbench for those interested in generating data of sufficiently high-quality to be considered for publication in the BOLD reference barcode database. The data contained within the BOLD database come from samples that were vouchered and have sequencing data of extremely high-quality. In order to be able to contribute sequencing data to this database for use by other researchers, a very stringent process must be obeyed. Because of this, certain sequences students generate may not meet the requirements for assessment by the BOLD-SDP software and may receive trace file quality assessments of “low” or “fail”. This does not preclude students from learning bioinformatics or analyzing the data they generate, however, and students should not be discouraged! If students receive sequencing data quality assessments that are not high enough to allow them to complete the full workflow within the BOLD-SDP portal, they can still perform a BLAST (basic local alignment search tool) search using the GenBank database (instructions included in the first section of the student instruction manual) and find their level of homology to samples in that database.

GenBank is a commonly used and well known database and does not have the same stringent requirements as BOLD for submission of sample sequences to its database. Hence sample identification based on sequence matches with GenBank sequences may not be as trustworthy. For example, sequences for known samples have mistakenly corresponded to sequences from marine bacteria that had contaminated the original samples, and samples have been incorrectly identified in the first place. These errors are not just made by students, but also sometimes by researchers! This is an important teaching point on handling samples and looking at database data. Students can still work through a homology search and see if their sequences might match any in a sequencing database and at what level. An 86% match? A 99% match? How comfortable does your student feel with guaranteeing an identification of a species if it has only 86% homology to a species in the database? The more experience students get in looking at their data against data in different databases, the more they will improve their critical thinking skills. It is not merely a matter of “my sequence matches this database record by $x\%$ and therefore must be y species.”

Important! Barcoding Fungi Using BOLD-SDP



The BOLD-SDP does not currently accept barcode sequences from fungi. Several steps in the workflow on BOLD-SDP do not provide the correct options for fungal specimens. However, it is still possible for students to follow a modified version of the protocol to acquire a sequence that can then be identified using the BOLD identification tools or a BLAST search.

Follow the steps in this guide with the following modifications:

1. Create a specimen for the fungal sample as described in step 2 of instructor guide. The specimen can be created in a course for animal samples.
2. When uploading traces as described in step 3 of the instructor guide, select any PCR primers and any sequencing primers.
3. In step 2.f of the student guide, choose the Fungal Identification tab instead of the Animal Identification tab.
4. After step 3.i of the student guide, when students are viewing their contig sequence, click **Save** at the top of the sequence editor window. The contig sequence will now be shown in Box B. Select the sequence and copy it.
5. Repeat steps 2.d through 2.j of the student guide, again choosing the Fungal Identification tab.
6. Skip steps 4 and 5 of the student guide entirely. **Note:** the ITS region, which is used to barcode fungi, may contain stop codons, since it is not a protein coding region. This is acceptable.

Timeline for Required Advance Preparation

Steps	Work time	Wait time
Create an Instructor Account on BOLD-SDP <ul style="list-style-type: none"> • Receive an email from BOLD with instructor username and password • Receive an email from BOLD with five registration keys 	5 min	<ul style="list-style-type: none"> • 1 hr • 1–2 days post registration
Register a new course with BOLD-SDP <ul style="list-style-type: none"> • Receive an email from BOLD with course username and password 	No time	<ul style="list-style-type: none"> • 1 hr
Create specimen folders into which each student group will upload sample data <ul style="list-style-type: none"> • Create specimen folders • Upload data into folders <ul style="list-style-type: none"> • Optional: Assess quality of the data and determine workflow students will follow 	Up to 3 hr	<ul style="list-style-type: none"> • 1 hr • 1 hr, plus up to 24 hr for quality score data to be calculated • 1 hr

Instructor Required Advance Preparation Steps

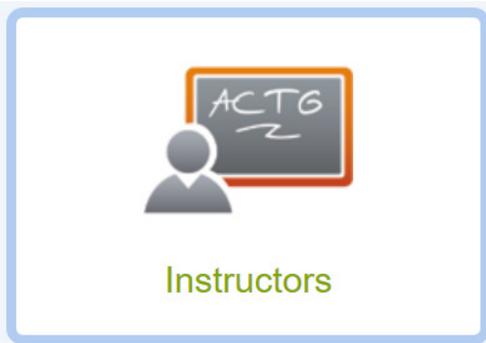
Step 1: Registering an account

- a. Access the homepage of BOLD-SDP at boldsystems.org/index.php/SDP_Home.

If, after reading these instructions, you require additional guidance to create a user account, note that there is a video tutorial that can be accessed by clicking the **Quick Start** tab or the **Quick Start Guide** on the BOLD-SDP homepage.

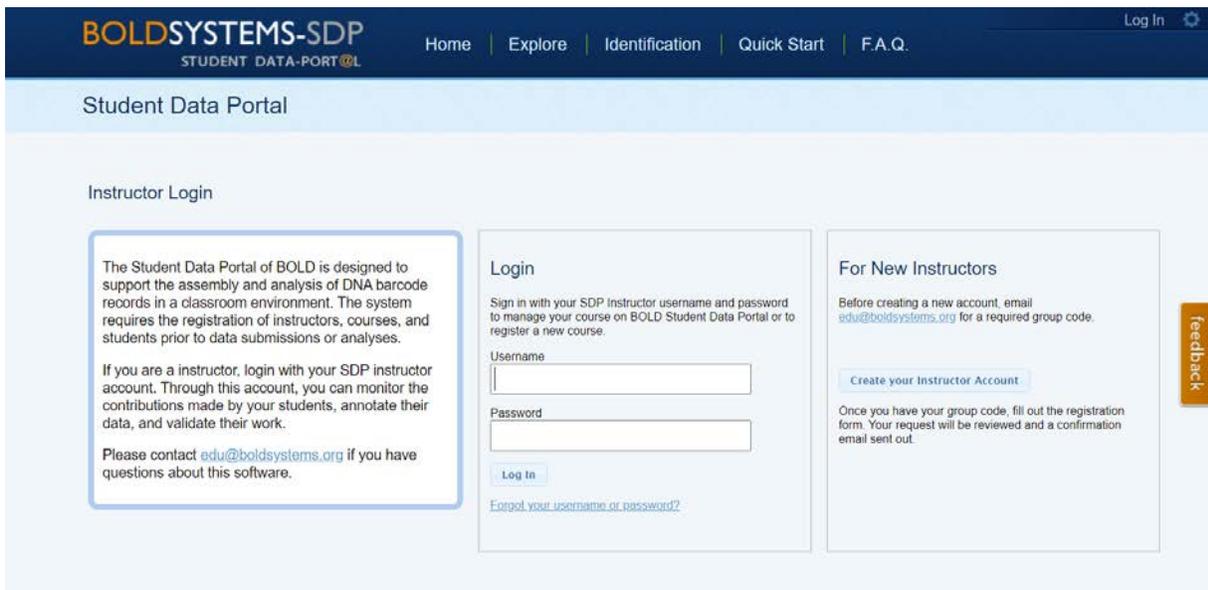


- b. Click the **Instructors** icon at the bottom of the homepage.

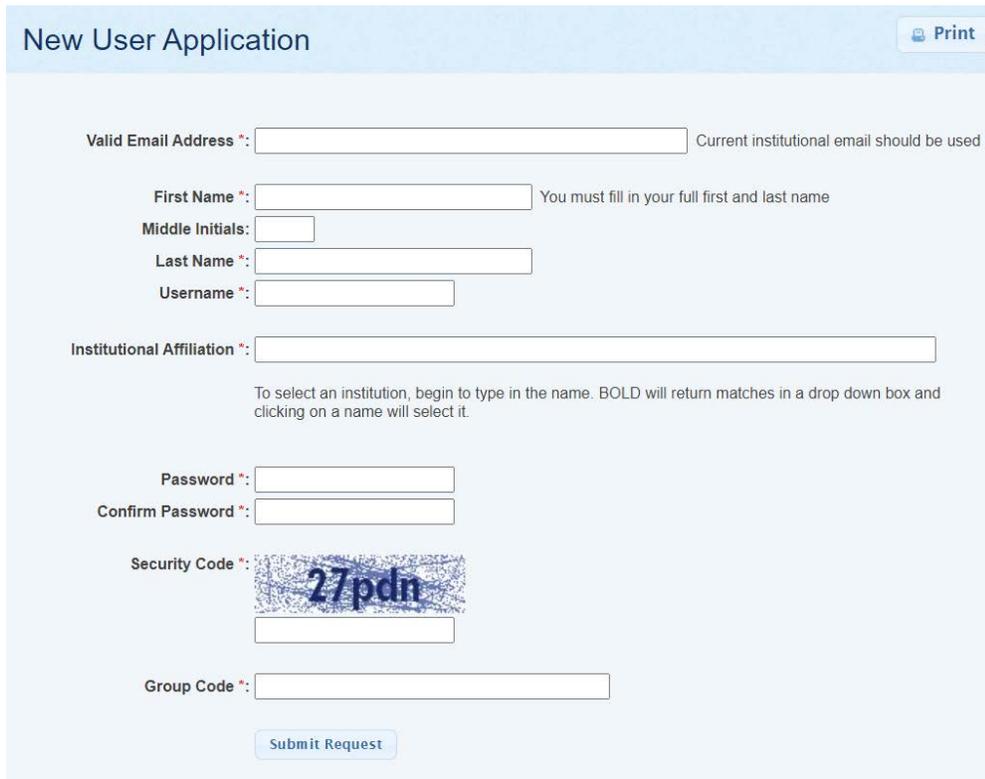


- c. If this is the first time you are using the BOLD-SDP website, you will need to create a new instructor account. Note that you can create an account without emailing BOLD Systems to request a group code because the required group code has already been obtained and is listed below in step 2.viii.

Click the **Create your Instructor Account** button.



This will open the New User Application page.



Instructor Preparation

Fill in all the information and click **Submit Request**.

- i. Enter your valid email address.
- ii. Enter your first name.
- iii. Enter your last name.
- iv. Enter your desired username.
- v. Enter your institutional affiliation. If your institution is not in the database, click **Add New Institution** and add the required information, then click **Submit Request**.

New Institution Application

Institution Name *:

Institution Code:

Address *:

City *:

Province/State *:

Country *:

Postal/Zip Code *:

Phone *:

Url:

Security Code *:

- vi. Enter a password and then enter it again in the Confirm Password box.
 - vii. Enter the security code as it is shown.
 - viii. Enter the group code **BOLD-EDU-SDP** and then click **Submit Request**.
- d. **1–2 Days** Within 1–2 days of submitting your request you will receive two separate emails from BOLD-SDP.
- Email 1 will say “Welcome to BOLD” in the subject line. This email will contain information on your **instructor username and password**
 - Email 2 will arrive later with “Welcome to BOLD-SDP!” in the subject line and will include five registration keys. Each registration key can be used for a separate class. If you need more keys, please email **edu@boldsystems.org**

- e. At this point you can register your class on BOLD-SDP. From the BOLD-SDP homepage, click the **Instructors** button.

BOLD SYSTEMS-SDP
STUDENT DATA-PORT@L

Home | Explore | Identification | Quick Start | F.A.Q.

An integrated workbench supporting the assembly, analysis and publication of DNA barcode data by students.

- System Overview
- Quick Start Guide
- Video Introduction

The BOLD Student Data Portal (BOLD-SDP) is a classroom-focused interface to the BOLD database. This platform provides instructors and students with the tools necessary to make contributions to the DNA barcode library used for identifying species. It also gives students the opportunity to integrate and analyze specimen and sequence data while providing instructors with tools to monitor student progress and evaluate their work. Students can explore the large database of DNA barcode records submitted by scientists around the world, but they will also be able to add their own data to the ever-growing DNA barcoding library. Contact us for support in using this system in a classroom setting (edu@boldsystems.org).

Students

A portal for students to assemble DNA barcode records in a simplified but powerful data collection and analysis environment.

Instructors

A management interface for instructors to register students and classrooms; review student work and utilize data validation workflows.

Explore

Search for and browse data records, images, and geographic distributions for species being barcoded.

- f. Log in with the **instructor username and password** that was emailed to you, then click **Log In**.

BOLD SYSTEMS-SDP
STUDENT DATA-PORT@L

Home | Explore | Identification | Quick Start | F.A.Q. Log In

Student Data Portal

Instructor Login

The Student Data Portal of BOLD is designed to support the assembly and analysis of DNA barcode records in a classroom environment. The system requires the registration of instructors, courses, and students prior to data submissions or analyses.

If you are an instructor, login with your SDP instructor account. Through this account, you can monitor the contributions made by your students, annotate their data, and validate their work.

Please contact edu@boldsystems.org if you have questions about this software.

Login

Sign in with your SDP instructor username and password to manage your course on BOLD Student Data Portal or to register a new course.

Username

Password

[Forgot your username or password?](#)

For New Instructors

Before creating a new account, email edu@boldsystems.org for a required group code.

Once you have your group code, fill out the registration form. Your request will be reviewed and a confirmation email sent out.

feedback

- g. You will now see a page where you can register your class information. Click **Register a New Course**.

Manage Courses

Register a New Course

Every course that you plan to include in the project must be registered independently. The registration process involves the entry of course information, including student names and their contact information (strict adherence to privacy policies is assured), and the listing of any instructors with whom you plan to collaborate.

Once registered, an instructor will be provided with account information for the course. Every student within that course will use the same account information (username and password) to access the site and to submit/analyze their data. The use of a single account simplifies the management of accounts for instructors, without compromising their ability to track contributions made by individual students.

Course Management

The management interface allows instructors to monitor the activity of individual students within a course, or to monitor multiple courses simultaneously. When combined with contribution reports for students, a live feed of student activity facilitates efficient tracking and enables instructors to provide early intervention and timely feedback for their students.

This interface also provides a convenient forum for instructors to review and approve student DNA barcode records before they are submitted to experts for additional review and publication.

Type in or copy and paste one of the five registration keys that was emailed to you.

Registration

Registration key *:

Instructor and Institution Information

Instructor:
[Add Co-instructor](#) +

Institution *:

County *:

City & Province/State *:

Course *:

Grade Level *:

School Year *:

Select Category *: Animal Plant

[Import from Excel sheet](#)

Students

Name: <input type="text"/>	Email: <input type="text"/>
Name: <input type="text"/>	Email: <input type="text"/>
Name: <input type="text"/>	Email: <input type="text"/>
Name: <input type="text"/>	Email: <input type="text"/>
Name: <input type="text"/>	Email: <input type="text"/>

- h. Your name should already be in the Instructor box. If you have a co-instructor, click **Add Co-Instructor**. (Co-instructors must also have their own instructor usernames and passwords in BOLD-SDP in order to access your account's data.) Please note that only the instructor can add more co-instructors. However, all other instructor privileges will be shared by anyone designated a co-instructor.
- i. Now enter your institution, district (or county), city, province/state, course name, grade level (9–16 entered as a numeral), and school year in the appropriate boxes.
- j. Select the appropriate category of samples that your students will be using.
- k. Enter the names of your students. Email addresses are not required, but can be entered if desired.
- l.  Click the **Submit** button. This will open a page where the course username and password will be shown. You should also receive an email with the same information. **Make sure to record your course username and password, as these will be required to work on your data within BOLD-SDP.**

Course Username: _____

Course Password: _____

Step 2: Creating specimen folders into which student groups will upload data

At this point, you have registered yourself, the instructor, with BOLD-SDP. You have received an instructor username and password and five registration keys, each of which can be used to generate a separate course username and password for each class, and you have registered your students and received your first course username and password.

The following steps outline how to create specimen folders so that forward and reverse sequence data can be uploaded to BOLD-SDP. This step can be performed either by the instructor or by the students, but must be performed before sequencing trace files can be uploaded and analyzed in BOLD-SDP.

- a. Access the homepage of BOLD-SDP at: boldsystems.org/index.php/SDP_Home. Click the **Students** button to access the Student Login page.

An integrated workbench supporting the assembly, analysis and publication of DNA barcode data by students.

- System Overview
- Quick Start Guide
- Video Introduction

The BOLD Student Data Portal (BOLD-SDP) is a classroom-focused interface to the BOLD database. This platform provides instructors and students with the tools necessary to make contributions to the DNA barcode library used for identifying species. It also gives students the opportunity to integrate and analyze specimen and sequence data while providing instructors with tools to monitor student progress and evaluate their work. Students can explore the large database of DNA barcode records submitted by scientists around the world, but they will also be able to add their own data to the ever-growing DNA barcoding library. Contact us for support in using this system in a classroom setting (edu@boldsystems.org).

Students

A portal for students to assemble DNA barcode records in a simplified but powerful data collection and analysis environment.

Instructors

A management interface for instructors to register students and classrooms; review student work and utilize data validation workflows.

Explore

Search for and browse data records, images, and geographic distributions for species being barcoded.

- b. Enter the course username and password into the appropriate spaces and click **Log In** to enter the Main Student Console page. Note that the password is case sensitive!

Student Data Portal

Student Login

The Student Data Portal of BOLD is designed to support the assembly and analysis of DNA barcode records in a classroom environment. The system requires the registration of instructors, courses, and students prior to data submissions or analyses.

If you are a student, you should login with a single account generated for your course. Through this account, you can work with fellow students to contribute new records to the Barcode Database. If you lose or forget your password, please contact your instructor.

If you and your classmates would like to conduct a DNA barcoding project and do not have any login information, please contact your instructor and refer him/her to this website.

Login

Sign in with your course username and password to start your scientific experience in the world of DNA barcoding.

Username

Password

[Log In](#)

- c. You should now see the Data Management Console.

Data Management Console

New Specimen



Upload Images



Upload Traces



Add Sequence



- d. Click the **New Specimen** icon.



The New Specimen page will open. For each student group, you will need to enter multiple pieces of information.

First, using the Student Attribution box, choose the student(s) who worked on this sample. If more than one student worked on this sample, click the **Add Student** button and select his/her name from the dropdown menu. Repeat the **Add Student** process until all students who worked on this sample have been added. This is important in order to later assign credit to the students who did the work for each sample in various steps of the process.

▼ **Student Attribution**

Please enter the names of all student contributors to this record.

Student name:

[Add Student](#)



In step A, **Specimen Identifier**, enter a name to uniquely identify the sample you are testing. The specimen identifier needs to be a name that is unique not only for the samples tested by your class, but also unique within the entire BOLD database. **The best way to ensure a unique specimen identifier name is to include the following information in the name: 1) year, 2) institution, 3) group name, and 4) some information about the sample.** After entering a name, press the **Tab** key. If the name is unique, a green arrow will appear on the right-hand side of the sample ID name. An example of a unique specimen identifier name is shown below.



Note: It is important to take extra care when entering your desired sample ID, as this information cannot be modified later. If a mistake is made, the entire specimen record can be deleted, but this option is accessible only when logged in as the instructor.

A **Specimen Identifier**

Sample ID:

In step B, Specimen Details, if you know the life stage, sex, and reproduction mode of the sample you are using, you can select those parameters. Otherwise, select the **Unknown** circle.

B Specimen Details

Life Stage: Adult Immature Unknown

Sex: Male Female Hermaphrodite Unknown

Reproduction: Sexual Asexual Cyclic Parthenogen
 Unknown

Notes:

In step C, Taxonomy, find out the appropriate phylum and choose that from the dropdown menu. Also, you can type comments in the Taxonomy Notes to specify where you collected your sample. Only the phylum data field is required — other taxonomic data are optional.

C Taxonomy

Phylum *: ▼

Class: ▼

Order: ▼

Family: ▼

Subfamily: ▼

Tribe: ▼

Genus: ▼

Species: ▼

Identification Method (morphology, barcode, etc):

Taxonomy Notes:

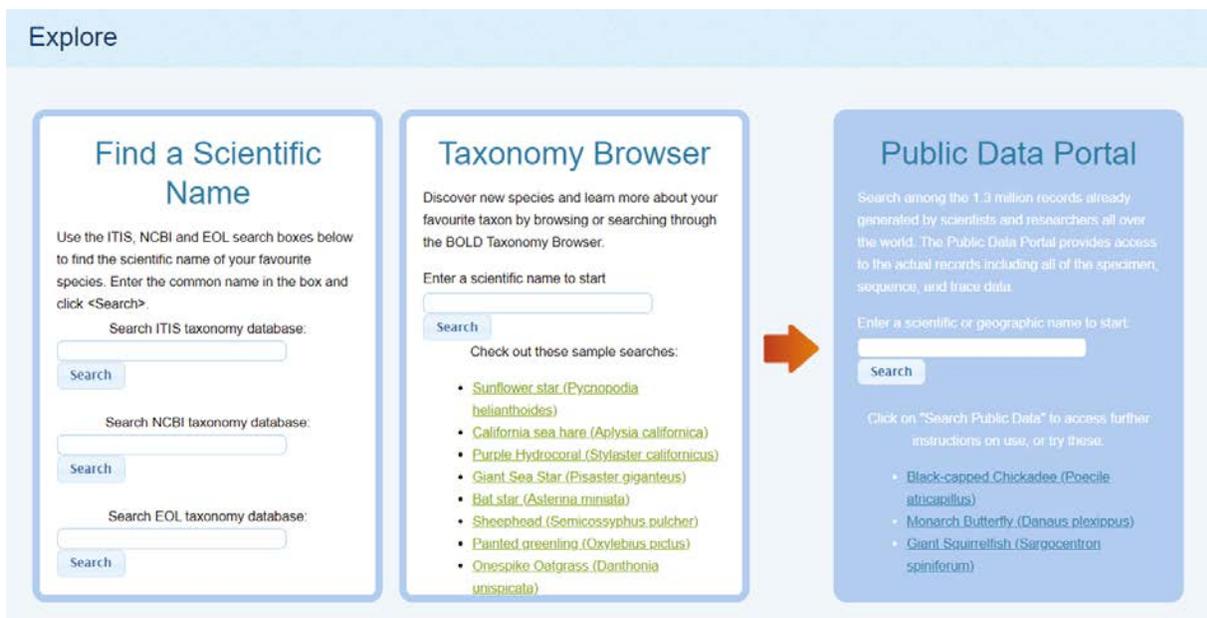
If you do not know the phylum of your sample, you can use a search browser in BOLD-SDP to find it.

1. Open a new browser page and go to the BOLD-SDP homepage at: boldsystems.org/index.php/SDP_Home.
 You do not need to be logged in to your class account in order to perform the following searches.

2. Click **Explore**.



Type the common name of your sample (for example, shrimp) into the Search ITS taxonomy database box, then click **Search**.



A new page will open with multiple species and subspecies. Find your sample on the list and click it to get its entire taxonomical tree, including the phylum to which it belongs — in the case of shrimp, Arthropoda.



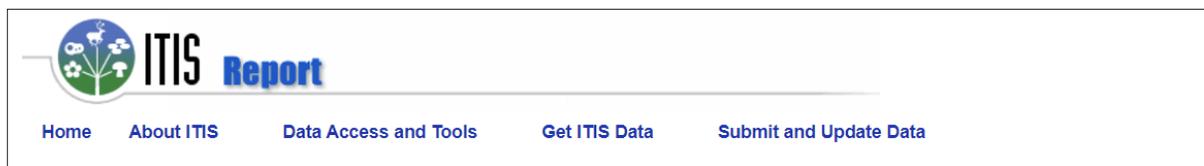
ITIS Search Results

Home About ITIS Data Access and Tools Get ITIS Data Submit and Update Data

Results of: Search in every Kingdom for Common Name containing 'shrimp'

Kingdom Animalia

- [abalone visored shrimp](#) – Species: *Betaeus harfordi* (Kingsley, 1878) – valid
- [Adonis shrimp](#) – Species: *Parapenaeopsis venusta* De Man, 1907 – valid
- [Aesop shrimp](#) – Species: *Pandalus montagui* Leach, 1814 – valid
- [African mud shrimp](#) – Species: *Solenocera africana* Stebbing, 1917 – valid
- [african spider shrimp](#) – Species: *Nematocarcinus africanus* Crosnier and Forest, 1973 – valid
- [akiami paste shrimp](#) – Species: *Acetes japonicus* Kishinouye, 1905 – valid
- [Alabama cave shrimp](#) – Species: *Palaemonias alabamae* Smalley, 1961 – valid
- [Alachua fairy shrimp](#) – Species: *Branchinella alachua* Dexter, 1953 – invalid
- [Alachua fairy shrimp](#) – Species: *Dendrocephalus alachua* (Dexter, 1953) – valid
- [alamang shrimp](#) – Species: *Acetes sibogae* Hansen, 1919 – valid
- [Alaska bay shrimp](#) – Species: *Neocrangon alaskensis* (Lockington, 1877) – invalid
- [Alaska coastal shrimp](#) – Species: *Heptacarpus moseri* (M. J. Rathbun, 1902) – valid
- [Alaskan pink shrimp](#) – Species: *Pandalus eous* Makarov, 1935 – valid
- [Aleutian coastal shrimp](#) – Species: *Heptacarpus maxillipes* (M. J. Rathbun, 1902) – valid
- [alkali fairy shrimp](#) – Species: *Branchinecta mackini* Dexter, 1956 – valid
- [American grass shrimp](#) – Species: *Perilimnopsis americana* (Kingsley, 1878) – valid



ITIS Report

Home About ITIS Data Access and Tools Get ITIS Data Submit and Update Data

[Go to Print Version](#)

***Trypaea australiensis* Dana, 1852**
 Taxonomic Serial No.: 552889

Download TWB Download DwC-A [\(Download Help\)](#) *Trypaea australiensis* TSN 552889

Taxonomy and Nomenclature

Kingdom:	Animalia
Taxonomic Rank:	Species
Synonym(s):	Callianassa australiensis (Dana, 1852)
Common Name(s):	Australian ghost shrimp [English]

Taxonomic Status:
 Current Standing: valid

Data Quality Indicators:
 Record Credibility Rating: verified - standards met

Taxonomic Hierarchy

Kingdom	Animalia – Animal, animaux, animals
Subkingdom	Bilateria
Infrakingdom	Protostomia
Superphylum	Ecdysozoa
Phylum	Arthropoda – Artrópode, arthropodes, arthropods
Subphylum	Crustacea Brünnich, 1772 – crustacés, crustáceo, crustaceans
Class	Malacostraca Latreille, 1802
Subclass	Eumalacostraca Grobben, 1892
Superorder	Eucarida Calman, 1904 – camarão, caranguejo, ermitão, lagosta, siri
Order	Decapoda Latreille, 1802 – crabs, crayfishes, lobsters, prawns, shrimp, crabes, crevettes, écrevisses, homards
Suborder	Pleocyemata Burkenroad, 1963
Infraorder	Thalassinidea Latreille, 1831
Superfamily	Callianassoidea Dana, 1852
Family	Callianassidae Dana, 1852 – ghost shrimps
Subfamily	Callianassininae Dana, 1852
Genus	Trypaea Dana, 1852
Species	<i>Trypaea australiensis</i> Dana, 1852 – Australian ghost shrimp

In step D, Collection Details, the minimum detail that needs to be entered is the Country/Ocean of sample collection. Use the dropdown menu to enter the information. Unless you collected the sample yourself from the field and are certain of the location where it was caught, you should list the location of the store/restaurant from which you purchased the sample.

Once all data have been entered, click the **Submit** box. You will see a Submission Confirmation page.

Repeat the above steps to generate sample IDs for all student samples tested. To get to the main console page in order to create more sample IDs, click the **Main Console** button.

Once all sample IDs have been entered, you can see the complete list by clicking the **View Class Records** button on the Submission Confirmation page or by clicking the **View Data** button on the Main Console Page.

Identification ▼	Sample ID ▼	Process ID ▼
<input type="checkbox"/> Chordata	2013-Bio-Rad-CGAK-Salmon	SDP27001-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-CGAK-Tuna	SDP27002-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-CCTZ-Trout	SDP27003-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-CCTZ-CookedSalmon	SDP27004-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-ACLSTW-coho	SDP27005-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-ACLSTW-tuna	SDP27006-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-ACLSTW-grouper	SDP27007-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-CGAK-pCOIControl	SDP27008-13

Once you have confirmed that all student samples have sample IDs, click the **Main Console** button to go back to the Main Console page.

Step 3: Uploading forward and reverse trace files into your account on BOLD-SDP**1 Day**

These steps may be performed by the instructor or by the students. Sample IDs must have been created for all student samples tested in order to upload traces. Also, after each trace is uploaded to BOLD-SDP, the software will analyze each and assign a quality ranking of high, medium, low, or failed. **It can take up to 24 hours for these quality assignments to be made.** Further analysis of data cannot be performed until the quality assignments have been made by the software. Therefore, if sequence trace files are uploaded by the students, no other analyses can be performed during that class period. **If a continuous workflow by students is desired, then all the sequence trace files should be uploaded to the appropriate Sample ID folder at least one day before the class period.**

- If you are not already logged in, log in to the Main Student Console page of BOLD-SDP by going to boldsystems.org/index.php/SDP_Home and clicking the **Students** button.
- Enter the course username and password into the appropriate spaces and click the **Log In** button to enter the Main Student Console page. Note that the password is case sensitive!

Student Data Portal

Student Login

The Student Data Portal of BOLD is designed to support the assembly and analysis of DNA barcode records in a classroom environment. The system requires the registration of instructors, courses, and students prior to data submissions or analyses.

If you are a student, you should login with a single account generated for your course. Through this account, you can work with fellow students to contribute new records to the Barcode Database. If you lose or forget your password, please contact your instructor.

If you and your classmates would like to conduct a DNA barcoding project and do not have any login information, please contact your instructor and refer him/her to this website.

Login

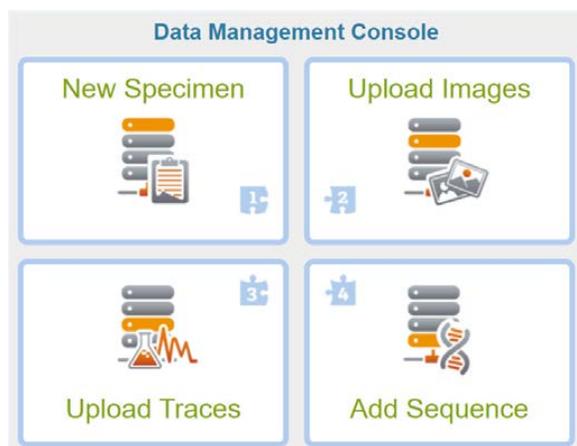
Sign in with your course username and password to start your scientific experience in the world of DNA barcoding.

Username

Password

[Log In](#)

- You should now see the Data Management Console.



If you have taken photographs of your sample, its packaging, or of the sample pieces before you performed the DNA extraction, those photographic images can be uploaded by using the **Upload Images** button. Otherwise, begin by clicking the **Upload Traces** button. Traces refer to the files you received from your sequencing facility. These files contain not only information on the base calls at each location of your PCR product, but also information on the quality of each base call.

- d. Click the **Upload Traces** button.



- e. Using the Student Attribution box, choose the student(s) who worked on this sample. If more than one student worked on this sample, click the **Add Student** button and select his/her name from the dropdown menu. Repeat the **Add Student** process until all students who worked on this sample have been added.

▼ Student Attribution

Please enter the names of all student contributors to this record.

Student name:

[Add Student](#)

- f. In Section A, the Specimen Identifier section of the Upload Traces page, enter the sample ID for your first sample. If you do not remember exactly the name of the sample ID for your sample, you can click the **Lookup** button, which will open in a new tab the Record List for your course, and find the sample ID you need. Below is an example of a sample ID list for a class's data.

Identification ▼	Sample ID ▼	Process ID ▼
<input type="checkbox"/> Chordata	2013-Bio-Rad-CGAK-Salmon	SDP27001-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-CGAK-Tuna	SDP27002-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-CCTZ-Trout	SDP27003-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-CCTZ-CookedSalmon	SDP27004-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-ACLSTW-coho	SDP27005-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-ACLSTW-tuna	SDP27006-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-ACLSTW-grouper	SDP27007-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-CGAK-pCOIControl	SDP27008-13

Note: If you click the sample ID you want to use from the Record List page, it will open another information page but will not fill in the information in the Specimen Identifier Sample ID box.

Once you know the name of your sample ID, close or minimize the Record List window.

Type or copy and paste the appropriate sample ID in the Specimen Identifier Sample ID box and press the **Tab** key. If your sample ID was successfully found, you will see three icons appear in the Specimen Identifier box and a green check will appear next to the Sample ID entry. The first of the icons shows how many sequencing traces have been uploaded into BOLD-SDP for this sample. The second icon shows how many photographic images have been uploaded for this sample into BOLD-SDP. The third icon shows how many contiguous sequences have been generated and saved for this sample in BOLD-SDP. For the case shown below, nothing has been uploaded or generated at this point.

A Specimen Identifier

Sample ID: ✓

0 x 0 x 0 x

- g. In Section B, the PCR Primers section of the Upload Traces page, use the dropdown menus to select the forward and reverse PCR primers you used to generate your PCR product. Here, if you are barcoding a fish sample, the forward PCR primers used were **C_FishF1t1** and the reverse PCR primers used were **C_FishR1t1**. If you are barcoding any other animal, including insects, the forward PCR primers used were LepF1 and the reverse PCR primers used were LepR1.

B PCR Primers

Forward: ▾

Reverse: ▾

- h. In Section C, the Attach Trace Files section of the Upload Traces page, use the dropdown menus to select the forward and reverse sequencing primers the sequencing facility used to generate your trace files (the sequencing data files). Here the forward sequencing primer used was **M13F-20** and the reverse sequencing primer used was **M13R**.

C Attach Trace Files

Forward: Sequencing Primer ▾
 No file chosen

Reverse: Sequencing Primer ▾
 No file chosen



Now, click the **Browse** button under the Forward section and choose your sequencing trace file that corresponds to the forward reaction sent to you from the sequencing facility. This file should end in .ab1. **It is critical that the sequencing trace file name does not contain any characters such as : ; or &. Remove these from the file names before trying to upload them to BOLD-SDP. It is also critical that the name be unique, as the BOLD-SDP portal does not allow duplicate file names. Try to follow the same general naming scheme as the sample ID (year-institution-initials-sample-forward).** Do the same for the reverse sequencing trace file (year-institution-initials-sample-reverse). Now click the **Submit** button to add your data to the BOLD-SDP database.

If everything was entered correctly, you will see a Submission Confirmation page.

Submission Confirmation

- Your trace file has been successfully submitted

Please choose from the options below:

View Class Records
Main Console
View Sequence

- i. If you have additional samples, click the **Main Console** button. This will open the Data Management page, where you will need to follow steps e–i to upload your traces for your next sample. Make sure to use the correct specimen ID for each of the sequencing traces being uploaded.
- j. Once all sequencing traces have been uploaded to BOLD-SDP, you can confirm that they are in the database by clicking the **View Class Records** button from the Submission Confirmation page or **View Data** from the Main Student Console page.

Identification ▼	Sample ID ▼	Process ID ▼	Length [Ambig] COI-5P ▼	Record Flags	Extra Info
<input type="checkbox"/> Chordata	2013-Bio-Rad-ACLSTW-coho	SDP27005-13	0	2	
<input type="checkbox"/> Chordata	2013-Bio-Rad-ACLSTW-grouper	SDP27007-13	0	2	

If sequencing trace files were successfully uploaded, there will be an icon showing the number of files for each sample ID.



Since students will need to know their sample ID(s), it might be helpful to print out a screen shot of the Record List page and give this to students.

1 Day

Note: If the instructor is uploading all student trace files, repeat this process until all trace files are uploaded to their appropriate specimen ID folders. Then **allow up to 24 hours for the database to calculate quality scores for each trace** before students can begin their analyses.

Step 4: Assessing quality values for student sequencing traces to determine bioinformatics workflow

The BOLD system is designed to carefully screen sequence trace data before species identification can be made. Because of this, the system does not allow many bioinformatics processes on lower quality sequence data. However, this does not preclude students from learning the steps of sequence analysis, even if their data is of lower quality.

- The first protocol in the Student Instruction Manual involves learning to look at sequence trace files, and this can be done for any sequencing trace, whether high- or low-quality
- The second protocol in the Student Instruction Manual involves determining the best match — either within the BOLD database and/or using a BLAST search of GenBank — of single sequencing trace files to determine how closely they match if only one sequence trace file is used at a time
- The subsequent protocols involve generating a contig (the best consensus sequence using both the forward and reverse sequencing trace files), trimming off low-quality data at the ends, correcting any discrepancies in the contig file, and then determining the best match in the BOLD database to this contig file. If the initial sequencing trace file quality designation was low or fail, these activities cannot be performed with the student data. However, sequencing trace files are available for download at bio-rad.com/barcoding, on the download tab, and can be used to complete all of these activities

Students should not be discouraged if their sequencing trace files are of lower quality, but instead should use this as an opportunity to learn about what factors can impact sample purity, sample processing, and sequencing results. There are many places where contamination can occur that are outside the control of the students. Sources of contamination include supermarkets, where samples may have bacterial growth or be cross-contaminated with other samples cut with the same knife. Some samples, such as fried samples or pickled and canned samples, may have damaged DNA. Pet store “dead fish” samples may have started to decay, destroying the integrity of the DNA within them.

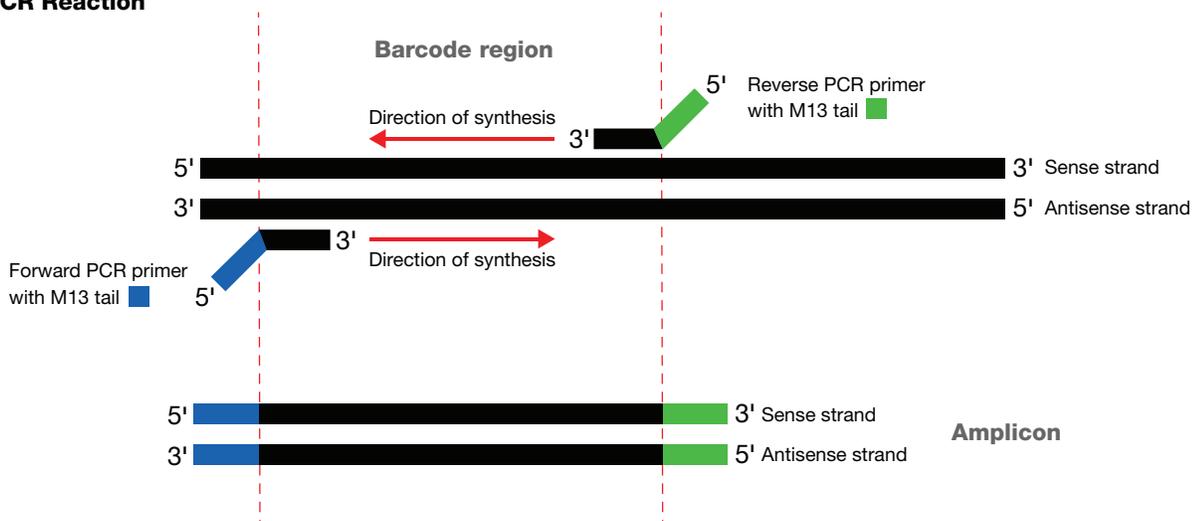
Student Instruction Manual

DNA Sequencing of Amplicons Using Dye Terminator Cycle Sequencing

Background — DNA sequencing is a procedure for determining the order in which nucleotides (adenine, guanine, cytosine, and thymine) appear, regardless of whether that DNA consists of small pieces or complete genomes. Over the last several decades, a variety of sequencing methods have been developed for different applications and research goals. A researcher's selection of a particular method is based on a variety of considerations, including speed, cost, accuracy, and the length of the DNA molecule to be sequenced. Dye terminator cycle sequencing — an automated variation of Sanger sequencing — is the method of choice for DNA barcoding. This PCR-based method of automated DNA sequencing is performed at a nominal cost by both commercial and university-based sequencing facilities.

Methodology — for DNA barcoding, two dye terminator sequencing reactions are performed separately for each amplicon (PCR product). The forward sequencing reaction will determine the nucleotide sequence of the sense strand of DNA, whereas the reverse sequencing reaction will determine the nucleotide sequence of the antisense strand. Unlike in conventional PCR, only a single oligonucleotide primer is used for each sequencing reaction (Figure 1).

PCR Reaction



Sequencing Reaction

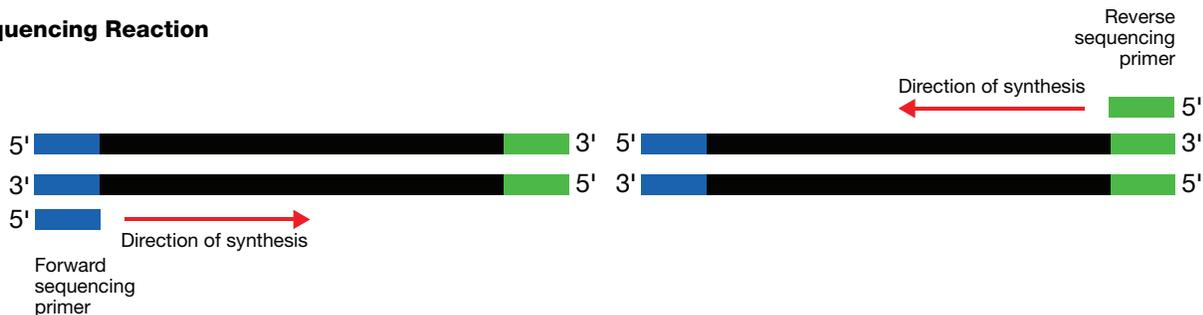


Fig. 1. Both amplification of a portion of DNA and sequencing reactions employ the polymerase chain reaction. PCR amplification of a portion of DNA uses a pair of primers used together in the same reaction with the end result being a final amplicon (double-stranded PCR product) of defined length. Sequencing reactions also use a template of double-stranded DNA; however, sequencing reactions use only one primer per reaction and create single strands of DNA terminated by ddNTPs rather than double-stranded PCR products.

The following components are common to both the forward and reverse sequencing reactions, which are performed in separate tubes:

1. Multiple copies of a double-stranded amplicon (the DNA template for each sequencing reaction)
2. A heat-stable DNA polymerase
3. dNTPs (the basic building blocks of DNA)
4. ddNTPs (fluorescently labeled terminator nucleotides that lack an –OH group at position 3 of the ribose ring)

Because the ddNTPs lack an –OH group at position 3 of the sugar (Figure 2), they cannot be involved in further extension of the sequencing reaction PCR product. Therefore once a ddNTP is incorporated, the reaction stops for that one chain. The only component that is different between the forward and reverse sequencing reaction is the single sequencing primer that is used.

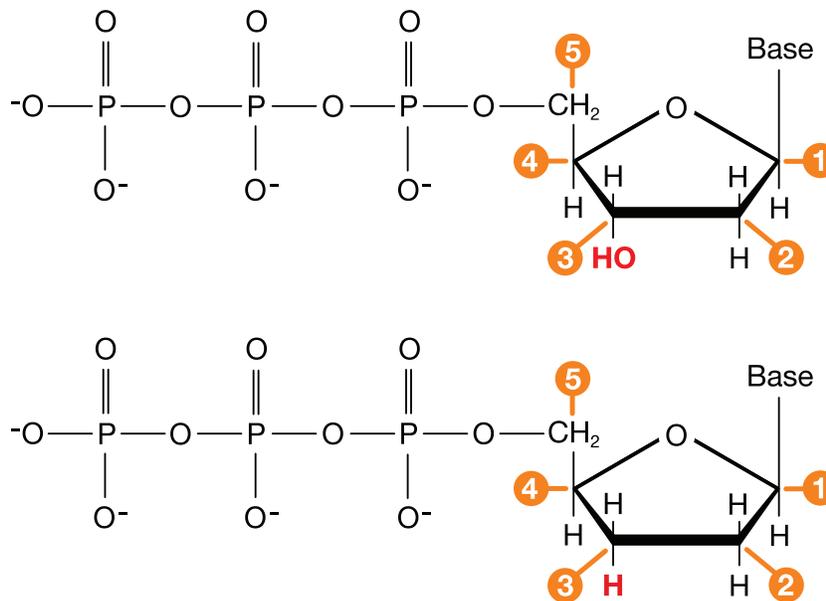


Fig. 2. Structure of dNTPs and ddNTPs. A. dNTPs have a 3-hydroxyl (3-OH) group, which is necessary for elongation of DNA. **B.** ddNTPs do not have a 3-OH; instead, the 3' position has been modified to have a hydrogen (–H) at that position. When a ddNTP is incorporated into a DNA molecule, the synthesis ends at that nucleotide and the DNA chain is terminated. Figure reprinted from Brown K (2018). *Biotechnology, A Laboratory Skills Course* (Bio-Rad Laboratories Inc.), p.199.

Each sequencing reaction progresses through the same major steps of a PCR reaction:

1. During the **denaturation** step, each sequencing reaction mixture is heated to ~95°C to disrupt the hydrogen bonds that hold the sense and antisense strands of the amplicon together.
2. During the **annealing** step, each reaction mixture is lowered to ~50°C, allowing the sequencing primer to bind to a complementary sequence on one strand of the amplicon. The sequencing primer that was added to the forward sequencing reaction binds, or anneals, to a complementary sequence on the antisense strand according to the base pairing rules. The sequencing primer that was added to the reverse sequencing reaction anneals to a complementary sequence on the sense strand.
3. During the **elongation** step, each reaction mixture is raised to ~72°C. At this temperature, a heat-stable DNA polymerase finds the 3' end of the sequencing primer and begins joining nucleotides that are complementary to those present in the template strand. For the forward sequencing reaction, the DNA polymerase joins nucleotides that are complementary to those in the antisense strand. For the reverse sequencing reaction, the DNA polymerase joins nucleotides that are complementary to those in the sense strand.

Instructor Guide

During this step of the sequencing reaction, the DNA polymerase cannot distinguish between dNTPs and ddNTPs present in the reaction mixture. Because a higher proportion of dNTPs are added to each sequencing reaction mixture, they are more likely to be incorporated into the growing DNA chain. However, when a ddNTP lacking a 3' -OH is incorporated, DNA synthesis stops, as no new nucleotides can be added to the growing chain.

The denaturation, annealing, and elongation steps are repeated multiple times, thereby ensuring that at the conclusion of each sequencing reaction, single-stranded DNA fragments of every possible length are generated. Importantly, each fragment terminates with one of the four ddNTPs, which are labeled with a different fluorescent tag.

Upon completion of each sequencing reaction, the fluorescently labeled DNA fragments are separated according to size using capillary electrophoresis, which is electrophoresis performed in a long and extremely narrow tube. As the DNA fragments migrate from smallest to largest through the capillary tube, they pass through a laser, which excites the fluorescent ddNTP at the terminal end of each fragment.

- DNA fragments terminated by ddATP emit green light
- DNA fragments terminated by ddTTP emit red light
- DNA fragments terminated by ddGTP emit yellow light
- DNA fragments terminated by ddCTP emit blue light

The light emitted from fluorescently labeled DNA fragments is detected by the sequencer and represented as a continuous series of colored peaks in an electropherogram, or trace file (Figure 3). The peak from the smallest fluorescently labeled DNA fragment is represented first in the trace file, whereas the peak from the largest fragment is represented last. The information contained in a trace file will be discussed in greater detail below.

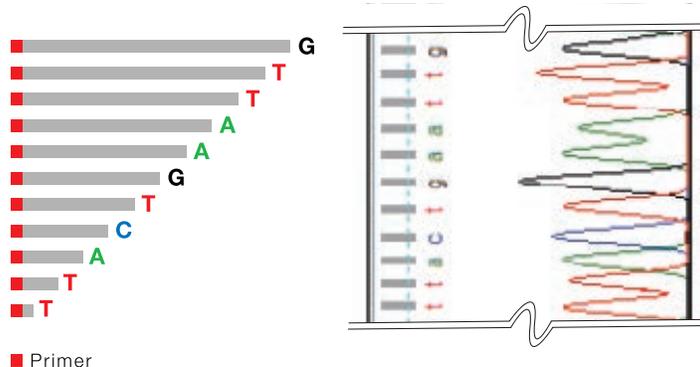


Fig 3. Electropherogram. (Left) DNA fragments of varying lengths each have a fluorescent ddNTP at the terminal end, that indicates the base at the final position. (Right) Peaks with colors that correspond to the identity of the terminal base indicate the DNA sequence. Figure reprinted from Brown K (2018). *Biotechnology, A Laboratory Skills Course* (Bio-Rad Laboratories, Inc.), p 199.

Sequence Assembly and Editing in BOLD-SDP

At this point, you have isolated DNA from a sample, amplified the barcoding region from that genomic DNA using PCR, analyzed your amplicon using agarose gel electrophoresis, and submitted your amplicon for two sequencing reactions — one in the forward direction and one in the reverse direction. You should now have two data files from the sequencing facility for each amplicon you sent for sequencing — one file representing the data from the forward sequencing reaction and one file representing the data from the reverse sequencing reaction. It is now time to analyze that data. The general workflow for analysis of your data is outlined below.

1. Assess the quality of DNA sequencing data.
2. Query either the BOLD database or GenBank for matches for forward and reverse sequencing data.
3. Assemble a single consensus sequence (contig) from your two sequencing reactions.
4. Manually compare any nucleotide calls that are different between the forward and reverse sequencing reactions.
5. Perform a search to determine the identity of your sample contig data.

Sequence Trace Files

As noted, the fluorescently labeled DNA fragments generated during dye terminator cycle sequencing migrate sequentially through a capillary according to their size (smallest to largest) and pass by a laser. Upon exposure to the laser, fragments terminated by a ddATP emit green light, fragments terminated by ddTTP emit red light, fragments terminated by ddCTP emit blue light, and fragments terminated by ddGTP emit yellow light. The light signals are detected by the DNA sequencer, processed by a software program, and represented as a series of colored peaks in a trace file (yellow light signals emitted from DNA fragments terminated by ddGTP are represented as black peaks in the trace file to make them more readable on a white background).

The software also 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. To compute quality scores, the algorithm examines several parameters associated with the shape and resolution of the peak as well as the signal-to-noise ratio at each position in the trace file. The resulting scores are logarithmically linked to error probabilities according to the following equation:

$$Q = -10 \log_{10} P$$

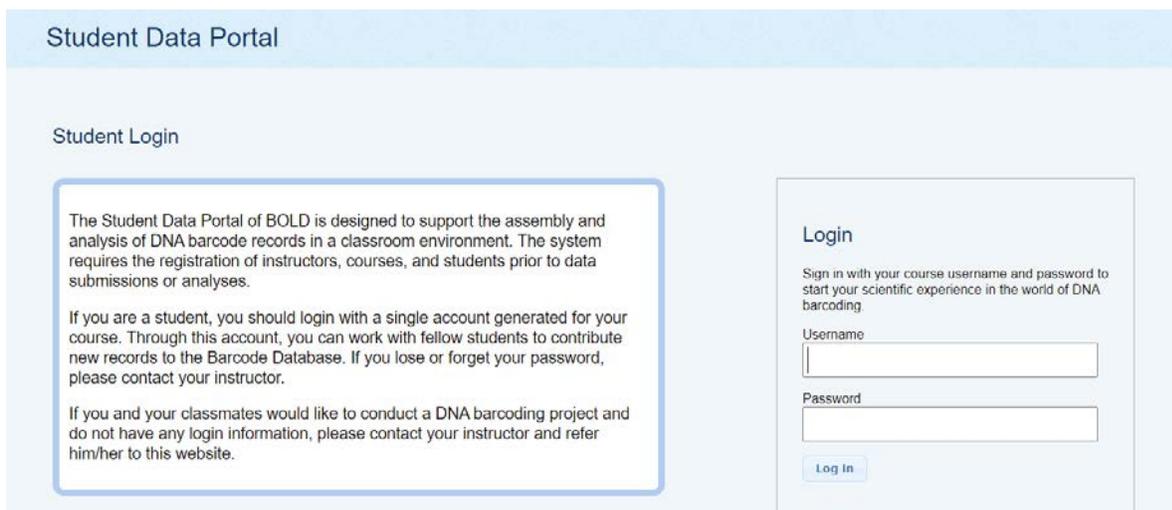
where P represents the probability of an incorrect base call.

Based on this equation, a quality score of 20 indicates that the probability of an incorrect base call is 1 in 100, whereas a quality score of 40 indicates that the probability of an incorrect base call is 1 in 10,000. Generally speaking, for submission of a vouchered sample sequence, quality scores below 50 are considered unacceptable. This stringency is much higher than for normal sequence assessment, due to the BOLD database's standard of including data of only the highest integrity.

Step 1. Assessing the quality of DNA sequencing data

One of the first steps in bioinformatics is to analyze the quality of the sequencing data. The ability to look at sequencing data and assess whether or not further analyses are worthwhile is a critical skill. For example, did both of your sequencing reactions result in approximately the same length of sequence data? Did both sequencing reactions work well to give you high-quality data? It is important to know the quality of a sequencing reaction because, for example, if there are a lot of ambiguous base calls, how can you differentiate between poor quality data or actual base differences between different samples?

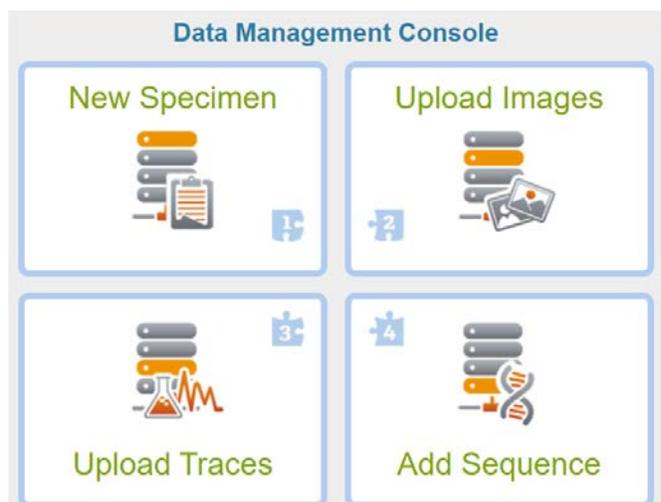
- a. Go to the Main Student Console page of BOLD-SDP at: boldsystems.org/index.php/SDP_Home and click the **Students** button to access the Student Login page.
- b. Your instructor will provide a course username and password. Enter these into the appropriate spaces and click **Log In** to enter the Main Student Console page. Note that the password is case sensitive!



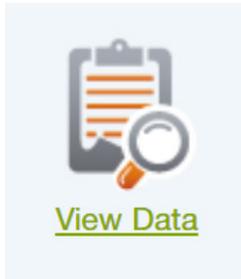
Course Username: _____

Course Password: _____

- c. You should now see the Data Management Console.



- d. In the right sidebar, click the **View Data** button.



- e. On the Record List page for your class, find the line with the sample ID that corresponds to your sample. If your instructor created these sample IDs, get the sample IDs from him/her. Record the sample IDs that correspond to your samples.

Identification ▼	Sample ID ▼	Process ID ▼
<input type="checkbox"/> Chordata	2013-Bio-Rad-CGAK-Salmon	SDP27001-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-CGAK-Tuna	SDP27002-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-CCTZ-Trout ←	SDP27003-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-CCTZ-CookedSalmon	SDP27004-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-ACLSTW-coho	SDP27005-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-ACLSTW-tuna	SDP27006-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-ACLSTW-grouper	SDP27007-13

Record Sample ID for Sample 1: _____

Record Sample ID for Sample 2: _____

Then, click the **Sample ID** link for your first sample.

This will open a new page that includes all the specimen data that either you or your instructor has entered for this sample. Recording this sort of data is critical for submission of new barcode data for species because it is a trail of evidence detailing where the organism was found and other information about the sample. This is similar to the evidence trail needed for crime scenes. How much would you trust a DNA sample if the crime scene investigator was not sure exactly what room and location the sample was found in and on what day? It could be from the wrong crime scene!

IDENTIFIERS	PHOTOGRAPHS
<p><u>Sample ID:</u> 2013-Bio-Rad-CCTZ-Trout</p> <p><u>Process ID:</u> SDP27003-13</p> <p><u>Institution Storing:</u> Coastal Marine Biolabs</p> <p><u>Field ID:</u> 2013-Bio-Rad-CCTZ-Trout</p> <p><u>Museum ID:</u></p> <p><u>Collection Code:</u></p>	<p>No images available</p>
TAXONOMY	GEOGRAPHY
<p><u>Identification:</u> Chordata</p> <p><u>Rank:</u> Phylum</p> <p><u>Identifier:</u> Student</p> <p><u>Identification Method:</u></p> <p><u>Identifier Institution:</u></p> <p><u>Identifier Email:</u></p> <p><u>Taxonomy Note:</u></p> <p><u>Rank:</u> Current Record (2013-Bio-Rad-CCTZ-Trout)</p> <p><u>Phylum:</u> Chordata</p> <p><u>Class:</u></p> <p><u>Order:</u></p> <p><u>Family:</u></p> <p><u>Subfamily:</u></p> <p><u>Genus:</u></p> <p><u>Species:</u></p> <p><u>Subspecies:</u></p>	<p><u>Country:</u> Atlantic Ocean</p> <p><u>Province/State:</u></p> <p><u>Region/Country:</u></p> <p><u>Sector:</u></p> <p><u>Exact Site:</u></p> <p><u>Lat/Lon:</u></p>
SPECIMEN DETAILS	COLLECTION DETAILS
<p><u>Voucher Status:</u></p> <p><u>Tissue Descriptor:</u></p> <p><u>Sex:</u></p> <p><u>Reproduction:</u></p> <p><u>Life Stage:</u></p> <p><u>Extra Info:</u></p> <p><u>Note:</u></p> <p><u>Associated Taxa:</u></p> <p><u>Associated Specimens:</u></p> <p><u>Reference Link:</u></p> <p><u>Contributors:</u></p> <ul style="list-style-type: none"> • Specimen: Cherie Chan, Tim Zimmerman • Trace Run Site: Generic Commercial Lab • Trace Upload: Cherie Chan, Tim Zimmerman 	<p><u>Collectors:</u></p> <p><u>Collection Event ID:</u></p> <p><u>Date Collected:</u></p> <p><u>Date Accuracy:</u></p> <p><u>Time Collected:</u></p> <p><u>Site Code:</u></p> <p><u>Habitat:</u></p> <p><u>Sampling Protocol:</u></p> <p><u>Coord. Source:</u></p> <p><u>Coord. Accuracy:</u></p> <p><u>Elevation:</u></p> <p><u>Elevation Accuracy:</u></p> <p><u>Depth:</u></p> <p><u>Depth Accuracy:</u></p> <p><u>Collection Notes:</u></p>

Once you are finished reviewing the data on the Specimen page, close or minimize the page.

f. Now click the Process ID link for your specimen on the Record List page.

Identification ▼	Sample ID ▼	Process ID ▼
<input type="checkbox"/> Chordata	2013-Bio-Rad-CGAK-Salmon	SDP27001-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-CGAK-Tuna	SDP27002-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-CCTZ-Trout	SDP27003-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-CCTZ-CookedSalmon	SDP27004-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-ACLSTW-coho	SDP27005-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-ACLSTW-tuna	SDP27006-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-ACLSTW-grouper	SDP27007-13
<input type="checkbox"/> Chordata	2013-Bio-Rad-CGAK-pCOIControl	SDP27008-13

A sequence page should open in a new window.

Run Date	Direction	Trace File	Seq Primer	Quality
2013-01-16	Forward	Tun-forwardM13For.ab1	M13F-20	low
2013-01-16	Reverse	Tun-reverseM13Rev.ab1	M13R	low

The PCR primer names, sequencing primer names, and trace file names should appear in the Sequencing Runs pane.

BOLD-SDP also displays a quality designation (high, medium, low, or fail) for each of the trace files. These designations are based on the average quality scores of the base calls in each trace file and are designations based on BOLD standards of data quality for uploading.



Note: Trace files with low or failed designations indicate that there might be a problem with the initial sample, isolation of the DNA, the PCR reaction, or the sequencing reaction. Trace files with low or failed designations generally cannot be used to assemble contigs using the BOLD Sequence Editor. You may nevertheless use the BOLD Identification System (BOLD-IDS) to determine the **possible** identity of the specimen from which your sequence data were generated. Those steps will be outlined in step 2.

- g. To examine the sequencing trace files, **select both check boxes** that appear next to their filenames and then click **View Trace Files**.

- h. The forward and reverse trace files are displayed in the top and bottom panes of the Trace Viewer page, respectively.



The Trace File Viewer displays quality values for individual base calls in the trace files using a histogram. The quality value for each base call can be determined by comparing the height of its blue shaded bar to the vertical scale on the right-hand side of the trace file window. Continuous stretches of low-quality base calls on the 5' and 3' ends of each trace file are displayed in gray.

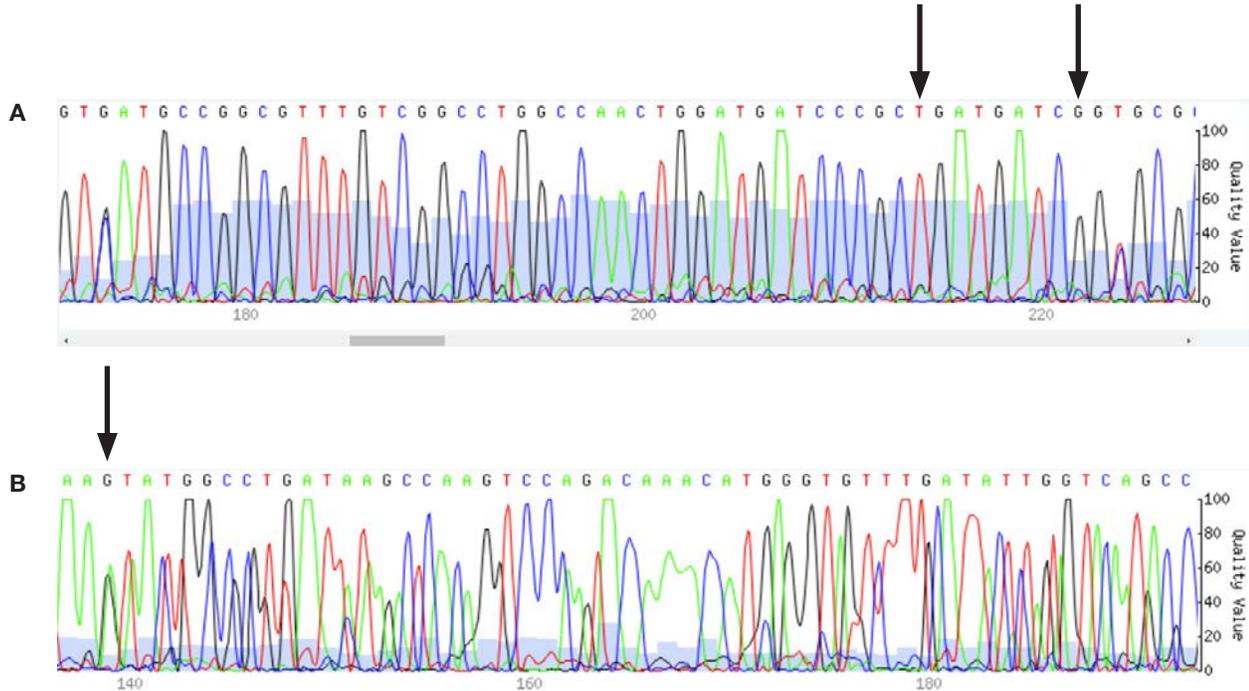


Fig. 4. A comparison of quality values in sequencing trace files. **A**, the T base call under the first arrow has a high quality value. The light blue shaded bar shows a quality value of ~60, which means there is only a 1 in 1,000,000 chance that this was the incorrect base call. The peak height and shape and the presence of other peaks at the same location (signal-to-noise ratio) impact the quality value score for each base call. The G base call under the second arrow has a lower quality value. The light blue shaded bar shows a quality value of ~25, which means there is about a 1 in 300 chance that this was the incorrect base call. Notice the peak for this base call is smaller than the surrounding peaks. **B**, for the second sequencing trace file example, all the quality scores are low (mostly below 20, meaning there is a greater than 1 in 100 chance the call is not correct). While there are some nice tall peaks at each location, there are also lots of underlying peaks. At the base call under the arrow, how confident would you be to call this a G versus an A? According to the quality value score (~10), there is a 1 in 10 chance this was called incorrectly! This is why only high-quality data are used for determining barcodes. This DNA sample probably had contamination from another sample, which led to two PCR products being generated and then being sequenced at the same time.

BOLD-SDP also computes quality statistics and displays them in tabular and graphical format above each trace file window (see Figure 5). Of these statistical values, the mean and standard deviation are the most informative. The mean refers to the average quality score for the base calls in a given trace file. The standard deviation (Stdev) is a measure of how close the quality scores for the base calls are to the mean. A low standard deviation value indicates that the quality scores are clustered near the mean, whereas a high standard deviation value indicates that the quality scores are dispersed over a large range of values. Lower standard deviation values therefore indicate a greater level of consistency in the quality of base calls, which imparts a higher degree of confidence in the overall accuracy of the trace file. The bar charts that appear above each trace file window show the percentage of base calls that correspond to different quality scores. The data displayed in these charts provide an indication of the range of quality scores for the base calls.

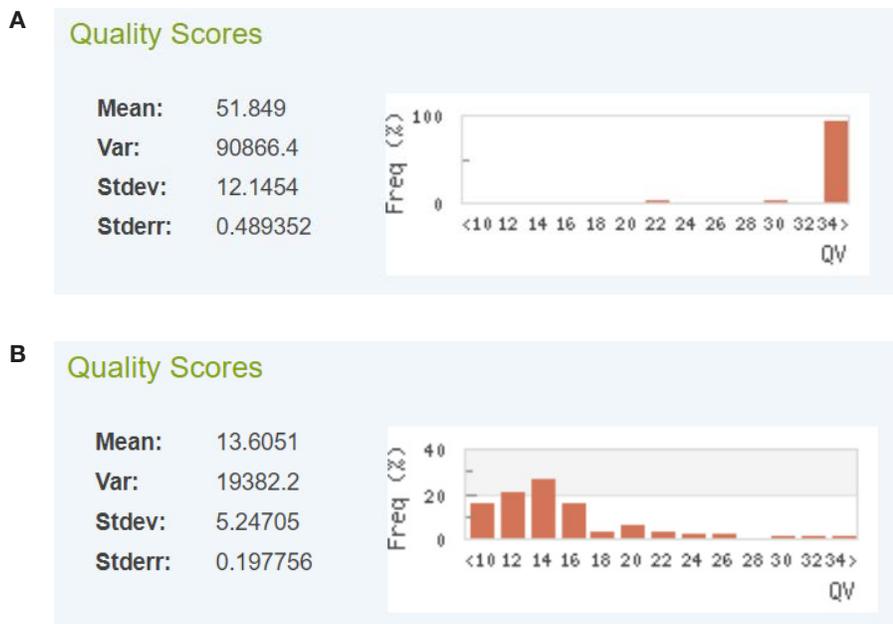
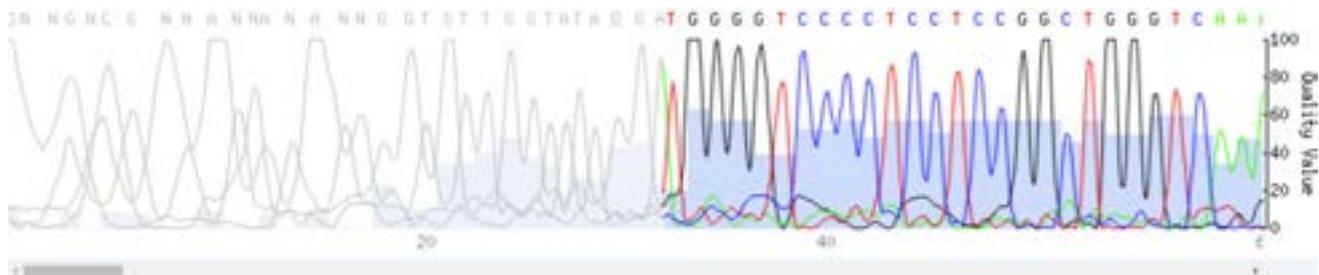


Fig. 5. Histograms and statistics for examples of forward and reverse sequencing trace quality scores. **A**, high-quality sequencing traces; **B**, a failed (low-quality) sequencing trace. Notice that the mean quality score for the high-quality sequencing traces is >50, which means that on average, there is less than a 1 in 100,000 chance that the base calls are incorrect. For the low-quality sequencing trace, the average quality score is 13 and this would mean that on average there is a 1 in 20 chance that any particular base call is incorrect. The standard deviations are comparable for both the high-quality and low-quality trace files, which means that overall, the quality is either mostly good (for the high-quality traces) or mostly bad (for the low-quality trace).

Record the mean quality scores and standard deviations for your first sample, then open up the trace and follow the instructions in steps f-h above for your second sample.

	Mean Quality Score	Standard Deviation
Sample 1: Forward sequence	_____	_____
Sample 1: Reverse sequence	_____	_____
Sample 2: Forward sequence	_____	_____
Sample 2: Reverse sequence	_____	_____

The scroll bar at the bottom of each trace file window allows you to examine the sequences along their entire length. Even in the absence of quality values for individual base calls in the trace files, the quality of the base calls can be inferred from the resolution of their corresponding peaks. Notice that the peaks in the beginning (5' end) of the forward trace file are broad, overlapping, and poorly resolved. This gray region of the trace file corresponds to low-quality base calls, which correlate with high error probabilities.



As you scroll to the right, the non-gray peaks appear sharp, well resolved, and non-overlapping. This region of the trace files corresponds to high-quality base calls, which correlate with low error probabilities. As you scroll even further to the right (toward the 3' end of the trace files), the peaks become lower in amplitude and begin to broaden and overlap. This gray region of the trace file corresponds to low-quality base calls.

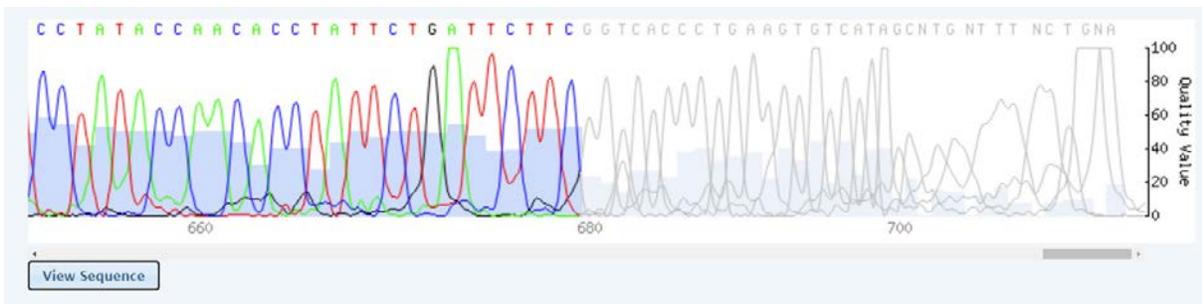
The low-quality base calls that appear at the beginning and end of a trace file arise from technical limitations of dye terminator sequencing. These limitations are caused by and complicated by a variety of different interacting factors associated with capillary electrophoresis and the underlying chemistry of this particular sequencing method. Regardless of their causes, low-quality base calls must be eliminated from the sequence in order to preserve its overall accuracy.

Step 2. Querying either the BOLD database or GenBank for matches to forward and reverse sequencing data

You have now reviewed your traces and assembled a contig. Now it is time to compare your contig sample sequencing trace to sequences that have been deposited into two different databases. The first is BOLD, which contains sequences for vouchered samples that require an extremely high level of quality to be considered for submission. The second is GenBank, which is one of the world’s largest repositories of DNA, RNA, and protein sequencing data. The BOLD database can be searched by using BOLD-SDP and if a sequence match is not found, then GenBank can be searched using BLAST (basic local alignment search tool) programs, which find short regions where pairs of sequences match.

The **blastn** program in GenBank is used to compare a nucleotide sequence to a database of nucleotide sequences. Here blastn will be used to compare your forward and reverse sequences to the GenBank database of all nucleotide sequences. Once the searches are complete, blastn counts all the nucleotides in the matching region and awards two points for every pair of bases that match. If one sequence has an insertion, a deletion, or a gap (more than one base missing) relative to the other, blastn takes points away from the score. The net result is that a blastn score is equal to two times the length of the matching region. The completed search will return a blastn score and E-value for each match of your query sequences relative to sequences in the GenBank database. The results also include an alignment of your sequences to each match in the database so that you can compare them. The meaning of the blastn scoring will be explained in more detail below.

- a. The forward and reverse trace files are displayed in the top and bottom panes of the Trace Viewer page, respectively. Click the **View Sequence** button for your upper sequence.



- b. A text window will display the sequence generated by one of your sequencing reactions. Wherever there is an **n**, the software could not determine what the base call should be for this location.

✕

```

CnnnnnnATTGnnnCCTCTATCTGGTATTCGGTGCCTGAGCCGGCATGGTCCG
TACAGCTTTAAGTTTACTCATCCGAGCAGAAGTCTAGTCAACCCGGTGCTCTTC
TGGGAGACGATCAAATTTATAACGTAATCGTTACAGCGCACGCGTTTGAATA
ATTTTCTTTATAGTAATGCCAATTATGATTGGAGGTTTGGGAACTGACTCATC
CCCTTAATGATCGGGGCTCCCGATATAGCATTCCCCGAATAAACAACATGAG
CTTCTGACTCCTTCCCCTTCATTCTCCTACTTCTGGCCTCTTCAGGTGTTGA
AGCCGGAGCTGGGACGGGTTGGACAGTCTACCCGCCCTAGCCGGCAACCTT
GCTCACGGGGAGCATCCGTAGACTTAACAATTTTCTCCCTTCATTTAGCTGG
GATCTCCTCAATTCTGGGGCTATTAACCTTATCACAACCATCATCAACATAA
AACCCCATGCCGTCTCTATGTACAAATTCCTTATTCTGTTGAGCTGTCTCTG
ATTACGCCGTGCTCCTGCTCCTGCTCACTCCCAGTTTTAGCCGCCGCGCATTAC
AATGCTTCTGACAGCCGAAACTTAAATACTGCCTTCTTTGACCCAGCCGGAG
GAGGGGACCCCATCCTATACCAACACCTATTCTGATTCTTCGGTCACCCTGAA
GTGTCATAGCnTGnTTnCTGnA
                    
```

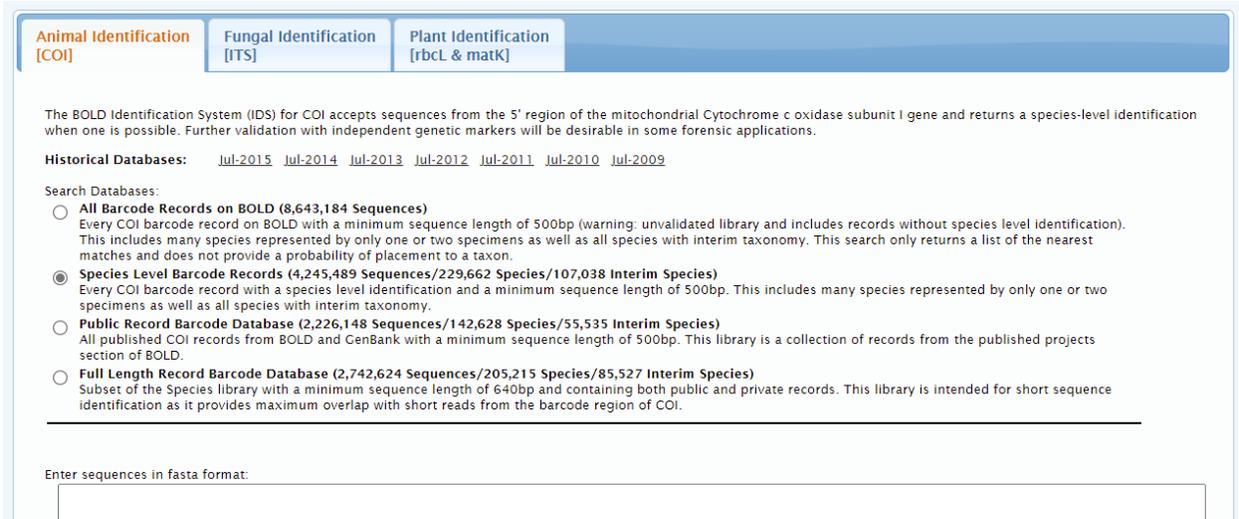
Close

- c. Highlight the text in this window and then copy it by typing **Control+C** (PC) or **Command+C** (Mac).

- d. Click the **Identification** tab.



- e. This will open up a new window for searching the BOLD database.



- f. Choose the correct tab for your sample. For animal samples, select **All Barcode Records on BOLD** and then paste your sequence (**Control +V** on a PC or **Command +V** on a Mac) into the box labeled **Enter sequences in fasta format** and then click **Submit**. A sequence in FASTA format contains a single line of description followed by lines of sequence data. While lacking some of the formatting associated with FASTA format sequences, your pasted nucleotide sequence alone is sufficient for submission.
- g. A window containing the search results will open if a match can be found. Because you submitted sequence data from only a single sequencing reaction, BOLD-IDS may be unable to return a conclusive species-level match for your identification request.



Student Instruction Manual

If a match is found, record the phylum, class, order, family, genus, species (or any level of taxonomical match that was found) and % similarity for the top match to your sequence.

Sequencing trace file name: _____

Phylum: _____

Class: _____

Order: _____

Family: _____

Genus: _____

Species: _____

% Similarity: _____

Repeat steps a–g for your reverse sequencing trace file.

Sequencing trace file name: _____

Phylum: _____

Class: _____

Order: _____

Family: _____

Genus: _____

Species: _____

% Similarity: _____

If you got two matches to the genus or species level, were they the same species and genus? Did they have the same % similarity?

If you have lower-quality data, it might not be possible to find a match using the BOLD database. The example listed above was for high-quality sequences. The one shown below is for low-quality sequences.

SEQUENCING RUNS: Generic Commercial Labs				
Run Date	Direction	Trace File	Seq Primer	Quality
PCR Primers: C_FishF1t1/C_FishR1t1				
<input type="checkbox"/> 2013-01-16	Reverse	Gro-ACLSTW-reverseM13Rev.ab1	M13R	fail
<input type="checkbox"/> 2013-01-16	Forward	Gro-ACLSTW-forwardM13For.ab1	M13F-20	fail

[View Trace Files](#)

In this case, both sequences have failed-quality data. The sequencing trace looks like an overlap of two different sequences, and this means that at some point, contamination occurred. This contamination might have occurred when the fish was prepared for sale, during isolation of a piece, during DNA isolation, during PCR, or during sequencing. Where the contamination occurred cannot be determined, just that there was contamination.

If the sequence for this trace is submitted for identification in the BOLD system, no very strong match is found.

Query: **unlabeled_sequence** Top Hit: **No match**

Search Request:

Type: COI FULL DATABASE (includes records without species designation)

Search Result:

[Tree Based Identification](#)

Similarity scores of the top 99 matches:

TOP 20 Matches:

Phylum	Class	Order	Family	Genus	Species	Subspecies	Similarity (%)	Status
Chordata	Actinopterygii	Perciformes	Serranidae	<i>Hyporthodus</i>	<i>acanthistius</i>		75.74	Private
Chordata	Actinopterygii	Perciformes	Serranidae	<i>Hyporthodus</i>	<i>acanthistius</i>		75.54	Published ↗
Chordata	Actinopterygii	Perciformes	Serranidae	<i>Hyporthodus</i>	<i>acanthistius</i>		75.54	Private
Chordata	Actinopterygii	Perciformes	Serranidae	<i>Hyporthodus</i>	<i>acanthistius</i>		75.54	Private

Display option: Top 20 ▾

The reason for such low matches can be found if a BLAST search is performed using GenBank.

- h. Open a new internet browser window and go to blast.ncbi.nlm.nih.gov/Blast.cgi. Click **Nucleotide BLAST** to open a window to search the nucleotide database (as opposed to a protein database).

NIH U.S. National Library of Medicine
National Center for Biotechnology Information Log in

BLAST® Home Recent Results Saved Strategies Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

A new feature was added to Primer-BLAST. We now offer the ability for user to run primer-blast from NCBJ assembly page..

Tue, 23 Feb 2021 12:00:00 EST [More BLAST news...](#)

Web BLAST



Nucleotide BLAST
nucleotide ▶ nucleotide

blastx
translated nucleotide ▶ protein



Protein BLAST
protein ▶ protein

BLAST Genomes

Enter organism common name, scientific name, or tax id Search

Human Mouse Rat Microbes

Standalone and API BLAST



Download BLAST
Get BLAST databases and executables



Use BLAST API
Call BLAST from your application



Use BLAST in the cloud
Start an instance at a cloud provider

- i. Under **Enter Query Sequence**, paste your sequence into the **Enter accession number(s)** box. Under **Choose Search Set**, make sure **Standard databases** is selected for **Database** and that **Nucleotide collection (nr/nt)** is selected in the dropdown box. All other boxes under **Choose Search Set** can be left blank. Under **Program Selection**, choose **Somewhat similar sequences (blastn)**. Then click the blue **BLAST** button to begin the search.

The screenshot shows the BLAST web interface with the following sections:

- Enter Query Sequence:** Contains a text area with a DNA sequence, a 'Query subrange' field with 'From' and 'To' sub-fields, and an 'Or, upload file' section.
- Choose Search Set:** Includes a 'Database' dropdown set to 'Nucleotide collection (nr/nt)', an 'Organism' field, and checkboxes for 'Exclude' and 'Limit to' options.
- Program Selection:** Features radio buttons for 'Highly similar sequences (megablast)', 'More dissimilar sequences (discontiguous megablast)', and 'Somewhat similar sequences (blastn)', which is selected.
- BLAST Button:** A large blue button labeled 'BLAST' with a summary of the search parameters.

Your results will appear after the database has been searched.

U.S. National Library of Medicine
National Center for Biotechnology Information
[Log in](#)

BLAST® » blastn suite » results for RID-DTXZUEAG013
[Home](#) [Recent Results](#) [Saved Strategies](#) [Help](#)

< Edit Search

[Save Search](#)
[Search Summary ▾](#)

[How to read this report?](#)
[BLAST Help Videos](#)
[Back to Traditional Results Page](#)

Job Title	Nucleotide Sequence
RID	DTXZUEAG013 <small>Search expires on 07-02 07:02 am</small> Download All ▾
Program	BLASTN Citation ▾
Database	nt See details ▾
Query ID	Ic Query_26473
Description	None
Molecule type	dna
Query Length	709
Other reports	Distance tree of results MSA viewer ?

Filter Results

Organism only top 20 will appear exclude

[+ Add organism](#)

Percent Identity to

E value to

Query Coverage to

[Filter](#) [Reset](#)

Descriptions

Graphic Summary

Alignments

Taxonomy

Sequences producing significant alignments
[Download ▾](#) [Select columns ▾](#) Show [?](#)

select all 100 sequences selected

[GenBank](#)
[Graphics](#)
[Distance tree of results](#)
[MSA Viewer](#)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	Epinephelus acanthistius voucher SIO 00-142 cytochrome oxidase subunit I (COI) gene, partial cds, mitochondrial	Hyporthodus aca...	357	357	70%	1e-93	75.75%	603	HQ010051.1
<input checked="" type="checkbox"/>	Hyporthodus acanthistius voucher REDES-OC33 cytochrome c oxidase subunit I (COI) gene, partial cds, mitoc...	Hyporthodus aca...	324	324	53%	9e-84	78.91%	577	MN880575.1
<input checked="" type="checkbox"/>	Hyporthodus acanthistius voucher HA1 cytochrome c oxidase subunit I (COI) gene, partial cds, mitochondrial	Hyporthodus aca...	324	324	53%	9e-84	78.91%	557	MN839521.1
<input checked="" type="checkbox"/>	Hyporthodus acanthistius isolate DF63 cytochrome c oxidase subunit 1 (cox1) gene, partial cds, mitochondrial	Hyporthodus aca...	324	324	53%	9e-84	78.91%	439	MN756312.1
<input checked="" type="checkbox"/>	Epinephelus acanthistius voucher SIO 00-142 cytochrome oxidase subunit I (COI) gene, partial cds, mitochondrial	Hyporthodus aca...	291	291	49%	5e-74	78.37%	528	GU125715.1
<input checked="" type="checkbox"/>	Epinephelus chabaudi voucher ADC12_166_37 #7 cytochrome oxidase subunit 1 (COI) gene, partial cds, mitoc...	Epinephelus cha...	287	287	70%	2e-72	72.76%	652	KF489580.1
<input checked="" type="checkbox"/>	Hyporthodus nigritus isolate Hnig_CK1303222 cytochrome oxidase subunit 1 (COI) gene, partial cds, mitochon...	Hyporthodus nig...	286	286	70%	2e-72	72.76%	653	KU739506.1
<input checked="" type="checkbox"/>	Hyporthodus nigritus isolate Hnig_30Oct2015_01 cytochrome oxidase subunit 1 (COI) gene, partial cds, mitoch...	Hyporthodus nig...	286	286	70%	2e-72	72.76%	653	KU739503.1

Understanding blastn results

The results from a blastn search include many different kinds of information and statistics. These bits of information include the size of the database, length of each query sequence, statistics that describe the number and percentage of matching bases, a BLAST score, and the E value.

On the Graphic Summary tab, there is a graphical representation of the results. A thicker bar represents the full length of the query sequence. Below the thick query bar are thinner bars that represent sequences from the database (subject sequences) that align with the query sequence. The description of the subject sequence is given in a box above the graph when the bar is moused over. The colors of these bars show the degrees of alignment using the color key at the top of the Graphic Summary tab. The colors are based on the blastn max scores. The subject sequence with the highest max score aligns at the top.

If the subject sequences do not continuously match the query, the colored bars are connected by thin gray lines representing regions where there is no homology to the query.

On the Descriptions tab there is a table that summarizes the statistics. Each row contains a matching sequence with the best-matching sequence at the top of the table.

Accession — an accession number is the unique identifier given to a DNA sequence when it is submitted to a database. (It can also refer to a submitted protein sequence.) The submitted data can be for an entire genome, a chromosome within a genome, an entire mitochondrial genome (such as JX135579.1 for *Hyporthodus octofasciatus*) or it can be for shorter sequences such as the sequences for vouchered samples (such as JF493437.1 for *Epinephelus chabaudi*).

As previously mentioned, though GenBank is one of the world's largest repositories of DNA, RNA, and protein sequencing data, this database lacks the same stringency employed by the BOLD database. BOLD requires additional data to be included with reference sequences published in its database to ensure accuracy of the species identified within it, while GenBank does not. For instance, within GenBank, some sequences may be labeled as voucher specimens, yet they do not adhere to the requirements for vouchered sample data as set forth by BOLD. The best way to determine whether a sequence match in GenBank has truly adhered to the stringent requirements of the BOLD database is to examine the Keywords line in the accession entry. Entries that contain the word BARCODE in the Keywords line adhere to BOLD data standards, while entries lacking this designation may not adhere to BOLD data standards.

This sequence for *Epinephelus acanthistius* does not contain the BARCODE designation, though the name of the entry includes the word voucher in its description. This sequence may not have the additional data required by BOLD to truly serve as a reference barcode for species identification.

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Search Advanced Help

GenBank Send to

Epinephelus acanthistius voucher SIO 00-142 cytochrome oxidase subunit I (COI) gene, partial cds; mitochondrial

GenBank: HQ010051.1
[FASTA](#) [Graphics](#)

[Go to:](#)

LOCUS HQ010051 603 bp DNA linear VRT 23-AUG-2010
 DEFINITION Epinephelus acanthistius voucher SIO 00-142 cytochrome oxidase subunit I (COI) gene, partial cds; mitochondrial.
 ACCESSION HQ010051
 VERSION HQ010051.1
 KEYWORDS .
 SOURCE mitochondrion Hyporthodus acanthistius (rooster hind)
 ORGANISM [Hyporthodus acanthistius](#)
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Actinopterygii; Neopterygii; Teleostei; Neoteleostei; Acanthomorpha; Eupercaria; Perciformes; Serranoidei; Serranidae; Epinephelinae; Epinephelini; Hyporthodus.
 REFERENCE 1 (bases 1 to 603)
 AUTHORS Gleason,L.U., Walker,H.J., Hastings,P.A. and Burton,R.S.
 TITLE Establishing a DNA Sequence Database for the Marine Fish Fauna of California
 JOURNAL Unpublished
 REFERENCE 2 (bases 1 to 603)
 AUTHORS Gleason,L.U., Walker,H.J., Hastings,P.A. and Burton,R.S.
 TITLE Direct Submission
 JOURNAL Submitted (30-JUL-2010) Marine Biology Research Division, Scripps Institution of Oceanography, 9500 Gilman Drive, La Jolla, CA 92093, USA

Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

Related information

Protein

Taxonomy

Recent activity [Turn Off](#) [Clear](#)

- Epinephelus acanthistius voucher SIO 00-142 cytochrome oxidase subunit I (C Nucleotide
- Hyporthodus octofasciatus mitochondrion, complete genome Nucleotide
- Entrez Sequences Quick Start - Entrez Sequences Help

This sequence for *Epinephelus chabaudi* does contain the BARCODE designation, and this indicates that the entry conforms to BOLD data standards.

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Search Help

Advanced

GenBank Send to Change region shown Customize view

Epinephelus chabaudi voucher Smith166.37-1 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial

GenBank: JF493437.1
[FASTA](#) [Graphics](#)

Go to:

LOCUS JF493437 637 bp DNA linear VRT 25-JUL-2016
 DEFINITION Epinephelus chabaudi voucher Smith166.37-1 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial.
 ACCESSION JF493437
 VERSION JF493437.1
 KEYWORDS BARCODE.
 SOURCE mitochondrion Epinephelus chabaudi (moustache grouper)
 ORGANISM [Epinephelus chabaudi](#)
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Actinopterygii; Neopterygii; Teleostei; Neoteleostei; Acanthomorpha; Eupercaria; Perciformes; Serranoidei; Serranidae; Epinephelinae; Epinephelini; Epinephelus.
 REFERENCE 1 (bases 1 to 637)
 AUTHORS Steinke,D., Zemlak,T.S., Connell,A.D., Heemstra,P.C. and Hebert,P.D.N.
 TITLE DNA Barcodes: Linking adults and immatures of marine fishes
 JOURNAL Unpublished
 REFERENCE 2 (bases 1 to 637)
 AUTHORS Steinke,D., Zemlak,T.S., Connell,A.D., Heemstra,P.C. and Hebert,P.D.N.

Analyze this sequence
 Run BLAST
 Pick Primers
 Highlight Sequence Features
 Find in this Sequence

Related information
 Protein
 Taxonomy
 Trace Archive

Recent activity
 Turn Off Clear
 Epinephelus chabaudi voucher Smith166.37-1 cytochrome oxidase subunit 1 (CC Nucleotide)

Description — the description refers to the source of the matching sequence. In the case of the *Hyporthodus octofasciatus*, or eightbar grouper, sequence match, the complete mitochondrial genome of this species of grouper has been sampled. Other sequences are from *COI* barcodes and come from different species, such as *Epinephelus acanthistius*, rooster hind (HQ010051.1), *Epinephelus chabaudi*, moustache grouper (JF493437.1), *Epinephelus ergastularius*, sevenbar grouper (DQ107881.1), and *Epinephelus albomarginatus*, white-edged grouper (GU804970.1).

Max score — each of the colored bars in the BLAST alignment graph (at the top of the BLAST search results page) have been assigned a score based on the extent of the match. The max score comes from the block of aligned sequence that had the highest score. Because the blastn score is about twice the number of matching nucleotides, it is possible to infer that the maximum score of 356 for the top sequence represents either approximately 700 matching bases or a longer region that contains gaps.

Total score — the total score is obtained by adding the scores from the region of the query sequence that matches any region on the sequence in the database. In this example, since the *COI* gene is a mitochondrial gene, there should not be long gaps of nonmatching sequence followed by large stretches of matching sequence, so this score will be comparable to the max score. Total score is more important when trying to match different genomic DNA sequences that include both exon sequences expected to have higher similarity and intronic sequences that are not expected to have much similarity between different species. Since there are no introns in mitochondrial genes, the total score should be the same as the max score for your DNA barcode sequences.

Query coverage — the query coverage corresponds to the fraction of the entire query sequence that is matched by parts of the subject sequence. In this case for the top match, 70% of the query matches the subject sequence (the sequence in GenBank). The query that was submitted was 709 bases long and 482 of these bases were found to align with the subject sequence in the database.

E-value — the E-value is reported as a power of 10 (expressed as an exponent; for example e-2 means 10^{-2}). For each subject sequence (match in the database), the E-value represents the number of equally good matches to the query sequence that would be expected in a database of the same size containing random sequences. When E-values are below 1, they can be understood as the probability that two sequences will match.

This would mean that with an E value of 0.01, there is a 1% chance of finding an equally good match in a database of random sequences. While low E-values are good, high E values suggest that it is possible to find an equally good match by random chance. In the top row of this example, the E-value is $1e-94$. This means that there is an essentially 0% chance of finding this match in a database of random sequences. In other words, a match is statistically very unlikely to occur by random chance.

Two additional factors have a strong influence on E-values: the length of the sequence, because it is easier to find a perfect match to a shorter sequence than it is to a longer sequence, and the size of the database, because it is easier to find a match in a larger database than it is in a smaller one.

Percent identity — this column shows the block of a sequence that has the highest percentage of matching bases. In this example, the maximum identity of any matching block is 76% with the *Epinephelus acanthistius* voucher *COI* sample (HQ010051.1). This sequence has 381 of 503 aligned bases matching and three gaps as well. This can be seen if you scroll down the GenBank search page until you reach this first match.

Links — the final column in the blastn alignment table contains links to other databases that are identified in a key above the table on the BLAST results page. In this example, there are no links to other databases.

On the Alignments tab, the sequence alignments are organized by subject sequence, with all the regions that match one subject sequence grouped together. The sets of alignments are presented in order of maximum score, with the set containing the longest and best alignment shown first.

- j. Record the top three species that align with your sequence and record each max score, query coverage, E value, and percent identity.

Species	Max Score	Query Coverage	E Value	Percent Identity

The take-home message from the example search is that the best match found was the same as that found in BOLD, but it should be noted that this search covered only 70% of the submitted sequence, and of that 70%, the maximum identity was only 76%. So, for a sequence containing 709 base pairs (bp), only 496 bp were used to make a match, and of those, only 381 did actually match the best case match. Based on this match, would you feel great confidence in concluding that the fish was the specific species of grouper known as the rooster hind? How confident would you be in concluding this was a member of the grouper genera?

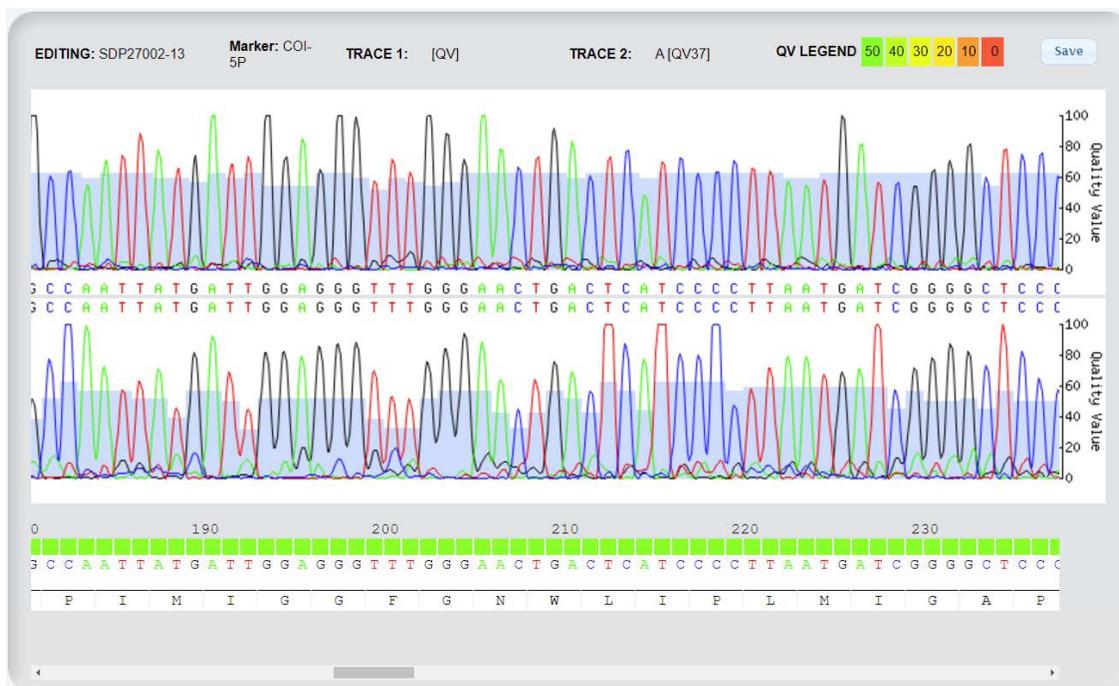
Step 3. Assembling contigs from trace files

Note: Steps 3–4 must be completed in their entirety.

Although low-quality or ambiguous base calls are normally found at the 5' and 3' ends of a trace file, they may also appear elsewhere in the sequence. If only a single trace file was generated for a given amplicon, it would be difficult or impossible to confidently determine the identity of a base call that is assigned a low quality score value (that is, a value <20). However, a second trace file contains duplicate data that can help determine its identity with a greater level of statistical certainty.

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 algorithm that operates within the BOLD-SDP Sequence Editor (and the Trace File Viewer described above) automatically reverses the sequence of base calls and peaks in the reverse trace file (which corresponds to the sequence of the antisense strand) so that they read in the opposite direction. It then converts each base call to its complementary nucleotide and recolors the corresponding peaks accordingly. This conversion therefore displays the reverse complement of the sequence read on the reverse trace file. For example, a base call of T that appears at the first position of the trace file above a red peak is replaced with a base call of A and moved to the last position above a green peak of the same shape and height as the original red peak. The quality value assigned to the original base call is also shifted to the last position. Next, the program aligns this sequence of complementary base calls and appropriately recolored peaks with the unaltered sequence of base calls and peaks of the forward trace file (which corresponds to the sequence of the sense strand). The largely overlapping DNA sequences are displayed in a project window of the BOLD-SDP Sequence Editor, as shown below.



The forward trace file appears in the top pane of the Online Sequence Editor window along with its sequence of base calls. The reverse complement of the reverse trace file appears in the lower pane along with its corresponding base calls. For both trace files, quality scores are represented graphically in the form of a histogram, where higher bars indicate higher quality scores and vice versa. The vertical scale on the right side of each trace file histogram displays the numerical quality values.

The nucleotide sequence at the bottom of the project window represents a contig – a continuous nucleotide sequence assembled from two overlapping DNA sequences (in this case, from the forward trace file and the reverse complement of the reverse trace file). The BOLD-SDP 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 bars that appear above each nucleotide in the contig are graphical representations of quality values, which are color-coded according to the legend in the upper right-hand corner of the window.

It is important to realize that the algorithm utilized by BOLD-SDP to make these comparisons is not perfect, so you must scan through the contig to ensure that no errors were made. This task is simplified by examining the quality scores that appear over each nucleotide in the contig. These scores represent the algorithm's confidence that the correct base call was chosen for inclusion in the contig. Low quality scores flagged with orange or red bars require human inspection.

To assemble contigs for your forward and reverse trace files, please follow the steps outlined below:

- a. Go to the Main Student Console page of BOLD-SDP at: boldsystems.org/index.php/SDP_Home and click the **Students** button to access the Student Login page.

BOLD SYSTEMS-SDP
STUDENT DATA-PORT

Home | Explore | Identification | Quick Start | F.A.Q.

An integrated workbench supporting the assembly, analysis and publication of DNA barcode data by students.

- System Overview
- Quick Start Guide
- Video Introduction

The BOLD Student Data Portal (BOLD-SDP) is a classroom-focused interface to the BOLD database. This platform provides instructors and students with the tools necessary to make contributions to the DNA barcode library used for identifying species. It also gives students the opportunity to integrate and analyze specimen and sequence data while providing instructors with tools to monitor student progress and evaluate their work. Students can explore the large database of DNA barcode records submitted by scientists around the world, but they will also be able to add their own data to the ever-growing DNA barcoding library. Contact us for support in using this system in a classroom setting (edu@boldsystems.org).

Students

A portal for students to assemble DNA barcode records in a simplified but powerful data collection and analysis environment.

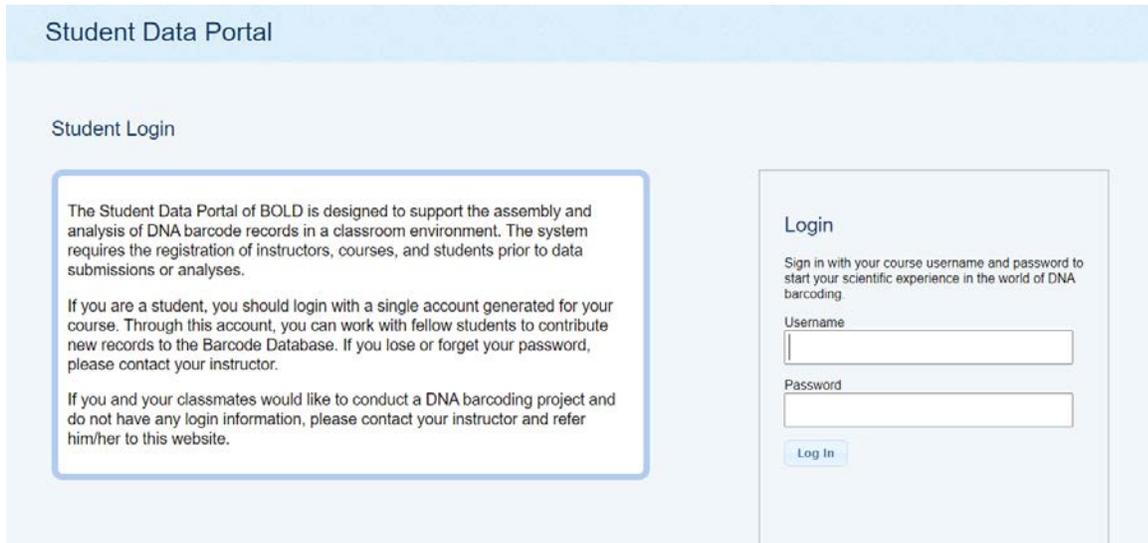
Instructors

A management interface for instructors to register students and classrooms; review student work and utilize data validation workflows.

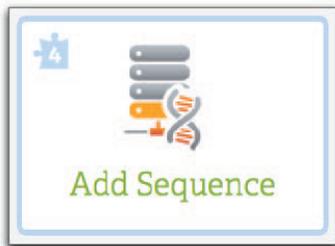
Explore

Search for and browse data records, images, and geographic distributions for species being barcoded.

- b. Enter the course username and password in the appropriate spaces and click **Log In** to enter the Main Student Console page. Please note that the password is case sensitive!

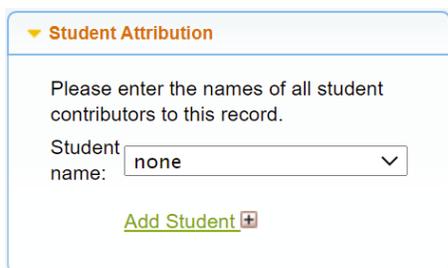


- c. On the Main Student Console page, click the **Add Sequence** icon.



- d. The Add Sequence Page will open. For each student group, you will need to enter multiple pieces of information.

Using the **Student Attribution box**, choose the student(s) who worked on this sample. If more than one student worked on this sample, click the **Add Student** button and select his/her name from the dropdown menu. Repeat the **Add Student** until all students who worked on this sample have been added.



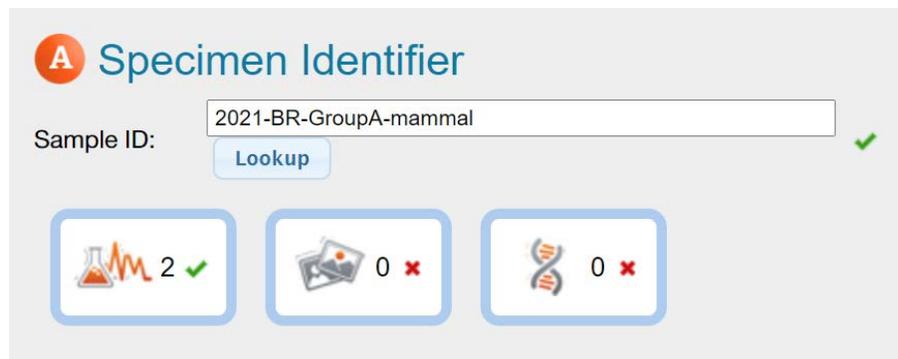
- e. In Section A of the Add Sequence page, type or copy and paste a sample ID that corresponds to the trace files that you wish to assemble and edit and then press the **Tab** key. If you forgot the sample ID, click **Lookup** to find its ID in your class record list. Record the sample IDs for your two samples below.

Sample 1 ID: _____

Sample 2 ID: _____



If you have typed in the sample ID correctly, you should see icons such as the ones below appear, once the system is ready to process your sequence trace files. There should also be a green check mark next to your sample ID entry. In the example below, the first icon means that there are two sequencing trace files for this sample ID. The second icon means that there are no photographs entered for this particular sample ID. The third icon means that there are no contig sequences generated for these two trace files at this point.



- f. In Section B of the Add Sequence page, click the **BOLD Sequence Editor** button to load the trace files associated with the specimen/sample into the BOLD-SDP Sequence Editor. If the BOLD-SDP software can align your sequences to produce a contig, a new window, the Online Sequence Editor, will appear.



If a contig cannot be generated, then you will get the following error message: **Assembly Failed! Your traces failed to produce a high-quality alignment.** This may be the result of low-quality traces or because the traces files are not of the same gene fragment

There are several possible reasons why a contig cannot be generated. One is that the uploaded sequence trace files did not come from the same sample. Make sure that your two sequencing traces are in the correct specimen ID folder. Another reason is lower quality data. Assembly programs are algorithms that have defined alignment parameters, such as how many base pair differences between the two sequences can be allowed before alignment is deemed a failure, or how many base pairs in a row must align in order to be confident that the contig sequence represents the best data. If a contig cannot be formed, you can speak with your instructor about possibly working on classmates' data along with them.

The BOLD-SDP Sequence Editor simplifies the editing process by automatically eliminating continuous stretches of low-quality base calls from the contig. This process is known as trimming. It is important to realize that although these base calls are not included in the contig, they are still displayed in the forward and reverse trace files and colored gray.

The scroll bar at the bottom of the assembly project window allows you to examine the trace files and contig along their entire lengths (moving from 5' to 3'). In the BOLD-SDP Sequence Editor, start by scanning the entire length of the assembly to identify low-quality bases, which are flagged with orange or red bars above the consensus sequence. Moving the mouse pointer over a base call in the trace files or consensus sequence will highlight the alignment position and display the base calls and associated quality scores/values at the top of the editor. Clicking a base will expose the editing tool, which enables a base call to be revised or made ambiguous. Do this by selecting one of the six options in the dropdown menu (**A**, **T**, **C**, **G**, **N**, or **-**).

Note: You cannot delete or insert a base call using this software; you can only change a base call to an N (ambiguous) or – (which serves the same purpose). This is because the program screens base calls and the codons they represent against known sequences.



- g. The first step in the editing process is to carefully inspect the color of the bars over each nucleotide in the contig (the consensus sequence the BOLD-SDP program generated by determining the best sequence data from the forward and reverse sequencing reactions), starting from the 5' end (left side).
- h. The bars are graphical representations of quality values, which are color-coded according to the legend in the upper right-hand corner of the window. It is important to watch for quality scores <20 (which are indicated by orange and red bars). If you discover an orange or red bar in the contig, **highlight the nucleotide** beneath it with your mouse. Notice that the corresponding base call in each trace file is also highlighted. Next, carefully inspect the quality scores for the corresponding base calls in both trace files. If the quality score for the base call is >20 in at least one trace file, the nucleotide in the contig can be regarded as correct.

It can be highly subjective deciding the base call for bases at the 5' and 3' ends of the sequence where there might not be overlap between the two sequencing files. It can also be highly subjective to make base calls where both sequencing reactions have yielded low-quality sequences. If evidence points to the base call being wrong, click the **Edit Base** box dropdown menu and choose the base call you feel is more appropriate.



Note: Changes you make will apply only to the Contig sequence. You cannot change the raw data in your original trace files. Should you make a change that you later do not feel confident about, you can perform the Sequence function again and the sequence trace files that appear will be your original raw data; you will need to redo all base calls you want to make for the final contig.

Step 4. Inspecting contigs for the presence of stop codons, trimming primers, and checking for contaminants

COI is a mitochondrial gene that directs the production of a protein subunit vital for cellular respiration. All mitochondrial protein-coding genes terminate in a stop codon — a triplet nucleotide that may take one of several forms depending on the taxon. During the process of transcription, the stop codon of a protein-coding gene is transcribed into messenger RNA. At the conclusion of translation, the stop codon binds a release factor, which signals the ribosome to dissociate and release the newly synthesized amino acid chain.

The 650 bp region of the *COI* gene that you amplified by PCR is located upstream of the stop codon found in the mitochondrial DNA template. Accordingly, stop codons should be absent in your edited contig. The presence of a stop codon indicates one of three likely possibilities: 1) a nucleotide was erroneously omitted in the contig, 2) an extra nucleotide was erroneously included, or 3) a base call was incorrectly made. Unlike *COI*, the ITS region that is used to barcode fungi does not code for a protein product and so the presence of a stop codon does not indicate a problem in the sequence data.

The BOLD-SDP Sequence Editor enables you to examine your sequence for the presence or absence of stop codons in a *COI* sequence — this is unnecessary for ITS sequences for the reason mentioned above. Because the *COI* barcode region that you amplified is also downstream of the start (ATG) codon found in the mitochondrial DNA template gene, the Auto Translator algorithm built into the BOLD-SDP Sequence Editor must first organize your contig into three reading frames. For reading frame 1, nucleotides are grouped into codons beginning with the first nucleotide in the contig. For reading frame 2, nucleotides are grouped into codons beginning with the second nucleotide in the contig (the first nucleotide is ignored). For reading frame 3, nucleotides are grouped into codons beginning with the third nucleotide in the contig (the first and second nucleotides are ignored). The translator then uses a translation matrix similar to a genetic code table used in classroom settings to determine the amino acid sequence of each reading frame. It then compares the three amino acid sequences to a database of known *COI* amino acid sequences to determine which reading frame is correct. The correct amino acid sequence is displayed at the bottom of the sequence editor project window.

- Examine your contig sequence for stop codons represented by an *. Look carefully at the three base calls that translate to a stop codon. If any of them are ambiguous, you can change them to the correct base call. It is possible to change all three base calls for the stop codon to an N and this should change the translation from a stop codon to an X. The X is not a stop codon, but it also does not stand for a specific amino acid. It just represents an unknown call.
- If no stop codons were detected in the amino acid translation of the contig, then click the **Save** button. A window will appear that lets you know that your edited sequence has been saved and you should proceed with the sequence uploader to submit your sequence to BOLD. Click **OK**. BOLD-SDP will take you back to the Add Sequence work page.



Your contig sequence should now be shown in Box B, Add Sequence.

B Add Sequence

```

TGGCACCTGTACCTACTATTTGGTGCCTGAGCAGGAATAGTGGGCACTGCCTTGAGCCT
ACTAATTCGCGCTGAAC TAGGTCAGCCGGAACCTACTTGGCGATGATCAAATCTATAA
TGTAATTGTTACAGCTCATGCCTTTGTAATAATCTTCTTTATAGTAATACCCATTATGAT
TGGGGGTTTTGGTAACTGACTCGTACCCTAATAATCGGAGCTCCCGATATGGCCTTCC
ACGTATAAACACATAAGTTTCTGACTACTTCCACCATCCTTCTATTACTACTGGCATC
CTCAATAGTAGAAGCCGGGGGGTACTGGATGAACCGTATACCCACCTTTAGCTGGAAA
CTTAGCCCATGCAGGAGCTTCAGTTGATCTAACAATTTTCTCCCTACACCTTGCAAGTGT
ATCATCAATCCTAGGGGCTATTAATTTCAATACCACAATTATTAACATAAAACCTCCCGC
AATGTCTCAATACCAACACCCCTGTTTGTCTGATCAGTACTAATCACAGCCGTACTACT
TCTACTATCCCTGCCAGTTCTAGCAGCTGGCATTACTATACTACTGACAGACCGCAACCT
GAACACAACCTTTTTTATCCAGCAGGTGGTGGAGACCTATCCTTTATCAACACTTGTT
CTG
                    
```

- c. Next, you will trim off the sequences at the 5' and 3' ends of your contig that correspond to the PCR primers. If the PCR primers were entered correctly when your sequencing trace files were uploaded, you should see a list of primer sequences. These were the primers that were mixed with PCR master mix for your PCR reaction. Automatically trim the primer sequences from your edited sequence by clicking the **Trim Primers** button in Section C, Process Sequence.

C Process Sequence

i) Trim Primers from Assembled Sequence

Primers used

- VF2_t1: TGTAAAACGACGGCCAGTCAACCAACCACAAAGACATTGGCAC
- FishF2_t1: TGTAAAACGACGGCCAGTCGACTAATCATAAGATATCGGCAC
- FishR2_t1: CAGGAAACAGCTATGACACTTCAGGGTGACCGAAGAATCAGAA
- FR1d_t1: CAGGAAACAGCTATGACACCTCAGGGTGTCCGAARAAACARAA

Trim Primers

- d. Once BOLD-SDP performs the trimming function, you can see which of the primers contained in the mixed primers you used for PCR actually matched best for your sample.
- e. Now it is time to check whether any contaminants were present in your sample. BOLD has a list of standard contaminant sequences from bacteria that may have been PCR-amplified instead of your sample DNA. This has been a common issue in barcoding samples and, in fact, GenBank has several sequences attributed to fish that are actually marine bacteria. Click the **Check for Contaminant** button to inspect your sequence for the presence of common lab contaminants, including human contaminants.

C Process Sequence

i) Trim Primers from Assembled Sequence

Primers used

- VF2_t1: TGTAAAACGACGGCCAGTCAACCAACCACAAAGACATTGGCAC
- FishF2_t1: TGTAAAACGACGGCCAGTCGACTAATCATAAAGATATCGGCAC
- FishR2_t1: CAGGAAACAGCTATGACACTTCAGGGTGACCGAAGAATCAGAA
- FR1d_t1: CAGGAAACAGCTATGACACCTCAGGGTGTCCGAARAAYCARAA

Trimmed Sequence

```

-----TCGGCACCTTTATCTAGTATTTG
TGTAAAACGACGGCCAGTCGACTAATCATAAAGATATCGGCAC-----
GTGCCTGAGCCGGTATAGTAGGCACAGCCCTCAGCCTACTCATTGAGCAGAACTAAGCC
-----
AACCGGGCCTCTCCTTGGAGACGACCAAATTTATAATGTAATCGTTACAGCACATGCCT
-----
TCGTAATGATTTTCTTTATAGTAATGCCAATTATAATTGGAGGTTTGGAAACTGATTAA
-----
TTCCCTAATGATTGGAGCCCCAGATATAGCATTTCCTCGTATAAATAACATAAGTTTCT
-----
GACTTCTTCCCCTTCTTCTACTACTACTTGCCTTCTGGAGTAGAAGCAGGTGCCG
-----
GAACCGGGTGAACAGTGTACCCGCCCTGGCCGGTAATTTAGCCACGCAGGAGCATCAG
-----
TTGACCTAACAACTTTTCACTTACCTAGCAGGTATTTCTCAATCCTCGGGGCAATCA
-----
ATTTTATTACCACAATTATTAATATGAAGCCCCCTGCCATCTCTCAGTACCAGACGCCCC
-----
TATTTGTATGAGCCGTCTAATTACCGCCGTCTTCTCCTTCTCTCTACCAGTTCTCG
-----
CTGCCGGCATCACAATGCTACTTACCGACCGAAATCTTAATACCACCTTCTTTGACCCGG
-----
CTGGAGGAGGGGATCCAATTTTACCAGCACTTATTCTGATTCTTCG-----
-----TTCTGATTCTTCGGTCACCCCTGAAG
-----
TGTCATAGCTGTTTCTG

```

Check for Contaminant

- f. Once the contaminant check is passed, click the **Submit** button to link the edited and validated barcode sequence to your specimen/sample.

D Submit to BOLD

✓ Primers trimmed

✓ No contaminant

Your sequence can now be submitted to BOLD.

Submit

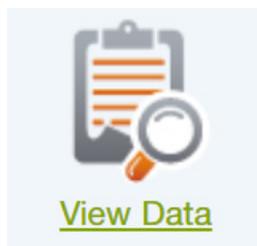
Cancel

Once the sequence has successfully been uploaded into BOLD, you will receive a confirmation page.

Step 5. Reviewing the sample record

The final step in the editing process is to verify that the edited sequence was integrated into the barcode record of the appropriate sample. To perform this function:

- Navigate to the Main Student Console page of BOLD-SDP.
- In the right sidebar of the Main Student Console page, click the **View Data** icon.



- On the Record List page, locate the row for the sample that you linked to the recently uploaded sequence.

Identification	Sample ID	Process ID	Length [Ambig]	Record Flags
<input type="checkbox"/> Chordata	2013-Bio-Rad-CGAK-Salmon	SDP27001-13	652 [0n]	
<input type="checkbox"/> Chordata	2013-Bio-Rad-CGAK-Tuna	SDP27002-13	652 [0n]	
<input type="checkbox"/> Chordata	2013-Bio-Rad-CCTZ-Trout	SDP27003-13	573 [0n]	
<input type="checkbox"/> Chordata	2013-Bio-Rad-CCTZ-CookedSalmon	SDP27004-13	651 [4n]	
<input type="checkbox"/> Chordata	2013-Bio-Rad-ACLSTW-coho	SDP27005-13	0	
<input type="checkbox"/> Chordata	2013-Bio-Rad-ACLSTW-tuna	SDP27006-13	529 [0n]	
<input type="checkbox"/> Chordata	2013-Bio-Rad-ACLSTW-grouper	SDP27007-13	0	
<input type="checkbox"/> Chordata	2013-Bio-Rad-CGAK-pCOIControl	SDP27008-13	655 [0n]	

- Click the Process ID link for the sample to open its sequence page in a new window.

IDENTIFIERS

Sample ID: 2013-Bio-Rad-ACLSTW-tuna
Process ID: SDP27006-13
Identification: Chordata

COI-5P

SEQUENCE DATA

Genbank Accession:
Translation Matrix: Vertebrate Mitochondrial
Last Updated: 2013-03-14

NUCLEOTIDE SEQUENCE

Sequence: 529 bp
 AGAAGCTGAGCGAGCCGGGCGCTCTCTAGGGGATGATCAGATTACAAGTAACTCGTCAC
 AGCCCATGCCCTCGTTATGATTTCTTTATAGTCATGCCGATTATGATCGGAGGCTTTGG
 AAACGTAATAATCCCTAAATGATCGGAGCCCGCTGATGGCATTCCCTGAATAAATAA
 CATAAGCTTCTGACTCCTCCGCCATCTTTCTCTCTCTCTCTCTCTCTGGAAGTTGA
 AGCCGGGCTGGACCGGCTGAACAGTTTATCCCTCTGGCCGGCACTCCGCCAGCG
 AGGAGCTCAGTTGATCTBACTATCTCTCCCTTCATTAGCCGGGATCTCTCAATTT
 AGGAGCCATTAATTTATAGACCACTTAAACATAAAGCCCGCAGTCTCTCTCAGTA
 CCAAGCCCACTTTTGTGAGCTGTGCTAGTCACTGCTCTTCTCTACTACTCTCTCT
 CCCCTTCTGGACGAGGCACTATGATTTACTTAGACCGAATCTA

Composition: A (117), G (98), C (154), T (160)
Ambiguous Characters: 0
Identify Sequence Using:

Full DB
Species DB
Published DB
Full Length DB

ILLUSTRATIVE BARCODE

SEQUENCING RUNS: Generic Commercial Labs

Run Date	Direction	Trace File	Seq. Primer	Quality
PCR Primers: C_FishF1t1/C_FishR1t1				
<input type="checkbox"/> 2012-12-27	Forward	Coho-ACLSTW-For.ab1	M13F-20	med
<input type="checkbox"/> 2012-12-27	Reverse	Coho-ACLSTW-Rev.ab1	M13R	med

[View Trace Files](#)

ANNOTATION

Add Tags & Comments | Comments: 0 | Associated Tags: No Tags

AMINO ACID SEQUENCE

Sequence: 186 residues
 ELSQPGALLGDDQIYVIVTAHAFVHIFVWIPVHIGGGFNNLIPVHIGAPDHAFFRHW
 WNFLLPPSFLLLLSSVIEAGAGTGHVYVPLAGHJAHAGASVDLTFSLHLAGSSSL
 GAINFTIIMKPPAISVQVPLFVHVAVLVAVLLLSLPLVLAAGITMLLDRNL

The edited nucleotide sequence can be found in the Nucleotide Sequence pane along with associated data, including sequence length (in base pairs), sequence composition (that is, the number of A, C, T, and G in the sequence), and the number of ambiguous characters or nucleotides (N).

The amino acid translation and total number of amino acid residues encoded by your nucleotide sequence are located in the lower left pane of the Sequence page in the Amino Acid Sequence pane.

The illustrative barcode in the upper right-hand corner of the Sequence page represents each nucleotide in your barcode sequence as a different colored line. A is represented with green lines, T with red lines, C with blue lines, and G with black lines.

- e. To compare the barcode sequence in your record with other barcode sequences in the BOLD species database, click the **Full DB** button at the bottom of the Nucleotide Sequence pane.

NUCLEOTIDE SEQUENCE

Sequence: **529 bp**

```

AGAACTGAGCCAGCCGGGCGCTCTTCTAGGGGATGATCAGATTACAACGTAATCGTCAC
AGCCCATGCCTTCGTTATGATTTTCTTTATAGTCATGCCGATTATGATCGGAGGCTTGG
AAACTGATTAATTCCCCTAATGATCGGAGCCCTGATATGGCATTCCCTCGAATAAATAA
CATAAGCTTCTGACTCCTTCGCCATCCTTTCTCCTCCTCCTATCTTCTCCTGGAGTTGA
AGCCGGGGCTGGCACCGGGTGAACAGTTTATCCCCCTCTGGCCGGCAACCTCGCCACGC
AGGAGCCTCAGTTGATCTGACTATCTTCTCCCTTCATTTAGCCGGGATCTCCTCAATTTT
AGGAGCCATTAATTTTATTACGACCATTATTAACATAAAGCCCCAGCTATCTCTCAGTA
CCAAACCCCACTTTTGTGGAGCTGTGCTAGTCACTGCTGTTCTTCTACTACTCTCTCT
CCCCGTTCTGGCAGCAGGCATTACTATGTTACTTACAGACCGAAATCTA

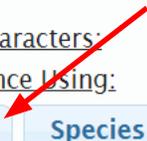
```

Composition: A (117), G (98), C (154), T (160)

Ambiguous Characters: 0

Identify Sequence Using:

Full DB
Species DB
Published DB
Full Length DB



A Specimen Identification Request window will open that contains different forms of information. The Search Result pane near the top of the page contains a summary statement of the search performed by the BOLD Identification System (BOLD-IDS), which is supported by the data displayed in other sections of the page.

Query: SDP27006-13|Chordata Top Hit: Chordata - Salmoniformes - *Oncorhynchus kisutch* (100%)

Search Request:

Type: COI FULL DATABASE (includes records without species designation)

Search Result:

Tree Based Identification

Similarity scores of the top 99 matches:

TOP 20 Matches: Display option: Top 20 ▾

Phylum	Class	Order	Family	Genus	Species	Subspecies	Similarity (%)	Status
Chordata	Actinopterygii	Salmoniformes	Salmonidae	<i>Oncorhynchus</i>	<i>kisutch</i>		100	Published ↗
Chordata	Actinopterygii	Salmoniformes	Salmonidae	<i>Oncorhynchus</i>	<i>kisutch</i>		100	Published ↗
Chordata	Actinopterygii	Salmoniformes	Salmonidae	<i>Oncorhynchus</i>	<i>kisutch</i>		100	Published ↗
Chordata	Actinopterygii	Salmoniformes	Salmonidae	<i>Oncorhynchus</i>	<i>kisutch</i>		100	Published ↗
Chordata	Actinopterygii	Salmoniformes	Salmonidae	<i>Oncorhynchus</i>	<i>kisutch</i>		100	Published ↗
Chordata	Actinopterygii	Salmoniformes	Salmonidae	<i>Oncorhynchus</i>	<i>kisutch</i>		100	Published ↗

The list of the top matches follows in a table.

If a match is found, record the phylum, class, order, family, genus, species (or any level of taxonomical match that was found), and % similarity for the top match to your sequence.

Sequencing trace file name: _____
Phylum: _____
Class: _____
Order: _____
Family: _____
Genus: _____
Species: _____
% Similarity: _____

How does this result compare to the results for your single sequence matches done previously? Would you expect the match to be better using your contig sequence or using single sequences? Why?

If you have <100% similarity with species in the database, what does this mean? What sorts of differences would you expect within species (in other words, between different individuals of the same species), within genera, or within families?

The world map at the bottom of the page shows the collection sites of specimens with sequences that are >98% similar to the barcode sequence of your specimen/sample.

Sampling Sites For Top Hits (>98% Match):



Did the identification match what you thought your species was? If it did not, what was the % similarity?

In addition to comparing your sample to samples in the full database, you can also compare your sample to the species database. The full database has all vouchered samples that have been submitted, but the species database is even more stringent and contains only samples for which both genus and species were confirmed. The full database (All Barcode Records database) is the unvalidated library, while the records in the species database (Species Level Barcode database) are all completely validated. So in terms of levels of stringency for submission, GenBank would be the least stringent, then the All Barcode Records database on BOLD, and finally the Species Level Barcode Records database on BOLD.

- f. From your Sequence page, click the **Species DB** button.

IDENTIFIERS

Sample ID: 2013-Bio-Rad-ACLSTW-tuna
 Process ID: SDP27006-13
 Identification: Chordata

COI-5P

SEQUENCE DATA

Genbank Accession:
 Translation Matrix: Vertebrate Mitochondrial
 Last Updated: 2013-03-14

NUCLEOTIDE SEQUENCE 529 bp

Sequence:
 AGAAGCTGAGCCAGCCGGCGCTCTTCTAGGGGATGATCAGATTTACAACGTAATCGTCAC
 AGCCCATGCCCTCGTATGATTTCTTATAGTCATGCCGATATGATCGAGGCTTTGG
 AAATGATTAATCCCTAATGATGGAGCCCTGATATGGGATCCCTCGAATTAATAA
 CATAAGCTTCGACTCCTCCGCCATCTTCTCCTCCTCTATCTTCTCTGGAGTTGA
 AGCAGGGCTGGCACCGGTGAACAGTTATCCCTCTGGCCGGCAACCTCGCCACGC
 AGGAGCTCAGTGGATCGACTATCTCTCCCTCATTAGCCGGGATCTCTCAATTTT
 AGGAGCCATTAATTTTATAGACCATTAATAACAAGCCCGAGCTATCTCTAGTA
 CCAAACCCCACTTTTGTGAGCTGTGCTAGTCACTGCTGCTCTCTACTCTCTCT
 CCCGTTCTGGCAGCAGGCATTAATGTTACTTACAGACCAAAATCTA

Composition: A (117), G (98), C (154), T (160)
 Ambiguous Characters: 0
 Identify Sequence Using:

Full DB **Species DB** Published DB Full Length DB

AMINO ACID SEQUENCE 186 residues

Sequence:
 ELSQPGALLGDDQIVIVVTAHAFVMIFFHMPIMIGGFNNLIPLMIGAPDMAFPRMWN
 MSFWLLPPSFLLLSSSGVEAGAGTGHVYVPLAGNLAHAGASVDLTI FSLHLAGISSIL
 GAINFIITIIINPKPPAISQYQPLFVNAVLTAVLLLSLPLVLAAGITHLLTDRNL

ILLUSTRATIVE BARCODE

0 199
 200 399
 400 526

SEQUENCING RUNS: Generic Commercial Labs

Run Date	Direction	Trace File	Seq. Primer	Quality
<input type="checkbox"/> 2012-12-27	Forward	Coho-ACLSTW-For.ab1	M13F-20	med
<input type="checkbox"/> 2012-12-27	Reverse	Coho-ACLSTW-Rev.ab1	M13R	med

PCR Primers: C_FishF11/C_FishR111
 View Trace Files

ANNOTATION

Add Tags & Comments Comments: 0 Associated Tags: No Tags

- g. This page will return even more specific data if a match can be found. When searching the species database, the % match at the phylum, class, order, family, genus, and species levels will all be listed.

Query: SDP27006-13|Chordata Top Hit: Chordata - Salmoniformes - *Oncorhynchus kisutch* (100%)

Search Result:

The submitted sequence has been matched to *Oncorhynchus kisutch*. This identification is solid unless there is a very closely allied congeneric species that has not yet been analyzed. Such cases are rare.

A species page is available for this taxon: [Species Page](#)

Closest matching BIN (within 3%): [BIN Page](#)

For a hierarchical placement - a neighbor-joining tree is provided: [Tree Based Identification](#)

Identification Summary:

Taxonomic Level	Taxon Assignment	Probability of Placement (%)
Phylum	Chordata	100
Class	Actinopterygii	100
Order	Salmoniformes	100
Family	Salmonidae	100
Genus	Oncorhynchus	100
Species	Oncorhynchus kisutch	100

Similarly Scores of Top 99 Matches:

TOP 20 Matches: Display option: [Top 20](#)

Phylum	Class	Order	Family	Genus	Species	Subspecies	Similarity (%)	Status
Chordata	Actinopterygii	Salmoniformes	Salmonidae	<i>Oncorhynchus</i>	<i>kisutch</i>		100	Published ↗
Chordata	Actinopterygii	Salmoniformes	Salmonidae	<i>Oncorhynchus</i>	<i>kisutch</i>		100	Published ↗
Chordata	Actinopterygii	Salmoniformes	Salmonidae	<i>Oncorhynchus</i>	<i>kisutch</i>		100	Published ↗
Chordata	Actinopterygii	Salmoniformes	Salmonidae	<i>Oncorhynchus</i>	<i>kisutch</i>		100	Published ↗
Chordata	Actinopterygii	Salmoniformes	Salmonidae	<i>Oncorhynchus</i>	<i>kisutch</i>		100	Published ↗

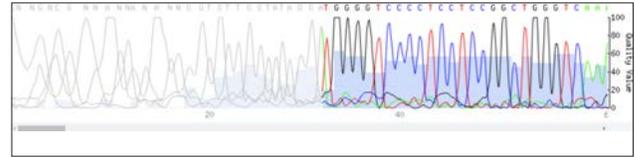
Compare your matches from single sequences using BOLD and/or GenBank versus using the All Barcode Records database and the Species Level Barcode Records database. Do you see any differences? Do they all point to the same genus and species? Did you find any strange matches from any of the databases? Why might those have occurred?

You have now completed the bioinformatics analysis of your samples. You isolated DNA, used PCR to amplify a barcoding gene, analyzed the PCR products using gel electrophoresis, and had your samples purified and sequenced. This is the same workflow performed by researchers participating in the International Barcode of Life project. The difference is the strict control and vouchering of the samples they use and the requirements for the sequencing data before it can be submitted. It is hard to know you have the correct sequence if you are not sure what you started with!

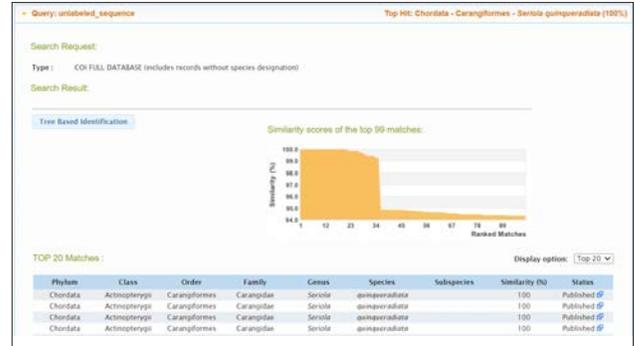
Did you find that the sample you tested was what you expected it to be? If not, do you have enough confidence in your data to determine whether there was a problem in the workflow or if the species of your sample had been misidentified? The more familiar you become with this process, the more confident you will be in your results. Maybe you did have a sample that was called red snapper but was really tilapia — you would not be the first. But hopefully with more barcoding, you will be the last!

Quick Guide

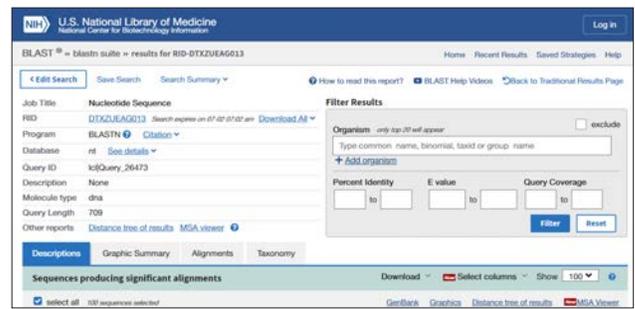
Assess quality of DNA sequencing data (look at chromatograms and examine quality scores)



Query BOLD database and/or GenBank for matches to forward and reverse sequencing data



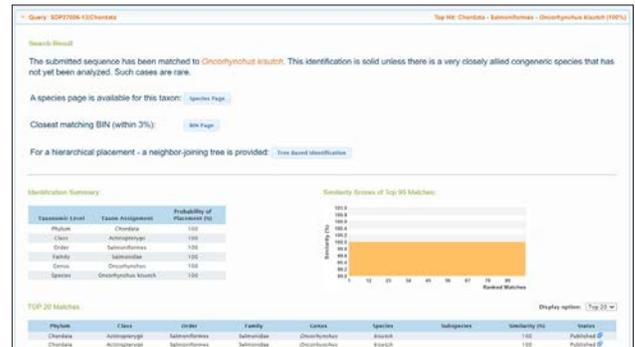
Assemble a single sequence (contig) from forward and reverse sequencing reaction data



Manually compare any base calls that have differences between the forward and reverse sequencing reaction data, check for stop codons, and save contig



Perform a search to determine the identity of your sample contig data



Student login

Course username: _____

Course password: _____

Sample ID for sample 1: _____

Sample ID for sample 2: _____

	Mean Quality Score	Standard Deviation
Sample 1: Forward sequence data	_____	_____
Sample 1: Reverse sequence data	_____	_____
Sample 2: Forward sequence data	_____	_____
Sample 2: Reverse sequence data	_____	_____

BOLD-SDP barcode record search against trace file data

Sample 1:

Forward sequencing trace file name: _____

Phylum: _____

Class: _____

Order: _____

Family: _____

Genus: _____

Species: _____

% Similarity: _____

Reverse sequencing trace file name: _____

Phylum: _____

Class: _____

Order: _____

Family: _____

Genus: _____

Species: _____

% Similarity: _____

Sample 2:

Forward sequencing trace file name: _____

Phylum: _____

Class: _____

Order: _____

Family: _____

Genus: _____

Species: _____

% Similarity: _____

Reverse sequencing trace file name: _____

Phylum: _____

Class: _____

Order: _____

Family: _____

Genus: _____

Species: _____

% Similarity: _____

Blastn database search against trace file data

Top 3 species that align with your sequences for sample 1:

Species	Max Score	Query Coverage	E Value	Percent Identity

Top 3 species that align with your sequences for sample 2:

Species	Max Score	Query Coverage	E Value	Percent Identity

BOLD-SDP database search against contig sequence

Database matches to assembled contig for sample 1:

Sample ID: _____
Phylum: _____
Class: _____
Order: _____
Family: _____
Genus: _____
Species: _____
% Similarity: _____

Database matches to assembled contig for sample 2:

Sample ID: _____
Phylum: _____
Class: _____
Order: _____
Family: _____
Genus: _____
Species: _____
% Similarity: _____

© 2021 Bio-Rad Laboratories, Inc.
BIO-RAD is a trademark of Bio-Rad Laboratories, Inc. All trademarks used herein are the property of their respective owner.



**Bio-Rad
Laboratories, Inc.**

Life Science
Group

Website bio-rad.com **USA** 1 800 424 6723 **Australia** 61 2 9914 2800 **Austria** 00 800 00 24 67 23 **Belgium** 00 800 00 24 67 23 **Brazil** 4003 0399
Canada 1 905 364 3435 **China** 86 21 6169 8500 **Czech Republic** 00 800 00 24 67 23 **Denmark** 00 800 00 24 67 23 **Finland** 00 800 00 24 67 23
France 00 800 00 24 67 23 **Germany** 00 800 00 24 67 23 **Hong Kong** 852 2789 3300 **Hungary** 00 800 00 24 67 23 **India** 91 124 4029300
Israel 0 3 9636050 **Italy** 00 800 00 24 67 23 **Japan** 81 3 6361 7000 **Korea** 82 2 3473 4460 **Luxembourg** 00 800 00 24 67 23
Mexico 52 555 488 7670 **The Netherlands** 00 800 00 24 67 23 **New Zealand** 64 9 415 2280 **Norway** 00 800 00 24 67 23 **Poland** 00 800 00 24 67 23
Portugal 00 800 00 24 67 23 **Russian Federation** 00 800 00 24 67 23 **Singapore** 65 6415 3188 **South Africa** 00 800 00 24 67 23
Spain 00 800 00 24 67 23 **Sweden** 00 800 00 24 67 23 **Switzerland** 00 800 00 24 67 23 **Taiwan** 886 2 2578 7189 **Thailand** 66 2 651 8311
United Arab Emirates 36 1 459 6150 **United Kingdom** 00 800 00 24 67 23

