Using DNA Barcodes to Identify and Classify Living Things

LABORATORY

Using DNA Barcodes to Identify and Classify Living Things

OBJECTIVES

This laboratory demonstrates several important concepts of modern biology. During this laboratory, you will:

- Collect and analyze sequence data from plants, fungi, or animals—or products made from them.
- Use DNA sequence to identify species.
- Explore relationships between species.

In addition, this laboratory utilizes several experimental and bioinformatics methods in modern biological research. You will:

- Collect plants, fungi, animals, or products in your local environment or neighborhood.
- Extract and purify DNA from tissue or processed material.
- Amplify a specific region of the chloroplast, mitochondrial, or nuclear genome by polymerase chain reaction (PCR) and analyze PCR products by gel electrophoresis.
- Use the Basic Local Alignment Search Tool (BLAST) to identify sequences in databases.
- Use multiple sequence alignment and tree-building tools to analyze phylogenetic relationships.

INTRODUCTION

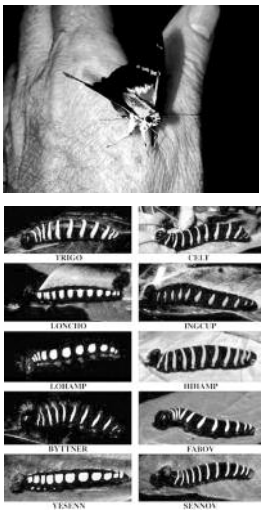
Taxonomy, the science of classifying living things according to shared features, has always been a part of human society. Carl Linneas formalized biological classification with his system of binomial nomenclature that assigns each organism a genus and species name.

Identifying organisms has grown in importance as we monitor the biological effects of global climate change and attempt to preserve species diversity in the face of accelerating habitat destruction. We know very little about the diversity of plants and animals—let alone microbes—living in many unique ecosystems on earth. Less than two million of the estimated 5–50 million plant and animal species have been identified. Scientists agree that the yearly rate of extinction has increased from about one species per million to 100–1,000 species per million. This means that thousands of plants and animals are lost each year. Most of these have not yet been identified.

Classical taxonomy falls short in this race to catalog biological diversity before it disappears. Specimens must be carefully collected and handled to preserve their dis-

tinguishing features. Differentiating subtle anatomical differences between closely related species requires the subjective judgment of a highly trained specialist—and few are being produced in colleges today.

Now, DNA barcodes allow non-experts to objectively identify species—even from small, damaged, or industrially processed material. Just as the unique pattern of bars in a universal product code (UPC) identifies each consumer product, a "DNA barcode" is a unique pattern of DNA sequence that identifies each living thing. Short DNA barcodes, about 700 nucleotides in length, can be quickly processed from thousands of specimens and unambiguously analyzed by computer programs.



DNA Barcoding revealed that what was once thought to be one species of butterfly is really ten species with caterpillars that eat different plants.

The International Barcode of Life (iBOL) organizes collaborators from more than 150 countries to participate in a variety of "campaigns" to census diversity among plant, fungi, and animal groups—including ants, bees, butterflies, fish, birds, mammals, mushrooms, and flowering plants—and within ecosystems—including the seas, poles, rain forests, kelp forests, and coral reefs. The 10-year Census of Marine Life, completed in 2010, provided the first comprehensive list of more than 190,000 marine species and identified 6,000 potentially new species.

There is a surprising level of biological diversity, literally in front of our eyes. For example, DNA barcodes showed that a well-known skipper butterfly (*Astraptes fulgerator*), identified in 1775, is actually ten distinct species. DNA barcodes have revolutionized the classification of orchids, a complex and widespread plant family with an estimated 20,000 members. The urban environment is also unexpectedly diverse; DNA barcodes were used to catalogue 54 species of bees and 24 species of butterflies in community gardens in New York City.

DNA barcodes are also used to detect food fraud and products taken from conserved species. Working with researchers from Rockefeller University and the American Museum of Natural History, students from Trinity High School found that 25% of 60 seafood items purchased in grocery stores and restaurants in New York City were mislabeled as more expensive species. One mislabeled fish was the endangered species, Acadian redfish. Another group identified three protected whale species as the source of sushi sold in California and Korea. However, using DNA barcodes to identify potential biological contraband among products seized by customs is still in its infancy.

Barcoding relies on short, highly variably regions of the genome. With thousands of copies per cell, mitochondrial and chloroplast sequences are readily amplified by polymerase chain reaction (PCR), even from very small or degraded specimens. A region of the chloroplast gene *rbc*L—RuBisCo large subunit—is used for barcoding plants. The most abundant protein on earth, RuBisCo (Ribulose-1,5-bisphosphate carboxylase oxygenase) catalyzes the first step of carbon fixation. A region of the mitochondrial gene *COI* (cytochrome c oxidase subunit I) is used for barcoding animals. Cytochrome c oxidase is involved in the electron transport phase of respiration. Thus, the genes used for barcoding are involved in the key reactions of life: storing energy in carbohydrates and releasing it to form ATP. *COI* in fungi is difficult to amplify, insufficiently variable, and some fungal groups lack mitochondria. Instead, the nuclear internal transcribed spacer (*ITS*), a variable region that surrounds the 5.8s ribosomal RNA gene, is targeted. Like organelle genes, there are many copies of *ITS* per genome, and the variability in fungi allows for their identification.

This laboratory uses DNA barcoding to identify plants, fungi, or animals—or products made from them. First, a sample of tissue is collected, preserving the specimen whenever possible and noting its geographical location and local environment. A small leaf disc, a whole insect, or samples of muscle are suitable sources. DNA is extracted from the tissue sample, and the barcode portion of the *rbc*L, *COI*, or *ITS* gene is amplified by PCR. The amplified sequence (amplicon) is submitted for sequencing in one or both directions.

The sequencing results are then used to search a DNA database. A close match quickly identifies a species that is already represented in the database. However, some barcodes will be entirely new, and identification may rely on placing the unknown species in a phylogenetic tree with near relatives. Novel DNA barcodes can be submitted to GenBank® (www.ncbi.nlm.nih.gov).

FURTHER READING

Benson D.A., Cavanaugh M., Clark K., Karsch-Mizrachi I, Lipman D.J., Ostell J., Sayers E.W. (2013). *Nucleic Acids Res*. GenBank®. 41(D1): D36–D42.

Hebert P.D., Cywinska A., Ball S.L., deWaard J.R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences* 270(1512): 313-21.

Hebert P.D.N., Penton E.H., Burns J.M., Janzen D.H., Hallwachs W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proc Natl Acad Sci U S A*. 101(41):14812-7.

Hollingsworth P.M. et al (2009). A DNA barcode for land plants. *Proc Natl Acad Sci U S A* 106(31): 12794-7.

Ratnasingham, S., Hebert, P.D.N (2007). Barcoding BOLD: The Barcode of Life Data System. *Molecular Ecology Notes* 7(3): 355-64.

Stoeckle M. (2003). Taxonomy, DNA, and the Bar Code of Life. *BioScience* 53(9): 2-3.

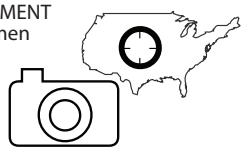
Van Den Berg C., Higgins W.E., Dressler R.L., Whitten W.M., Soto-Arenas M.A., Chase M.W. (2009) A phylogenetic study of laeliinae (*Orchidaceae*) based on combined nuclear and plastid DNA sequences. *Annals of Botany* 104(3): 417-30.

OVERVIEW OF EXPERIMENTAL METHODS

I. COLLECT, DOCUMENT, AND IDENTIFY SPECIMENS

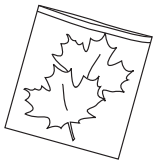
COLLECT specimen

DOCUMENT specimen



IDENTIFY specimen

COLLECT tissue sample



II. ISOLATE DNA FROM PLANT, FUNGAL, OR ANIMAL SAMPLES

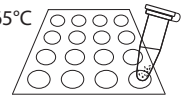
ADD specimen tissue sample

ADD lysis solution

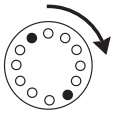
GRIND sample in solution

INCUBATE 10 min

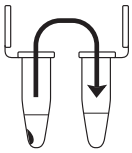
65°C



CENTRIFUGE 1 min

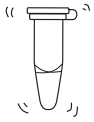


TRANSFER supernatant to fresh tube



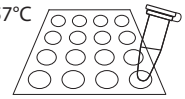
ADD silica resin

MIX

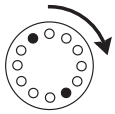


INCUBATE 5 min

57°C



CENTRIFUGE 30 sec

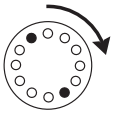


REMOVE supernatant

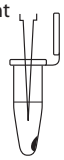
ADD wash buffer

VORTEX

CENTRIFUGE 30 sec



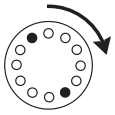
REMOVE supernatant



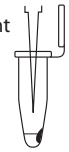
ADD wash buffer

VORTEX

CENTRIFUGE 30 sec



REMOVE remaining supernatant

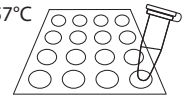


ADD dH_2O

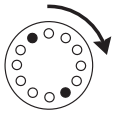
MIX by pipetting in and out

INCUBATE 5 min

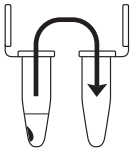
57°C



CENTRIFUGE 30 sec



TRANSFER supernatant to fresh tube



STORE at -20 °C

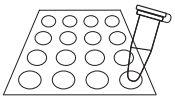
III. AMPLIFY DNA BY PCR

ADD primer mix

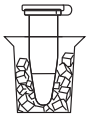
ADD DNA

ADD mineral oil (if necessary)

AMPLIFY in thermal cycler

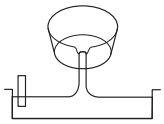


STORE at -20 °C



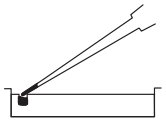
IV. ANALYZE PCR PRODUCTS BY GEL ELECTROPHORESIS

POUR
gel

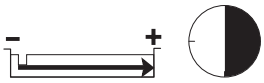


SET
20 min

LOAD
gel

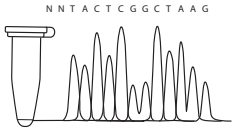


ELECTROPHORESE
130 volts
30 min

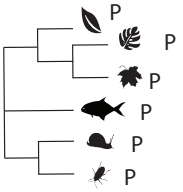


SEQUENCE PCR PRODUCT AND ANALYZE RESULTS

SEND
sample
for
sequencing



ANALYZE
results
using
bioinformatics



EXPERIMENTAL METHODS

I. Collect, Document, and Identify Specimens

The DNA isolation and amplification methods used in this laboratory work for a variety of plants, fungi, and animals—and many products derived from them.

Your collection of specimens may support a census of life in a specific area or habitat, an evaluation of products purchased in restaurants or supermarkets, or may contribute to a larger "campaign" to assess biodiversity across large areas. It may make sense for you to use sampling techniques from ecology. For example, a quadrat samples the plant and/or animal life in one square meter (or ¼ square meter) of habitat, while a transect collects samples along a fixed path through a habitat.

Use common sense when collecting specimens. Respect private property; obtain permission to collect in non-public places. Respect the environment; protect sensitive habitats, and collect only enough of a sample for barcoding. Do not collect specimens that may be threatened or endangered. Be wary of poisonous or venomous plants and animals. Consult your teacher if you are in doubt about the safety or conservation status of a potential specimen. You will also need a small sample for classical taxonomic analysis and to act as a reference sample if you plan to submit your data to GenBank®.

Do not take more sample than you need. Only a small amount of tissue is needed for DNA extraction—a piece of plant leaf about ¼ inch in diameter or a piece of animal or fungal tissue the size of a pencil eraser.

Minimize damage to living plants by collecting a single leaf or bud, or several needles. When possible, use young, fresh leaves or buds. Flexible, non-waxy leaves work best. Tougher materials, such as pine needles or holly leaves, can work if the sample is kept small and is ground well. Dormant leaf buds can often be obtained from bushes and trees that have dropped leaves. Fresh frozen leaves work well. Dried leaves and herbarium samples are variable.

Avoid twigs or bark. If woody material must be used, select flexible twigs with soft pith inside. As a last resort, scrape a small sample of the softer, growing cambium just beneath the bark. Roots and tubers are a poor choice, because high concentrations of storage starches and other sugars can interfere with DNA extraction.

For fungi, obtain fruit bodies (such as mushrooms) when possible, as DNA is easier to obtain from fruiting bodies than mycelia. Only include multiple fruiting bodies in the same sample when they are clearly growing together and appear similar, and avoid contamination by other fungi. Fresh samples work well for DNA isolation, while dried samples give variable results. Fungal fruiting is weather and climate dependent, so their abundance will vary.

Small invertebrate animals, such as insects, can be collected whole and euthanized in a kill jar by placing them in a freezer for several hours. Samples of muscle tissue can be taken from animal foods—such as fish, poultry, or red meat. Blood, internal organs, and bone marrow are all good sources of DNA. Fresh and frozen samples, and those recently preserved in ethanol, work well. However, bone, skin, leather, feather, dessicated, and processed samples are challenging.

REAGENTS, SUPPLIES, & EQUIPMENT

To Share

Collection tubes, jars, or bags
Tweezers, scalpel, and scissors
Smartphone with camera or digital camera with GPS (optional)
Field guide or taxonomic key

1. Collect specimens according to a strategy or campaign outlined by your teacher. "Field Techniques Used by Missouri Botanical Garden" has many good methods for collecting and preparing plant specimens: www.mobot.org/mobot/molib/fieldtechbook/handbook.pdf.

If a camera is not available, make sketches of the location and sample.

2. Use a smartphone or digital camera to photograph your specimen in its natural environment, or where it was obtained or purchased.
   a. Take wide, medium, and close-up views.
   b. Include a person for scale in wide and medium shots. Include a ruler or coin for scale in close-ups.

Please be aware that details described in steps 3 and 4 may change as the devices, software, and websites develop over time.

3. A global positioning system (GPS)-enabled phone or camera stores latitude, longitude, and altitude coordinates along with other metadata for each photo. Visualize or extract this geotag information:
   a. In Apple *iPhoto*, click on "i" (image properties) to plot the photo on a map. Click on "Photo," then "Show extended photo info" to find GPS coordinates.
   b. *GeoSetter*, photo metadata freeware for PCs, will plot your photo on a map.
   c. In Google *Picasa* photo editor, click on "i' to find GPS coordinates.
   d. Your smartphone's manual should explain how to use the GPS feature to obtain coordinates.
   e. Many smartphones also have apps that make it easy to harvest GPS coordinates.

A smartphone app can continuously record your location, making it easy to document a collection trip or a sampling transect.

4. Share your collection location by dropping a pin on a Google map.
   a. Sign in to or create a *Google Maps* account.
   b. Create and name a new map.
   c. Zoom in as much as possible on the collection location.
   d. Click on the pin icon to create a pin, then click the collection location.
   e. Give a title to the pin, and add any collection notes in the description field.
   f. To add a link to a photo or other url, click on the picture icon under the "Rich text" option.
   g. Click on "Done" to save your pin drop.
   h. Click on "Collaborate" or "Share" to share your map with others.
5. Use a field guide or taxonomic key to identify your specimen as precisely as possible: kingdom > phylum > class > order > family > genus > species. Taxonomic keys for local plants, fungi, or animals are often available online, at libraries, or from universities, natural history museums, and botanical gardens.

6. Check to see if your specimen is represented in the Barcode of Life Database, BOLD (www.boldsystems.org) or GenBank® (www.ncbi.nlm.nih.gov):
   a. Search by entering genus and species names in the search bar at top right. If the species is represented in the database, the "Taxonomy Browser" will list the number and sources of specimen records.
   b. Click on "Download Public Sequences" for a fasta file of available barcode sequences.
   c. Click on "Taxonomy Browser" at top left to explore barcode records by group.
7. Use tweezers, scalpel, or scissors to collect a small sample of tissue.
8 Freeze your sample at -20°C until you are ready to begin Part II.

II. Isolate DNA from Plant, Fungal, or Animal Samples

REAGENTS, SUPPLIES, & EQUIPMENT

For each group
(volumes for isolating DNA from 2 samples)

Distilled water (350 µL)
Lysis solution [6 M Guanidine Hydrochloride GuHCl] (700 µL)
Silica resin (10 µL)
Specimen tissue sample(s) (from Part I)
Wash buffer (2.5 mL)

To share

Container with cracked or crushed ice
Microcentrifuge
Microcentrifuge tube rack
6 microcentrifuge tubes (1.5 mL)
Micropipettes and tips (1–1000 µL)
Permanent marker
2 Plastic pestles
Vortexer (optional)
Water bath or heating block at 65°C and 57°C

This standard DNA extraction method is inexpensive and has the advantage of working reproducibly with almost any kind of plant, fungus, or animal specimen.

1. Obtain plant, fungal, or animal tissue ~10–20 mg or ¼ inch diameter from your sample. If you are working with more than one sample, be careful not to cross-contaminate specimens. (If you only have one specimen, make a balance tube with the appropriate volume of water for centrifuge steps.)
2. Place sample in a clean 1.5 mL tube labeled with an identification number.
3. Add 300 µL of lysis solution to each tube.
4. Twist a clean plastic pestle against the inner surface of 1.5 mL tube to *forcefully* grind the tissue for 2 minutes. Use a clean pestle for each tube if you are doing more than one sample.
5. Incubate the tube in a water bath or heat block at 65°C for 10 minutes.
6. Place your tube and those of other groups in a balanced configuration in a microcentrifuge, with cap hinges pointing outward. Centrifuge for one minute at maximum speed to pellet debris.
7. Label a clean 1.5 mL tube with your sample number. Transfer 150 µL of the supernatant (clear solution above pellet at bottom of tube) to the fresh tube. Be

The large end of a 1000 µL pipette tip will punch leaf disks of this size. Animal tissue should be about ¼ the size of a pencil eraser. Using more than the recommended amount can inhibit the DNA extraction or amplification.

Lysis solution dissolves membrane-bound organelles including the nucleus, mitochondria, and chloroplast.

Grinding the tissue breaks up the cell walls and other tough material. When fully ground, the sample should be liquid, but there may be some particulate matter remaining.

careful not to disturb the debris pellet when transferring the supernatant. Discard old tube containing the debris.

Silica resin is a DNA binding matrix that is white. In the presence of the lysis solution the silica resin binds readily to nucleic acids.

8. Add 3 µL of silica resin to tube. Mix well by pipetting up and down. Close and incubate the tube for 5 minutes in a water bath or heat block at 57 °C.

Centrifugation pellets the silica resin, which is now bound to nucleic acid. The pellet will appear as a tiny teardrop-shaped smear or particles on the bottom side of the tube underneath the hinge.

9. Place your tube and those of other groups in a balanced configuration in a microcentrifuge, with cap hinges pointing outward. Centrifuge for 30 seconds at maximum speed to pellet the resin. Use a micropipette with fresh tip to remove all supernatant, being careful not to disrupt the white silica resin pellet at the bottom of the tube.

Wash buffer removes contaminants from the sample while nucleic acids remain bound to the resin. The silica resin is not soluble in the wash buffer. The silica resin may stay as a pellet or break up during the washing.

10. Add 500 µL of ice cold wash buffer to the pellet. Close tube and mix well by vortexing or by pipetting up and down to resuspend the silica resin.
11. Place your tube and those of other groups in a balanced configuration in a microcentrifuge, with cap hinges pointing outward. Centrifuge for 30 seconds at maximum speed to pellet the resin. Use a micropipette with fresh tip to remove all supernatant, being careful not to disrupt the white silica resin pellet at the bottom of the tube.

Washing twice is much more effective than washing once with twice the volume.

12. Once again, add 500 µL of ice cold wash buffer to the pellet. Close tube and mix well by vortexing or by pipetting up and down to resuspend the silica resin.
13. Place your tube and those of other groups in a balanced configuration in a microcentrifuge, with cap hinges pointing outward. Centrifuge for 30 seconds at maximum speed to pellet the resin.
14. Use a micropipette with fresh tip to remove the supernatant, being careful not to disrupt the white pellet at the bottom of the tube. Spin the tube briefly to collect any drops of supernatant and then remove these with a micropipette.

In the presence of water or TE buffer, nucleic acids are eluted from the Silica resin.

For long-term storage it is recommended DNA samples be stored in TE buffer (Tris/EDTA). Tris provides a pH 8.0 environment to keep DNA and RNA nucleases less active. EDTA further inactivates nucleases by binding cations required by nucleases.

15. Add 100 µL of distilled water (or TE buffer) to the silica resin and mix well by vortexing or by pipetting up and down. Incubate the mixture at 57°C for 5 minutes.
16. Place your tube and those of other groups in a balanced configuration in a microcentrifuge, with cap hinges pointing outward. Centrifuge for 30 seconds at maximum speed to pellet the resin.
17. Label a clean 1.5 mL tube with your sample number. Transfer 90 µL of the supernatant (clear solution) to the fresh tube. Be careful not to disturb the pellet when transferring the supernatant. Discard old tube containing the resin.

Sample DNA eluted in TE may require a 1:10, 1:20 or 1:50 dilution in water prior to PCR, if the initial amplification of the target gene from the eluted DNA is unsuccessful (this may occur particularly in plant samples).

In Part III, you will use 2 µL of DNA for each PCR reaction. This is a crude DNA extract and contains nucleases that will eventually fragment the DNA at room temperature. Keep the sample cold to limit this activity.

18. Store your sample on ice or at -20°C until you are ready to begin Part III.

III. Amplify DNA by PCR

To amplify a DNA barcode region, choose the most appropriate set of primers for each sample. The table below lists available primer sets, the type of organism they target, and the PCR protocol for each set. For detailed information on the primer sets, go to Primer Sequences & References.

REAGENTS, SUPPLIES, & EQUIPMENT

For each group
Appropriate primer/loading dye mix (25 µL)* per reaction
DNA from specimen(s) (from Part II)*
Ready-To-Go PCR Bead in 0.2- or 0.5-mL PCR tube per reaction *OR* NEB *Taq* 2X Master Mix (12.5 µL)* per reaction

To share
Container with cracked or crushed ice
Micropipettes and tips (1–100 µL)
Microcentrifuge tube rack
Permanent marker
Thermal cycler

**Store on ice*

1. Obtain PCR tube containing Ready-To-Go PCR Bead. Label the tube with your identification number.
2. Use a micropipette with a fresh tip to add 23 µL of one of the following primer/loading dye mixes to each tube. Allow the beads to dissolve for 1 minute.

   Plant cocktail: *rbc*L primers (rbcLaF / rbcLa rev)

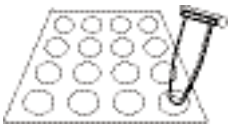
   Fungi cocktail: *ITS* primers (ITS1F/ITS4)

   Fish cocktail: *COI* primers (VF2_t1/ FishF2_t1/ FishR2_t1/ FR1d_t1)

   Vertebrate (non-fish) Cocktail:
   (VF1_t1/VF1d_t1/VF1i_t1/VR1d_t1/VR1_t1/VR1i_t1)

   Invertebrate cocktail: (LCO1490/ HC02198)

If the reagents become splattered on the wall of the tube, pool them by pulsing the sample in a microcentrifuge or by sharply tapping the tube bottom on the lab bench.



3. Use a micropipette with fresh tip to add 2 µL of your DNA (from Part II) directly into the appropriate primer/loading dye mix. Ensure that no DNA remains in the tip after pipetting.
4. Store your sample on ice until your class is ready to begin thermal cycling.
5. Place your PCR tube, along with those of the other students, in a thermal cycler that has been programmed with the appropriate PCR protocol.

Primers	Profile
Plant cocktail (rbcLaF / rbcLa rev) **Vertebrate (non-fish) cocktail** (VF1_t1/VF1d_t1/VF1i_t1/ VR1d_t1/VR1_t1/VR1i_t1)	Initial step: 94°C 1 minute 35 cycles of the following profile: Denaturing step: 94°C 15 seconds Annealing step: 54°C 15 seconds Extending step: 72°C 30 seconds One final step to preserve the sample: 4°C *ad infinitum*
Fish cocktail (VF2_t1/ FishF2_t1/ FishR2_t1/ FR1d_t1)	Initial step: 94°C 1 minute 35 cycles of the following profile: Denaturing step: 94°C 15 seconds Annealing step: 54°C 15 seconds Extending step: 72°C 30 seconds One final step to preserve the sample: 4°C *ad infinitum*

Primers	Profile
Invertebrate cocktail (LCO1490/ HC02198)	Initial step: 94°C 1 minute 35 cycles of the following profile: Denaturing step: 95°C 30 seconds Annealing step: 50°C 30 seconds Extending step: 72°C 45 seconds One final step to preserve the sample: 4°C *ad infinitum*
Fungi cocktail (ITS1F/ITS4)	Initial step: 94°C 1 minute 35 cycles of the following profile: Denaturing step: 94°C 1 minute Annealing step: 55°C 1 minute Extending step: 72°C 2 minutes One final step to preserve the sample: 4°C *ad infinitum*

6. After thermal cycling, store the amplified DNA on ice or at -20 °C until you are ready to continue with Part IV.

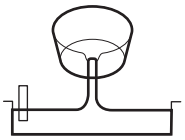
IV. Analyze PCR Products by Gel Electrophoresis

REAGENTS, SUPPLIES, & EQUIPMENT

For each group
- 2% agarose in 1x TBE (hold at 60°C) (50 mL per gel)
- pBR322/*Bst*NI marker (20 µL per gel)*
- PCR products from Part III*
- SYBR Green DNA stain (6 µL per group)
- 1x TBE buffer (300 mL per gel)

*Store on ice.

To share
- Container with cracked or crushed ice
- Gel-casting tray and comb
- Gel electrophoresis chamber and power supply
- Latex gloves
- Masking tape
- Microcentrifuge tube rack
- 3 Microcentrifuge tubes (1.5mL)
- Micropipette and tips (1–100 µL)
- Digital camera or photodocumentary system
- Microwave
- UV transilluminator <!> and eye protection
- Water bath for agarose solution (60°C)



Avoid pouring an overly thick gel, which makes visualization of the DNA more difficult.

The gel will become cloudy as it solidifies.

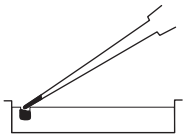
Do not add more buffer than necessary. Too much buffer above the gel channels electrical current over the gel, increasing running time.

1. Seal the ends of the gel-casting tray with masking tape, or other method appropriate for the gel electrophoresis chamber used and insert a well-forming comb.
2. Pour the 2% agarose solution into the tray to a depth that covers about one-third the height of the comb teeth.
3. Allow the agarose gel to completely solidify; this takes approximately 20 minutes.
4. Place the gel into the electrophoresis chamber and add enough 1x TBE buffer to cover the surface of the gel.
5. Carefully remove the comb and add additional 1x TBE buffer to fill in the wells and just cover the gel, creating a smooth buffer surface.
6. Use a micropipette with a fresh tip to transfer 5 µL of each PCR product (from

part III) to a fresh 1.5 mL microcentrifuge tube. Add 2 µL of SYBR Green DNA stain to tube.

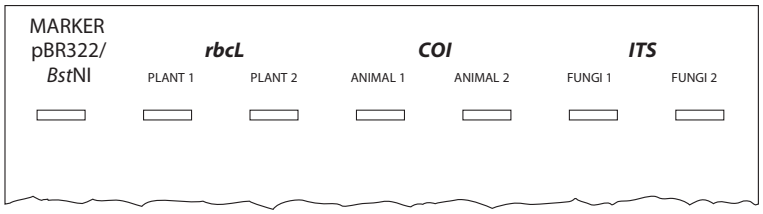
A 100-bp ladder may also be used as a marker.

7. Add 2 µL of SYBR Green DNA stain to 20 µL of pBR322/*Bst*NI marker.
8. Orient the gel according to the diagram below, so the wells are along the top of the gel. Use a micropipette with a fresh tip to load 20 µL of pBR322/*Bst*NI size marker into the far left well.



Expel any air from the tip before loading, and be careful not to push the tip of the pipette through the bottom of the sample well.

9. Use a micropipette with a fresh tip to load each sample from Step 6 in your assigned wells, similar to the following diagram:



The samples you load may not be exactly the same as those shown.

10. Store the remaining 20 µL of your PCR product on ice or at -20°C until you are ready to submit your samples for sequencing.

Transillumination, where the light source is below the gel, increases brightness and contrast.

11. Run the gel for approximately 30 minutes at 130V. Adequate separation will have occurred when the cresol red dye front has moved at least 50 mm from the wells.
12. View the gel using UV transillumination. Photograph the gel using a digital camera or photodocumentary system.

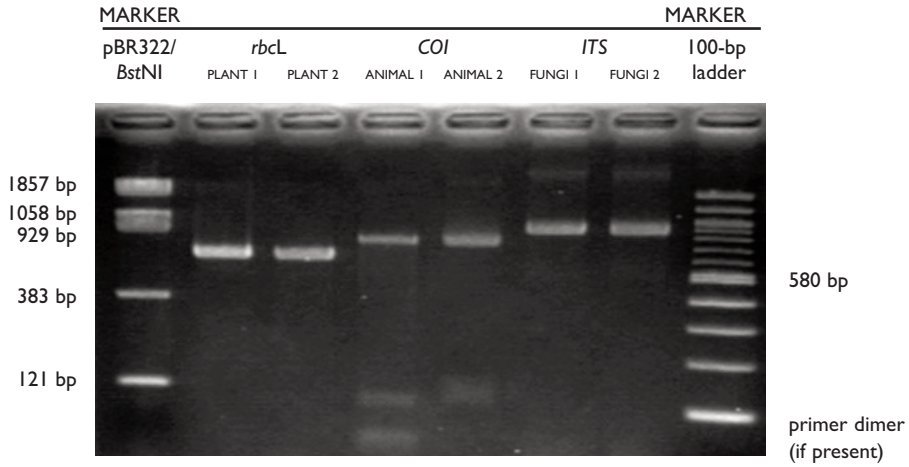
RESULTS AND DISCUSSION

I. Think About the Experimental Methods

1. Describe the effect of each of the following steps or reagents used in DNA isolation (Part I of Experimental Methods):
   i. Collecting fresh or dried specimens
   ii. Using only a small amount of tissue
   iii. Grinding tissue with pestle
   iv. Lysis solution
   v. Heating or boiling.

II. Interpret Your Gel and Think About the Experiment

1. Observe the photograph of the stained gel containing your PCR samples and those from other students. Orient the photograph with the sample wells at the top. Use the sample gel shown on the next page to help interpret the band(s) in each lane of the gel.
2. Locate the lane containing the pBR322/*Bst*NI markers on the left side of the gel. Working down from the well, locate the bands corresponding to each restriction fragment: 1857, 1058, 929, 383, and 121 bp. The 1058- and 929-bp fragments will



be very close together or may appear as a single large band. The 121-bp band may be very faint or not visible.

Additional faint bands at other positions occur when the primers bind to chromosome loci other than the intended locus and give rise to "nonspecific" amplification products.

3. Looking across the gel at the PCR products, do the bands all appear to be the same bp size and intensity?
4. It is common to see a diffuse (fuzzy) band that runs ahead of the 121-bp marker. This is "primer dimer," an artifact of the PCR that results from the primers overlapping one another and amplifying themselves.

If you have a very faint product or none at all, your teacher will help you decide if your sample should be sent for sequencing.

5. Which samples amplified well, and which ones did not? Give several reasons why some samples may not have amplified; some of these may be errors in procedure.
6. Generally, DNA sequence can be obtained from any sample that gives an obvious band on the gel.

BIOINFORMATICS

I. Use BLAST to Find DNA Sequences in Databases (Electronic PCR)

1. Perform a BLAST search as follows:
   a. Do an Internet search for "ncbi blast."
   b. Click the link for the result BLAST: *Basic Local Alignment Search Tool.* This will take you to the Internet site of the National Center for Biotechnology Information (NCBI).
   c. Under the heading "Basic BLAST," click "nucleotide blast."
   d. Enter the primer set you used into the search window. These are the query sequences. (See box at top of next page.)
   e. Omit any non-nucleotide characters from the window because they will not be recognized by the BLAST algorithm.
   f. Under "Choose Search Set," select "NCBI Genomes (chromosome)" from the pull-down menu.
   g. Under "Program Selection," optimize for "Somewhat similar sequences (blastn)."

The following primers were used in this experiment:

Plant *rbc*L gene
- rbcLa f 5'- ATGTCACCACAAACAGAGACTAAAGC-3' (forward primer)
- rbcLa rev 5'- GTAAAATCAAGTCCACCRCG-3' (reverse primer)

Vertebrate (non-fish) *COI* gene
- VF1_t1 5'-TCTCAACCAACCACAAAGACATTGG-3' (forward primer)
- VR1d_t1 5'-TAGACTTCTGGGTGGCCRAARAAYCA-3' (reverse primer)

Fish *COI* gene
- VF2_t1 5'-CAACCAACCACAAAGACATTGGCAC-3' (forward primer)
- FishR2_t1 5'-ACTTCAGGGTGACCGAAGAATCAGAA-3' (reverse primer)

Fungi *ITS*
- ITS1 F 5'-TCCGTAGGTGAACCTGCGG-3' (forward primer)
- ITS4 R 5'-TCCTCCGCTTATTGATATGC-3' (reverse primer)

Invertebrate *COI* gene
- LCO1490_F 5'-GGTCAACAAATCATAAAGATATTGG-3' (forward primer)
- HC02198_R 5'-TAAACTTCAGGGTGACCAAAAAATCA-3' (reverse primer)

h. Click "BLAST." This sends your query sequences to a server at the National Center for Biotechnology Information in Bethesda, Maryland. There, the BLAST algorithm will attempt to match the primer sequences to the DNA sequences stored in its database. A temporary page showing the status of your search will be displayed until your results are available. This may take only a few seconds or more than a minute if many other searches are queued at the server.

2. The results of the BLAST search are displayed in three ways as you scroll down the page:
   a. First, a *Graphic Summary* illustrates how significant matches, or "hits," align with the query sequence. **Why are some alignments longer than others?**
   b. This is followed by *Descriptions of sequences producing significant alignments*, a table with links to database reports.
      - The accession number is a unique identifier given to a sequence when it is submitted to a database, such as Genbank®. The accession link leads to a detailed report on the sequence.
      - Note the scores in the "E Value" column on the right. The Expectation, or E, value is the number of alignments with the query sequence that would be expected to occur by chance in the database. The lower the E value, the higher the probability that the hit is related to the query. For example, an E value of 1 means that a search with your sequence would be expected to turn up one match by chance.
      - **What is the E value of your most significant hit, and what does it mean? What does it mean if there are multiple hits with similar E values?**
      - **What do the descriptions of significant hits have in common?**
   c. Next is an *Alignments* section, which provides a detailed view of each primer

sequence (*Query*) aligned to the nucleotide sequence of the search hit (*subject*). Notice that hits have matches to one or both of the primers:

	Forward Primer	*Reverse Primer*
Plant	nucleotides 1-26	nucleotides 27-46
Vertebrate (non-fish)	nucleotide 1-25	nucleotides 26-53
Fish	nucleotides 1-25	nucleotides 26-51
Fungi	nucleotide 1-19	nucleotides 20-39
Invertebrate	nucleotides 3-25	nucleotides 26-51

3. Predict the length of the product that the primer set would amplify in a PCR reaction (*in vitro*).
   a. In the *Alignments* section, select a hit that matches both primer sequences.
   b. **Which nucleotide positions do the primers match in the subject sequence?**
   c. The lowest and highest nucleotide positions in the subject sequence indicate the borders of the amplified sequence. Subtracting one from the other gives the difference between the coordinates.
   d. However, the PCR product includes both ends, so add 1 nucleotide to the result that you obtained in Step 3.c. to determine the exact length of the fragment amplified by the two primers.
   e. **What value do you get if you calculate the fragment size for other species that have matches to the forward and reverse primer? Do you get the same number?**
4. Determine the type of DNA sequence amplified by the primer set:
   a. Click on the accession link (beginning with "*ref*") to open the data sheet for the hit used in Question 3 above.
   b. The data sheet has three parts:
      - The top section contains basic information about the sequence, including its basepair (bp) length, database accession number, source, and references to papers in which the sequence is published.
      - The bottom section lists the nucleotide sequence.
      - The middle section contains annotations of gene and regulatory FEATURES, with their beginning and ending nucleotide positions ("xx..xx"). These features may include genes, coding sequences (CDS), regulatory regions, ribosomal RNA (rRNA), and transfer RNA (tRNA).
   c. Identify the feature(s) located between the nucleotide positions identified by the primers, as determined in 3.b. above.

II. Determine Sequence Relationships Using the Blue Line

The following directions explain how to use the Blue Line of *DNA Subway* to analyze novel DNA sequences generated by a DNA sequencing experiment. If you did not sequence your own DNA sample, you can follow these directions to use DNA sequences produced for other students. You can find supplementary instructions by clicking on the "manual" link on the *DNA Subway* homepage.

DNA Subway is an intuitive interface for analyzing DNA barcodes. Generally, you progress in a stepwise fashion through the button "stops" on each "branch line." An "R" indicates that analysis is available. A blinking "R" indicates an analysis is in process. A "V" means that results are ready to view.

You can analyze relationships between DNA sequences by comparing them to a set of sequences you have compiled yourself, or by comparing your sequences to others that have been published in databases such as GenBank® (National Center for Biotechnology Information). Generating a phylogenetic tree from DNA sequences derived from related species can also allow you to draw inferences about how these species may be related. By sequencing variable sections of DNA (barcode regions) you can also use the Blue Line to help you identify an unknown species, or publish a DNA barcode for a species you have identified, which is not represented in published databases like GenBank® (www.ncbi.nlm.nig.gov/genbank).

1. Create a *DNA Subway* Project and Upload DNA Sequences
   a. Log in to *DNA Subway* at www.dnasubway.org. If you do not have an account, you will need to register first to save and share your work.
   b. Select "Determine Sequence Relationships" (Blue Line) to begin a project.
   c. Select "*rbc*L" or "*COI*" from the "Select Project Type" section. (*rbc*L (plant) sequences must be analyzed separately from *COI* (animal) sequences.) If you are analyzing a barcode region that is not listed, select "DNA."
   d. "Select Sequence Source" provides several ways to obtain sequences for barcode analysis:
      - *Upload sequence(s) in ab1* (files ending with .ab1) *or FASTA* format. Click "Browse" to navigate to a folder on your desktop or drive containing your sequence(s). Select a sequence by clicking on its file name. Select more than one sequence by holding down the ctrl key while clicking file names. Once you have selected the sequences you want, click "Open".
      - *Enter a sequence in FASTA format*. Below is an example of this format. The ">" symbol demarcates the sequence name. The sequence is started on the next line.

        >sequence name
        atcgccccttaatattgcctt…..

      - *Import a sequence/trace from the DNALC*. Click your tracking number. Select one or more files from the list. Click to "Add" selected files.
      - *Select a sample sequence*.
   e. Provide a title in the *Name Your Project* section.
   f. Write a short description of your project in the *Description* section (optional).
   g. Click "Continue" to load the project into *DNA Subway*.
2. View and Build Sequences

   There are many plants, animals, and fungi which do not have a documented barcode sequence. For instance, there are an estimated 350,000 species of

angiosperms (flowering plants), but as of June 2013 there were only about 74,000 *rbc*L angiosperm sequences in GenBank®. For other species, diversity in the barcode sequences are not well characterized. This means that there are opportunities to submit novel sequences and contribute to the global barcoding effort. Only samples that have high quality sequence for both the forward and reverse reads are good enough to ensure a low error rate and can be published to GenBank®, so the sequence quality must be checked. Sequences for which there is only one high quality read are not considered high enough quality to publish. These sequences and those with no high quality sequence can still be analyzed even though the results are not publishing quality.

a. On the *Assemble Sequences* branch line, Click "Sequence Viewer" to display the sequences you have input in the project creation section. If you did not upload trace files, you can scroll to see the sequence. If you uploaded trace files, click on the file names to view the trace files.

   - The DNA sequencing software measures the fluorescence emitted in each of four channels—A, T, C, G—and records these as a trace, or electropherogram. In a good sequencing reaction, the nucleotide at a given position will be fluorescently labeled far in excess of background (random) labeling of the other three nucleotides, producing a "peak" at that position in the trace. Thus, peaks in the electropherogram correlate to nucleotide positions in the DNA sequence.
   - A software program called *Phred* analyzes the sequence file and "calls" a nucleotide (A, T, C, G) for each peak. If two or more nucleotides have relatively strong signals at the same position, the software calls an "N" for an undetermined nucleotide.
   - *Phred* also examines the peaks around each call and assigns a quality score for each nucleotide. The quality scores corresponds to a logarithmic error probability that the nucleotide call is wrong, or, conversely, to the accuracy of the call.

***Phred* Score**	**Error**	**Accuracy**
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

   - The electropherogram viewer represents each *Phred* score as a blue bar. The horizontal line equals a *Phred* score of 20, which is generally the cutoff for high-quality sequence. Thus any bar at or above the line is considered a high-quality read. **What is the error rate and accuracy associated with a *Phred* score of 20?**
   - Every sequence "read" begins with nucleotides (A, T, C, G) interspersed with Ns. In "clean" sequences, where experimental conditions were near optimal, the initial Ns will end within the first 25 nucleotides. The remaining sequence will have very few, if any, internal Ns. Then, at the end of the

read the sequence will abruptly change over to Ns.

- Large numbers of Ns scattered throughout the sequence indicate poor quality sequence. Sequences with average *Phred* scores below 20 will be flagged with a "Low Quality Score Alert." You will need to be careful when drawing conclusions from analyses made with poor quality sequence. **What do you notice about the electropherogram peaks and quality scores at nucleotide positions labeled "N"?**
- **Note:** The exclamation icon (!) indicates poor quality sequence.

b. Use the "X" and "Y" buttons to adjust the level of zoom. You can undo zooming by pressing the "Reset" button.

c. Examine the quality of the sequence(s). Any sequence for which the forward or reverse has the warning icon indicating a low quality score is not of good enough quality to publish and any determination of novelty will be tentative as sequencing errors could appear to be novel polymorphisms.

d. Click "Sequence Trimmer" to trim your sequences; this automatically removes Ns from the 5' and 3' ends of selected sequences. Click again to view the trimmed sequences. **Why is it important to remove excess Ns from the ends of the sequences?**

e. If you wish to view trimmed sequences, click on the file name.

3. Pair and Build Consensus for Forward and Reverse Reads

a. If you have two reads for a sample, pair the sequences by checking the box to the right of each read for the sample. By default, *DNA Subway* assumes that all reads are in the forward orientation, and displays an "F" to the right of the sequence. If any sequence is not in that orientation, click the F to reverse complement the sequence. The sequence will display an "R" to indicate the change.

b. After checking the second read, a dialogue box will appear asking if you wish to designate the sequences as a pair. Alternatively, Click "Try auto pairing" to pair sequences which have identical sample names, but appended with an F or R based on sequencing direction.

c. Click "Save" to save your pair assignments.

d. Once you have created sequence pairs, click "Consensus Editor" to make a consensus sequence from both sequences in the selected pairs. To examine the consensus sequence click "Consensus Editor" again, and then click on the link to the pair you wish to examine. How does the consensus sequence optimize the amount of sequence information available for analysis? Why does this occur?

e. If there are any mismatched nucleotides between the first and second sequence, these will be highlighted yellow in the consensus editor window. **Do differences tend to occur in certain areas of the sequence? Why?**

f. Large numbers of yellow mismatches—especially in long blocks—may indicate that you have incorrectly paired sequences from two different sources (organisms), or that you failed to reverse complement the reverse strand.

- Return to *Pair Builder* to check your pairs and reverse complements.
- Click the red "x" to redo a pairing, and toggle "F" and "R" settings, as needed.

h. A large number of mismatches in properly paired and reverse complemented sequences indicate that one or both sequences is of poor quality. Often, one of the sequencing reactions produces a high quality read that can be used on its own. To determine this:

- Examine the distribution of Ns to see if they are mainly confined to one of the two sequences.
- Examine the electropherograms to see if one of the two sequences is of good quality.
- If one of the sequences seems of good quality, return to *Pair Builder*, and click the red x to undo the pairing.
- Continue on to Step 4.

i. Few or no internal mismatches indicate good quality sequence from forward and reverse reads. If you like, you can check the consensus sequence at yellow mismatches and override the judgment made by the software:

- Click a highlighted mismatch to see the electropherograms and graphic summarizing *Phred* scores for each read.
- Click the desired nucleotide in the black rectangle to change the consensus sequence at that position. You should only change the consensus if you have a strong reason to believe the consensus is wrong.
- Click the button to "Save Change(s)."

4. BLAST Your Sequence

A BLAST search can quickly identify any close matches to your sequence in sequence databases. In this way, you can often quickly identify an unknown sample to the genus or species level. It also provides a means to add samples for a phylogenetic analysis.

a. On the *Add Sequences* branch, click "BLASTN". Then, click the "BLAST" button next to the sequence you want to query against DNA databases.

b. The returned list has information about the 20 most significant alignments (hits):

- Accession number, a unique identifier given to each sequence submitted to a database. Prefixes indicate the database name—including gb (GenBank®), emb (European Molecular Biology Laboratory), and dbj (DNA Databank of Japan).
- Organism and sequence description or gene name of the hit. Click the genus and species name for a link to an image of the organism, with additional links to detailed descriptions at Wikipedia and Encyclopedia of Life (EOL).
- Several statistics allow comparison of hits across different searches. The number of mismatches over the length of the alignment gives a rough idea of how closely two sequences match. The Bit Score formula takes into

account gaps in the sequence; the higher the score the better the alignment. The Expectation or E-value is the number of alignments with the query sequence that would be expected to occur by chance in the database. The lower the E-value, the higher the probability that the hit is related to the query. For example, an E value of 0 means that a search with your sequence would be expected to turn up no matches by chance. **Why do the most significant hits typically have E-values of 0?** (This is not the case with BLAST searches with primers.) **What does it mean when there are multiple BLAST hits with similar E values?**

- Examine the last column in the report called "Mismatches." For barcodes, this is the informative column, with the best hits being those with the lowest number of mismatches. Note that hits with low numbers of mismatches can sometimes be lower on the list, as the bit scores are used to arrange the hits in the table. High bit scores can occur when the alignment length is longer, even when there are more mismatches than for other hits.
- If there are zero mismatches between your sequence and a BLAST result, it is unlikely that your sequence is unique. Instead, the identical sequences probably match because they are in the same taxonomic group as your sample. Check to see if the matching sequences are from species that seem reasonable for your sample. If your best matches include some mismatches, you may have identified a novel barcode. The more mismatches you find, the more likely that your sequence is unique, especially in regions of the sequence with high quality scores. However, sequencing errors could explain the difference, so it will be important to reexamine the trace files at any sites with mismatches to ensure that the consensus at those locations is of high quality.

c. Add BLAST sequence data to your phylogenetic analysis by checking the box(es) next to any accession number(s), then clicking on "Add BLAST hits to project" at the bottom of the BLAST results window.

5. Add Sequences to Your Analysis

a. Click "Upload Data" to add additional sequence data to your analysis without starting a new project. Use "Upload Sequence(s)" to upload *ab1* trace files or *FASTA*-formatted sequences stored locally on your computer; Use "Enter Sequences(s)" to paste or type sequences in *FASTA* format.

b. If you would like to import sequences from non-local sources you can use "Import Sequence" to search a sequence database using a sequence identifier. For GenBank® sequences you can search by identification number (GI or Version). Search BOLD by species name, or search the DNALC sequence database by tracking number for sequences you processed with GENEWIZ through the DNALC system.

c. If your sequence had no hits with zero mismatches, you may use NCBI BLAST to confirm that the sequence is novel. Click on the BLASTN button and then double-click on the sequence (the actual nucleotides) that you identified as possibly novel to select them. Right-click (PC) or command-click (Mac) and then select copy to move the sequence to your clipboard.

- In a web browser go to http://blast.ncbi.nlm.nih.gov. From this page click on "nucleotide blast."
- Paste your sequence into the "Enter Query Sequence" window under "enter accession number(s), gi(s), or FASTA sequence(s)."
- Under "Program Selection" select "Highly similar sequences (megablast)"; next click "BLAST."
- On the results page you will get a list of results very similar to what was returned by *DNA Subway*.
- Scrolling down the page, you will find alignments of your sequence (Query) to the sequences from the closest matches in GenBank® (Sbjct).
- Analyze the results of the BLAST search, which are displayed in three ways as you scroll down the page:
  - First, a graphical overview illustrates how significant matches (hits) align with the query sequence. Matches of differing lengths are indicated by color-coded bars. For barcoding results, it is likely that most matches will be red, indicating high scores, and cover most of the width of the table, showing matches that span the length of your query sequence.
  - This is followed by a table with "Descriptions of sequences producing significant alignments" much like the table for BLAST results in *DNA Subway*.
  - Next is an "Alignments" section, which provides a detailed view of each primer sequence ("Query") aligned to the nucleotide sequence of the search hit ("Sbjct," "subject").
  - From the table, identify any matches that are 100% identical or any matches with high identity that appear to represent species or sequences you have not identified previously. Select these sequences by clicking on the box to the left of each hit. After selecting sequences, click Download, ensure *FASTA* (complete sequence) is selected, and then click Continue.
  - Open the resulting *FASTA* file (named seqdump). *Double-click* the sequences to select them all, then *right-click* (PC) or *command-click* (Mac) and *select* copy to move the sequence to your clipboard. Add these sequences to your project using the Upload Data function, as in step 1.
- Click "Sequence Viewer" back on *DNA Subway*, and view the trace file for the forward read of your query sequence. Locate the position on your table where the query sequence differed from the GenBank® match. Determine if the nucleotides you identified as different were of high quality (e.g. not sequencing errors). Because of sequence trimming, you may have to search for the polymorphic site, as the numbers from the BLAST alignment and in the trace file may not correspond.

d. You may also choose to search for your sequence at the International Barcode of Life (IBOL) database, BOLD (Barcode of Life Online Database); their records are not all in GenBank®.

- Click the BLAST button and then *double-click* the nucleotides for the sequence you are analyzing. *Right-click* (PC) or *command-click* (Mac) and then *select* copy to move the sequence to your clipboard.
- In a web browser go to http://boldsystems.org. From this page, click on "Identification."
- Select the tab that corresponds to the appropriate kingdom for the sample (animal, plant, or fungal).
- Under the Animal Identification [COI] tab, select "Species Level Barcode Records." On the Fungal Identification [ITS] tab, select "ITS Sequences." On the Plant Identification [rcbL & matK] tab, select "Plant Sequences."
- Paste the sequence into the search box labeled "Enter sequences in fasta format"; next click "Submit."
- Again, a results table is produced. The column labeled "similarity" indicates how similar your sequence was to the records in the BOLD, with a 100% match indicating they were exact matches. Some records in BOLD are not public, or are not accompanied by species-level identifications. Scrolling down the list of matches you will see a pairwise alignment of your sequence (Query) to the matched sequences (Subj). Once again, identify any new hits that may be identical to your sequence. For published hits, you can download the sequence by clicking the link to the right of "Published," then clicking "FASTA" and saving the file. This FASTA file can be uploaded, as described above, in step 1.

e. Click "Reference Data" (optional) to include additional sequences. Depending on the project type you have created, you will have access to additional sequence data that may be of interest. For example, if you are doing a DNA barcoding project using the *rbc*L gene, samples of *rbc*L sequence from major plant groups (Angiosperms, Gymnosperms, etc.) will be provided. Choose any data set to add it to your analysis; you will be able to include or exclude individual sequences within the set in the next step.

6. Analyze Sequences: Select and Align

Many unknown species can be rapidly identified by a BLAST search. In this case, a phylogenetic analysis adds depth to your understanding by showing how your sequence fits into a broader taxonomy of living things. If your BLAST search fails to identify your sequence, phylogenetic analysis can usually identify it to at least the family level.

a. Click "Select Data" to display all the sequences you have brought into your analysis, including "user data," BLAST hits, or reference data. Check off sequences you wish to include in an alignment. In general, to determine the relationship of your sequence to species with known barcodes, it is best to concentrate on similar sequences. For instance, you should align sequences from samples that you believe are the same species and any close matches from database searches. You may also use the "Select all" feature to include all sequences; to deselect all sequences, click "Select all" a second time. You may run new alignments or download different sequences at any time after

selecting a new set of sequences.

- To download selected sequences to a FASTA file click the "Download" button and save the resulting file.
- To save your selections, click "Save Selections" in the blue dialog box that appears when you make any selections.

b. Click "MUSCLE" to run the MUSCLE multiple alignment software. This software will align all sequences that were included in the "Select Data" step. Click "MUSCLE" again to open the created multiple alignment. An alignment that is suitable for creating a phylogenetic tree will have an overall high Sequence Conservation Score (represented by the height of the gray bars along the Sequence Conservation row at the top).

- Scroll through your alignments to see similarities between sequences. Nucleotides are color coded, and each row of nucleotides is the sequence of a single organism or sequencing reaction. Columns are matches (or mismatches) at a single nucleotide position across all sequences. Dashes (-) are gaps in sequence, where nucleotides in one sequence are not represented in other sequences.
- Note that the 5' (leftmost) and 3' (rightmost) ends of the sequences are usually misaligned, due to gaps (-) or undetermined nucleotides (Ns). **What causes these problems?**
- Note any sequence that introduces large, internal gaps (-----) in the alignment. This is either poor quality or unrelated sequence that should be excluded from the analysis. To remove it, return to Select Data, uncheck that sequence, and save your change. Then click "MUSCLE" to recalculate.

c. You will need to "trim" the alignment. To trim, click the "Trim Alignment" button on the upper-left of the Alignment Viewer. **Why is it important to remove sequence gaps and unaligned ends?**

7. Analyze Sequences: Create a Phylogenetic Tree

a. Click "PHYLIP ML" to generate a phylogenetic tree using the maximum likelihood method. Click "PHYLIP ML"again and a tree will open in a new window, and the MUSCLE alignment used to produce it will open in another window.

b. A phylogenetic tree is a graphical representation of relationships between taxonomic groups. In this experiment, a *gene tree* is determined by analyzing the similarities and differences in DNA sequence.

c. Look at your tree.

- The branch tips are the DNA sequences of individual species or samples you analyzed. Any two branches are connected to each other by a node, which represents the common ancestor of the two sequences.
- The length of each branch is a measure of the evolutionary distance from the ancestral sequence at the node. Species or sequences with short branches from a node are closely related, while those with longer branches are more distantly related.

- A group formed by a common ancestor and its descendants is called a clade. Related clades, in turn, are connected by nodes to make larger, less closely-related clades.
- Click on a node to highlight sequences in that clade. Click the node again to deselect the clade. **What assumptions are made when one infers evolutionary relationships from sequence differences?**
- Generally, the clades will follow established phylogenetic relationships ascending from genus > family > order > class > phylum. However, gene and phylogenetic trees do disagree on some placements, and much research is focused on "reconciling" these differences. **Why do gene and phylogenetic trees sometimes disagree?**

d. Find and evaluate your sequence's position in the tree.

- If your sequence is closely related to any of the reference or uploaded sequences, it will share a single node with those species.
- If your sequence is identical to another sequence, the two will diverge directly from the node without branches.
- If your sequence is distantly related to all of the species in your tree, your sequence will sit on a branch by itself—with the other sequences grouping together as a clade.
- Look at the scientific names of sequences within the most closely associated clade. If all members share the same genus name, you have identified your sequence as belonging to that genus. If different genus names are represented, check and see if they belong to the same family or order.

e. Return to the menu, and click on "PHYLIP NJ" to generate a phylogenetic tree using the neighbor joining method. **How does it compare to the maximum likelihood tree? What does this tell you?**

f. If neither tree places your sequence within an identifiable clade—or if that clade is only at order level—you will need to add more sequences that may increase the resolution of your analysis. Return to Step 5, and add more reference sequences or obtain sequences within the order or family clade that contained your sequence. Then repeat Steps 6-7 to select, align, and generate trees from your refined data set.

8. Exporting Sequences to GenBank®

If you do not identify any identical hits through searches in *DNA Subway*, GenBank®, and BOLD, and you have determined that your sequence is of high quality, you may have a novel sequence.

Once you have identified a potentially novel sequence there are additional steps that you can take, including publishing your sequence to GenBank® through *DNA Subway*. It is not required that a sequence be novel to publish it to GenBank®. However, discretion should be used, and sequences that are already present in GenBank® multiple times for a particular species or without vetted metadata (definitive species identification, collection information, etc.) should not be published.

Note: Only high quality consensus sequences that have been generated by a submitter, and which have not been previously submitted can be exported to GenBank®.

a. Click "Export to GenBank®" in the project window.
b. Click "New submission." (If you are working with an animal sample, you need to specify if it is from a vertebrate, invertebrate, or echinoderm) then Click "Proceed."
d. If you have already collected information of your samples in the DNALC Barcoding Samples Database, write the sample's code number. Its information will be retrieved automatically. If not, you can enter the sample information manually in the next step; click "Continue."
e. Verify and fill in the information required in the "Specimen info" window; click "Continue".
f. Add photos of the sample if you have any available.
g. Verify your submission information, make any appropriate changes if necessary, and finally click "Submit." You will receive a notification that your sequence has been submitted to NCBI and a specialist there will check it. If your submission passes NCBI's verification procedure, you will receive a notification that your sequence has been published in GenBank®.

ANSWERS TO RESULTS AND DISCUSSION QUESTIONS

I. Think About the Experimental Methods

1. **Describe the purpose of each of the following steps or reagents used in DNA isolation (Part II or Part IIa of Experimental Methods):**

   i. **Collecting fresh or dried specimens**

   Fresh samples are easier to isolate DNA from than other samples, so they maximize the chances of success. Dried specimens are common for plant and fungal collections and often contain intact DNA, although this DNA can be more difficult to isolate. Other samples, such as those that are processed or degraded, can have less intact DNA or PCR inhibitors.

   ii. **Using only a small amount of tissue**

   Using a small amount of tissue reduces carry-forward of PCR inhibitors present in the sample. These include metal ions (plants and animals), polysaccharides, and secondary metabolites (plants).

   iii. **Grinding tissue with pestle**

   Grinding disrupts plant cell walls and animal chitin or connective tissue. It also produces small clumps of cells that are more easily lysed to release DNA.

   iv. **Lysis solution**

   GuHCl lysis solution is a chaotropic agent, which interferes with hydrogen bonds and other interactions that stabilize stuctures. This dissolves the cell membrane and membrane-bound organelles (nucleus, mitochondria,

chloroplast, etc.). In addition, GuHCL denatures biomolecules by disrupting hydrogen bonds with water surrounding them. This allows positively charged ions to form a salt bridge between the negatively charged silica and the negatively charged DNA backbone in high salt concentration.

v. **Heating or boiling**

Heating to 65°C with the GuHCl lysis solution helps to break down the cell and nuclear membranes and also denatures enzymes that can degrade the purified DNA. Heating to 57°C helps with the binding and release of DNA to the silica resin in the presence of the GuHCl lysis solution and distilled water respectively.

II. Interpret Your Gel and Think About the Experiment

3. **Looking across the gel at the PCR products, do the bands all appear to be the same bp size and intensity?**

rbcL, COI, and *ITS* primers amplify differently sized products that migrate to different positions on the gel. However, each barcode primer set is optimized to amplify the same region across a range of species. Although the size of products for each primer can vary, the majority of PCR products will be of similar base-pair size and, therefore, will migrate to the same position on the gel. However, the intensity of staining (thickness of bands) will vary between reactions. This is related to the mass of DNA product produced by the PCR reaction and the volume of the reaction that is successfully loaded in the well.

5. **Which samples amplified well, and which ones did not? Give several reasons why some samples may not have amplified; some of these may be errors in procedure.**

It may be difficult to extract enough DNA from tough leaves or dry materials. Some primer sets may not work with certain groups of organisms; for example, *rbc*L primers work less well with non-vascular plants (mosses and liverworts).

Major problems in PCR amplification typically occur at several points in the procedure: a) grinding step did not sufficiently disrupt the tissue, b) supernatant transferred after protein precipitation carried forward too many inhibitors, c) the nucleic acid pellet was lost after the precipitation step, or d) the small volume of DNA template was not pipetted directly into the PCR reaction (it was left in pipette or on wall of PCR tube).

ANSWERS TO BIOINFORMATICS QUESTIONS

I. Use BLAST to Find DNA Sequences in Databases (Electronic PCR)

2.a. **Why are some alignments longer than others?**

The main difference in length occurs between hits that align to both primers versus those that align only to the forward or reverse primer. The lengths and colors of the alignment bars tell how much of your query matched sequences in the database. Where the forward and reverse primer matches, you will see a black vertical line between the forward and reverse primer in the graphic summary.

Typically, most of the significant alignments will have complete matches to the forward and reverse primers.

2.b. **What is the E value of the most significant hit and what does it mean? What does it mean if there are multiple hits with similar E values?**

The lowest E value obtained for a match to both primers should be in the range of 0.001 to 2e-04, or 0.0002. This might seem high for a probability, but in fact each of these values means that a match of this quality would be expected to occur by chance less than once in this database! For example, a score of 0.33 would mean that a single match would be expected to occur by chance once in every three searches. E values are based on the length of the search sequence, and thus the relatively short primers used in this experiment produce relatively high E values. Searches with longer primers or long DNA sequences return E values with smaller values. Multiple hits with similar E values are from closely related species.

What do the descriptions of significant hits have in common?

For the plant primers, the sequence sources should all be chloroplast genomes. For the vertebrate, fish, and invertebrate primers, the hits should all be mitochondrial genomes. For the fungi primers, the hits should all be to the nuclear internal transcribed spacer of the 5.8s ribosomal RNA gene.

3.b. **Which nucleotide positions do the primers match in the subject sequence?**

The answers will vary for each hit and primer set.

For *Silene Conoidea* (NC_023358.1), the plant primers match 43684–43709 and 43111–43130, respectively.

For *Pucrasia macrolopha* (NC_020587.1), the vertebrate (non-fish) primers match 6589–6613 and 7272–7298 respectively.

For *Mallotus villosus* (NC_015244.1), the fish primers match 5556–5584 and 6233–6258 respectively.

For *Candida orthopsilosis* (NC_018301.1), the fungi primers match 344066–344085 and 344559–344577 respectively.

For *Choristoneura longicellana* (NC_019996.1), the invertebrate primers match 1474–1498 and 2155–2180 respectively.

3.e. **What value do you get if you calculate the fragment size for other species that have matches to the forward and reverse primer? Do you get the same number?**

The length range of the products produced from the primers will be between 450 to 800 nucleotides.

For the plant primers, using *Silene Conoidea* (NC_023358.1) as an example gives 43709 – 43111 = 599 nucleotides. These are the absolute nucleotide coordinates for this blast hit, and the total length will vary. The range in possible lengths should be between 550 and 600 nucleotides.

For the vertebrate (non-fish) primers, *Pucrasia macrolopha* (NC_020587.1) as an example gives 7298 – 6589 = 710 nucleotides.

For the fish primers, *Mallotus villosus* (NC_015244.1) as an example gives 6258

– 5556 = 703 nucleotides.

For the fungi primers, *Candida orthopsilosis* (NC_018301.1) as an example gives 344577 – 344066 = 512 nucleotides.

4.c. **Identify the feature(s) located between the nucleotide positions identified by the primers, as determined in 3.b. above.**

Depending on the hit, the name of features may vary. However, for plant primers, the feature is usually a gene named *rbc*L that codes for a product called "ribulose 1,5-bisphosphate carboxylase/oxygenase large subunit." For the vertebrate (non-fish), fish, and invertebrate primers, the feature is usually a gene named *COI* or COXI, which codes for cytochrome C oxidase subunit I. For the fungi primers, the feature is usually the nuclear internal transcribed spacer (*ITS*), a variable region that surrounds the 5.8s ribosomal RNA gene.

II. Identify Species and Phylogenetic Relationships Using *DNA Subway*

2.a. **What is the error rate and accuracy associated with a *Phred* score of 20?**

A *Phred* score of 20 equals 1 error in 100 or 99% accuracy.

What do you notice about the electropherogram peaks and quality scores at nucleotide positions labeled N?

At N positions, peaks representing different nucleotides have similar amplitudes (heights) and overlap, or no single peak rises above the background of lower amplitude peaks. Quality scores are very low.

2.b. **Why is it important to remove excess Ns from the ends of the sequences?**

Each N is scored as a misalignment, causing experimental sequences to appear to be less related to reference sequences than they actually are. This will significantly impact tree building, potentially placing related sequences in different clades.

3.e. **How does the consensus sequence optimize the amount of sequence information available for analysis? Why does this occur?**

The consensus sequence extends the length of the sequence and improves the accuracy of the sequence in regions where one read is of low quality. Sequence immediately following each primer has many errors and this sequence should be trimmed from the results. The read from the opposite strand usually extends into this region and provides data for the sequence at either end of the amplicon that would otherwise be lost. Also, the sequence quality can be low at different positions because of high GC content or other characteristics of the DNA. Often, the sequence quality from one direction is better than from the other direction. By selecting the best sequence for these regions, the overall quality of the consensus will be better than either forward or reverse sequences.

3.f. **Do differences tend to occur in certain areas of the sequence? Why?**

Differences cluster at the 5' and 3' ends because the sequence quality at the ends is poor.

4.b. **Why do the most significant hits typically have E values of 0? (This is not the case with BLAST searches with primers.) What does it mean when there are multiple BLAST hits with similar E values?**

The lower the E value, the lower the probability of a random match and the higher the probability that the BLAST hit is related to the query. Searching with a long (500 bp or more) barcode sequence increases the number of significant alignments with high scores compared to searches with short primers. It is common to have multiple hits with identical or very similar E values. Of course, identical matches to the same species would be expected to have an E value of zero. However, other hits with 0 or very low E values are often found for members of the same genus. In some families of plants, fungi, or animals, the barcode regions used in this experiment are not variable enough to make a conclusive species determination. Similar E values would also be obtained when two sequences have the same number of sequence differences, but at different positions.

6.b. **What causes these problems?**

The quality of sequences may be low at either end, contributing to gaps and Ns, and the length of the sequences in the databases may also be of different lengths, which can lead to gaps.

6.c. **Why is it important to remove sequence gaps and unaligned ends?**

Gaps and unaligned ends are scored as mismatches by the tree-building algorithms, making sequences appear less related than they actually are, forcing related sequences into different clades.

7.c. **What assumptions are made when one infers evolutionary relationships from sequence differences?**

The major assumption is that mutations occur at a constant rate; the "molecular clock" provides the measure of evolutionary time. Since branch lengths of a phylogenetic tree represent mutations per unit of time, an increase in the mutation rate at some point in evolutionary time would artificially lengthen branch lengths. If the barcode region mutates more frequently in one clade, then a larger number of differences would be incorrectly interpreted as increased phylogenetic distance between it and other clades. Also, although there is a chance that any given nucleotide has undergone multiple substitutions (for example A>T>C or A>T>A), tree-building algorithms only evaluate nucleotide positions as they occur in the sequences being compared. If the sequences being evaluated do not include a variation that happened during evolution, it will not be taken into account, and the algorithm will assume the minimum number of substitutions. Since the chance of multiple substitutions increases over time, the phylogenetic tree will tend to overestimate relatedness between distantly related species that diverged extremely long ago.

Why do gene and phylogenetic trees sometimes disagree?

Traditional phylogenetic trees are primarily based on morphological (physical) features. Related clades share morphological features by descent from a common ancestor. However, unrelated groups may develop a similar morphological feature when they independently adapt to similar challenges or environments. (For example, bats and birds have wings, but this feature arose independent of a common ancestor.) Gene trees can call attention to situations—at many taxonomic levels—where morphological similarities have been misinterpreted as a close phylogenetic relationship. Also, gene trees may identify new species that cannot

be differentiated by morphology alone.

7.e. **How does it compare to the maximum likelihood tree? What does this tell you?**

The trees will likely have a different arrangement of nodes and place some sequences on different nodes. This tells you that there are multiple possible solutions for most phylogenetic trees, and different algorithms will calculate different optimum trees.

PLANNING AND PREPARATION

The following table will help you to plan and integrate the different experimental methods.

Experiment Part	Day	Time	Activity	
I. Collect, Document, and Identify Specimens	-1	varies	**Lab:**	Collect tissue or processed material
II. Isolate DNA from Plant, Fungal, or Animal Samples	1	30-60 min	**Pre-lab:**	Aliquot distilled water, lysis solution, silica resin, and wash buffer
				Set up student stations
		80 min	**Lab:**	Isolate DNA
III. Amplify DNA by PCR	2	15 min	**Pre-lab:**	Prepare and aliquot primer mix
				Set up student stations
		10 min	**Lab:**	Set up PCR reactions
		70 min	**Post-Lab:**	Amplify DNA in thermal cycler
IV. Analyze PCR Products by Gel Electrophoresis	3	30 min	**Pre-lab:**	Dilute TBE electrophoresis buffer
				Prepare agarose gel solution
				Set up student stations
		30 min	**Lab:**	Cast gels
		45+ min		Load DNA samples into gels
				Electrophorese samples
				Photograph gels

Collecting Samples

Brainstorm in class to identify one or more organized campaigns that students can be involved with. Students may select samples of their own choosing; this can be done as a homework assignment or in class if the season permits. Alternatively, teachers can provide samples.

Obtain permission to collect on private property, parks, or nature preserves.

One application of DNA barcoding is to survey species from a particular location or habitat. Since accounting for every plant and animal in a habitat is usually impossible, samples are collected to generally represent the habitat. A common sampling unit is a quadrat, a 1-meter square frame that is laid over the ground and from which each different plant and animal is collected for barcoding. Quadrats make it possible to compare samples from different locations or habitats. Nets are useful for collecting flying insects or swimming invertebrates. A sample of freshwater or marine organisms can be strained from a defined amount of water.

Plants and Fungi

Avoid collecting woody parts, which are difficult to break up, and starchy storage tissue, which includes metabolites that may interfere with PCR. If fresh green plants are not available, DNA is readily isolated from frozen or dried material. Students may also bring in items from the grocery store. Fresh produce works well, and many processed foods containing plant material will also work. It is difficult to isolate DNA from fatty or oily foods, such as peanut butter.

For fungi, obtain fruit bodies (such as mushrooms) when possible. Avoid contamination by other fungi, such as moldy mushrooms or multiple species growing together. Fresh samples from soft mush-

rooms work well for DNA isolation, while dried samples and hard fungi give variable results. Fungal fruiting is weather and climate dependent, so their abundance will vary, although both fresh and dried mushrooms are readily available from stores.

Animals

Insects offer great opportunities for barcoding. A kill jar is a simple and humane way to collect and kill insects. Make a kill jar from a wide-mouth plastic jar with tight fitting lid. Cut enough discs of paper towel to make a ½ inch stack in the bottom of the jar, then soak the toweling with acetone (nail polish remover). Keep the jar tightly capped, away from flames. Alternatively, kill insects by placing them in the freezer for at least one hour. Larger animals may be safely sampled, and without injury, by isolating DNA from hair, feathers, or dung. Hair roots (follicles) and flesh scraped from the base of the feather shaft are reliable sources. Fresh meats and fish from the grocery store are good sources of DNA for barcoding. Many processed foods, and food scraps obtained at no cost, are good sources of DNA.

Isolating DNA

Part II works well for plant, fungal or animal samples, and uses a silica resin DNA isolation method.

Alternates Part IIa and b (online only) work well for animal and plant or fungal samples, respectively, but use reagents from the Qiagen® DNeasy Blood, and Tissue Kit catalogue number 69506 (250 preps) and Qiagen® DNeasy Plant Kit, catalogue number 69106 (250 preps).

Ethylenediaminetetraacetic Acid (EDTA) (0.5 M, pH 8.0)

Makes 100 mL.
Store at room temperature (indefinitely).

1. Add 18.6 g of EDTA (disodium salt dihydrate, MW 372.24) to 80 mL of deionized or distilled water.
2. Adjust the pH by slowly adding ~2.2 g of sodium hydroxide (NaOH) pellets (MW 40.00); monitor with a pH meter or strips of pH paper. (If neither is available, adding 2.2 g of NaOH pellets will make a solution of ~pH 8.0.)
3. Mix vigorously with a magnetic stirrer or by hand.
4. Add deionized or distilled water to make a total volume of 100 mL of solution.
5. Make sure that the bottle cap is loose and autoclave for 15 min at 121°C.
6. After autoclaving, cool the solution to room temperature and tighten the lid for storage.

Note: Use only the disodium salt of EDTA. EDTA will only dissolve after the pH has reached 8.0 or higher.

6 M Guanidine Hydrochloride Solution

Makes 100 mL.
Store at room temperature (for 6 months).

1. Dissolve 57.32 g of guanidine hydrochloride (m.w. = 95.53) in 50 mL of deonized or distilled water.
2. Add deionized or distilled water to make a total volume of 100 mL of solution.

Silica Resin Solution

Makes 50 mL.
Store at 4°C.

1. Dissolve 25 g of silicon dioxide (m.w. = 60.08) in 35 mL of deionized or distilled water.
2. Add deionized or distilled water to make a total volume of 50 mL of solution.

Note: The silica resin must be rinsed with 50 mL distilled water 3-4 times by centrifugation prior to bringing the total volume to 50 mL.

Sodium Chloride (NaCl) (5 M)

Makes 500 mL.
Store at room temperature (indefinitely).

1. Dissolve 146.1 g of NaCl (MW 58.44) in 250 mL of deionized or distilled water.
2. Add deionized or distilled water to make a total volume of 500 mL of solution.

Tris/EDTA (TE) Buffer

Makes 100 mL.
Store at room temperature (indefinitely).

1. In a 200-mL beaker mix the following:

   99 mL of deionized or distilled water
   1 mL of 1 M Tris pH 8.0
   200 μL of 0.5 M EDTA
2. Mix well.

Tris-HCl (1 M, pH 8.0 and 8.3)

Makes 100 mL.
Store at room temperature (indefinitely).

1. Dissolve 12.1 g of Tris base (MW 121.10) in 70 mL of deionized or distilled water.
2. Adjust the pH by slowly adding concentrated hydrochloric acid (HCl) for the desired pH listed below.

   pH 8.0: 5.0 mL
   pH 8.3: 4.5 mL
3. Monitor with a pH meter or strips of pH paper. (If neither is available, adding the volumes of concentrated HCl listed here will yield a solution with approximately the desired pH.)
4. Add deionized or distilled water to make a total volume of 100 mL of solution.
5. Make sure that the bottle cap is loose and autoclave for 15 min at 121°C.
6. After autoclaving, cool the solution to room temperature and tighten the lid for storage.

Note: A yellow-colored solution indicates poor-quality Tris. If your solution is yellow, discard it and obtain a Tris solution from a different source. The pH of Tris solutions is temperature dependent, so make sure to measure the pH at room temperature. Many types of electrodes do not accurately measure the pH of Tris solutions; check with the manufacturer to obtain a suitable one.

Wash Buffer

Makes 500 mL.
Store at -20°C (indefinitely).

1. Combine the following:

   Deionized or distilled water, 234 mL
   1 M Tris (pH 7.4), 10 mL
   5 M NaCl, 5 mL
   0.5 M EDTA, 1 mL
   100% Ethanol, 250 mL

2. Mix thoroughly.

Primer Strategy and Design

DNA barcoding relies on finding a universal chromosome location (locus) that has retained enough sequence conservation through evolutionary history that it can be identified in many organisms, but that also has enough sequence diversity to differentiate organisms to at least the family level. Regions of the chloroplast *rbc*L gene, mitochondrial *COI* gene, and nuclear *ITS* region generally fulfill these requirements.

Primers are designed to target conserved sequences that flank the variable barcode regions. However, even the conserved flanking regions have accumulated enough sequence differences over evolutionary time that it is impossible to identify universal primer sets that will work across all taxonomic groups of plants and animals. Thus, barcode primers often need to accommodate sequence variation, or degeneracy, at one or several nucleotide positions.

The degeneracy problem is often solved when the oligonucleotide primers are synthesized. Traditionally, a mixture of primers is synthesized – each having a different nucleotide in any of the variable positions. However, synthetic nucleotides are now available that pair with multiple nucleotides and can be incorporated at variable positions in a single primer.

The table below shows the letter abbreviation given for degenerate nucleotides. For example, a primer with the sequence 'ATCC**R**' contains both ATCC**A** and ATCC**G**.

W	=	A or T	B	=	C or G or T
S	=	G or C	D	=	A or G or T
M	=	A or C	H	=	A or C or T
K	=	G or T	V	=	A or C or G
R	=	A or G	N	=	A or C or G or T
Y	=	C or T			

Even degenerate primers cannot ensure amplification in taxonomic groups in which all or part of a particular primer sequence is deleted. Thus, broad surveys of unknown plants or animals typically employ multiple primer sets against slightly different flanking regions, which are combined in a multiplex PCR reaction.

The *rbc*L primer set used in this laboratory will work well for most green plants. We suggest starting with one of three *COI* primer sets for animals: one for fish, one for vertebrates, and one for other invertebrates (DMI). For fungal species, use the *ITS* primer set, which has the highest chance of success for identifying a broad range of fungi. These primer sets will not uniformly work across all groups. If the primers in this laboratory do not work with a group of organisms you are studying, consult the primer list (http://www.boldsystems.org/views/primerlist.php) at the Barcode of Life Online Database web site for alternatives.

Barcode Primer Sequences

Plant cocktail

5'-TGTAAAACGACGGCCAGTATGTCACCACAAACAGAGACTAAAGC-3' (forward primer–rbcLaf-M13)
5'-CAGGAAACAGCTATGACGTAAAATCAAGTCCACCRCG-3' (reverse primer–rbcLa-revM13)

Fish Cocktail

5'-TGTAAAACGACGGCCAGTCAACCAACCACAAAGACATTGGCAC-3' (forward primer–VF2_t1)
5'-TGTAAAACGACGGCCAGTCGACTAATCATAAAGATATCGGCAC-3' (forward primer–FishF2_t1)
5'-CAGGAAACAGCTATGACACTTCAGGGTGACCGAAGAATCAGAA-3' (reverse primer–FishR2_t1)
5'-CAGGAAACAGCTATGACACCTCAGGGTGTCCGAARAAYCARAA-3' (reverse primer–FR1d_t1)

Vertebrate cocktail (non-fish)

5'-TGTAAAACGACGGCCAGTTCTCAACCAACCACAAAGACATTGG-3' (forward primer–VF1_t1)
5'-TGTAAAACGACGGCCAGTTCTCAACCAACCACAARGAYATYGG-3' (forward primer–VF1d_t1)
5'-TGTAAAACGACGGCCAGTTCTCAACCAACCAIAAIGAIATIGG-3' (forward primer–VF1i_t1)
5'-CAGGAAACAGCTATGACTAGACTTCTGGGTGGCCRAARAAYCA-3' (reverse primer–VR1d_t1)
5'-CAGGAAACAGCTATGACTAGACTTCTGGGTGGCCAAAGAATCA-3' (reverse primer–VR1_t1)
5'-CAGGAAACAGCTATGACTAGACTTCTGGGTGICCIAAIAAICA-3' (reverse primer–VR1i_t1)

Invertebrate cocktail

5'-TGTAAAACGACGGCCAGTGGTCAACAAATCATAAAGATATTGG-3' (forward primer LCO1490)
5'-CAGGAAACAGCTATGACTAAACTTCAGGGTGACCAAAAAATCA-3' (reverse primer HC02198)

Fungi Cocktail

5'-TGTAAAACGACGGCCAGTTCCGTAGGTGAACCTGCGG-3' (ITS1 F)
5'-CAGGAAACAGCTATGACTCCTCCGCTTATTGATATGC-3' (ITS4 R)

Ready-to-Go PCR Beads

Ready-To-Go™ PCR beads limit reagent waste and optimize PCR reactions in a classroom setting. Each bead contains reagents so that when brought to a final volume of 25 µL, the reaction contains 2.5 units of *Taq* DNA polymerase, 10 mM Tris-HCl (pH 9.0), 50 mM KCl, 1.5 mM MgCl2, and 200 µM of each dNTP.

The lyophilized *Taq* DNA polymerase in the bead becomes active immediately upon addition of the primer/loading dye mix and template DNA. In the absence of thermal cycling, "nonspecific priming" at room temperature allows the polymerase to begin generating erroneous products, which can show up as extra bands in gel analysis. Therefore, work quickly. Be sure the thermal cycler is set and have all experimenters set up their PCR reactions as a coordinated effort. Add primer/loading dye mix to all reaction tubes, then add each student template, and begin thermal cycling as quickly as possible. Hold reactions on ice until all student samples are ready to load into the thermal cycler.

NEB *Taq* 2X Master Mix

The NEB Taq 2X master mix is a cost-effective alternative to PCR beads and works well in a classroom setting. Taq 2X Master Mix is an optimized ready-to-use solution containing *Taq* DNA Polymerase, dNTPs, $MgCl_2$, KCI and stabilizers. The Master Mix is used at a 1X final concentration with DNA template and primers in a

total reaction volume of 25 µL. It is stable for fifteen freeze-thaw cycles when stored at -20°C or for three months at 4°C, so for frequent use, an aliquot may be kept at 4°C.

Primer/Loading Dye Mix (for Ready-to-Go PCR Beads)

The primer/loading dye mix customizes the PCR reaction for DNA barcoding. The mix incorporates the appropriate primer pair (0.26 picomoles/µL of each primer), 13.8% sucrose, and 0.0081% cresol red. The inclusion of the loading dye components, sucrose and cresol red, allows the amplified product to be directly loaded into an agarose gel for electrophoresis.

Makes enough for 50 reactions. Store at –20°C for 1 year.

Mix in a 1.5-ml tube:

640 µL of distilled water
460 µL of Cresol Red Loading Dye (see recipes below)
20 µL of 15 pmol/ µL 5' primer
20 µL of 15 pmol/ µL 3' primer

(For multiplex primers, add 20 µL of each primer, and reduce volume of distilled water by 20 µL for each additional primer.)

Primer/Loading Dye Mix (for NEB Taq 2X Master Mix (#M0270)

The primer/loading dye mix customizes the PCR reaction for DNA barcoding. The mix incorporates the appropriate primer pair (0.526 picomoles/µL of each primer), 13.8% sucrose, and 0.0081% cresol red. The inclusion of the loading dye components, sucrose and cresol red, allows the amplified product to be directly loaded into an agarose gel for electrophoresis.

Makes enough for 50 reactions. Store at –20°C for 1 year.

1. Mix in a 1.5-ml tube:

   600 µL of distilled water
   460 µL of Cresol Red Loading Dye (see recipes below)
   40 µL of 15 pmol/ µL 5' primer
   40 µL of 15 pmol/ µL 3' primer

   (For multiplex primers, add 40 µL of each primer, and reduce volume of distilled water by 40 µL for each additional primer.)

Alternative PCR protocol for NEB *Taq* 2X Master Mix

1. Obtain a PCR tube and use a micropipette with a fresh tip to add 12.5 µLof the master mix to each tube.
2. Use a micropipette with a fresh tip to add 10.5 µL of the appropriate primer/loading dye mix (for NEB *Taq* 2X Master Mix) to each tube. Mix well by pepetting up and down.

Plantcocktail:	*rbc*L primers (rbcLaF / rbcLa rev)
Fungi cocktail:	*ITS* primers (ITS1F/ITS4)
Fish cocktail:	*COI* primers (VF2_t1/ FishF2_t1/ FishR2_t1/ FR1d_t1)
Vertebrate (non-fish):	(VF1_t1/ VF1d_t1/ VF1i_t1 / VR1d_t1/ VR1_t1/ VR1i_t1)
Invertebrate cocktail:	(LCO1490/ HC02198)

3. Use a micropipette with fresh tip to add 2 µL of DNA (from Part II) directly into the appropriate primer/loading dye mix. Ensure that no DNA remains in the tip after pipetting. Mix well by pepetting up and down.
4. Store your sample on ice until your class is ready to begin thermal cycling.
5. Place your PCR tube, along with those of the other students, in a thermal cycler that has been programmed with the appropriate PCR protocol.

1% Cresol Red Dye

Makes 50 mL.
Store at room temperature (indefinitely).

1. Mix in a 50-mL tube:

   500 mg cresol red dye
   50 mL of distilled water

Cresol Red Loading Dye

Makes 50 ml.
Store at –20°C (indefinitely).

1. Dissolve 17 g of sucrose in 49 mL of distilled water in a 50-mL tube.
2. Add 1 mL of 1% cresol red dye and mix well.

Thermal Cycling

Amplification of *rbc*L and *COI* is simplified by the large number of chloroplast and mitochondrial genomes, which are present at 100-1,000s of copies per cell. *ITS* is also present at high copy number in most fungi. Thus, the barcode regions are amplified more readily than most nuclear loci and small amount of specimen collected provides enough starting template to produce large quantities of the target sequence, reducing the concentration of contaminants that might inhibit PCR. The recommended amplification times and temperatures will work adequately for most common thermal cyclers, which ramp between temperatures within a single heating/cooling block. IMPORTANT: Follow manufacturer's instructions for Robocycler or other brands of thermal cyclers that physically move PCR reaction tubes between multiple temperature blocks. These machines have no ramping time between temperatures, and may require longer cycles.

Troubleshooting for Failed PCR Reactions

When PCR reactions fail, there are many possible reasons. A common source of difficulty is low quality DNA caused by using too much sample for the isolation. If you suspect your students have used too much material and their PCR failed, consider re-isolating the DNA with the standard protocol while ensuring the students use less material. For dried, degraded, or processed samples, PCR may fail due to low yield, in which case using the appropriate alternative method may allow amplification. For the silica isolation, other possible sources of trouble include evaporation of ethanol from the wash buffer, which can be avoided by storing the wash buffer at low temperature in a sealed container, and failing to remove wash buffer before elution, which can be avoided by carefully removing the wash buffer and drying briefly before elution. Mixing the DNA/silica mixture during incubation may also increase yields.

The PCR may also fail due to changes in the sequence at the primer binding sites, making it impossible to

amplify even with high quality DNA. Consulting the literature on the taxa you are studying may help determine whether this is the case. It may also help to re-amplify after lowering the annealing temperature a few degrees, as this may allow the primers to anneal even if the primer binding sites have mutated.

Gel Electrophoresis

CAUTION: Be sure to electrophorese only 5 µL of each amplified product. The remaining 20 µL must be retained for DNA sequencing: 10 µL for the forward read and, potentially, 10 µL for the reverse read.

Plasmid pBR322 digested with the restriction endonuclease BstNI is an inexpensive marker and produces fragments that are useful as size markers in this experiment. The size of the DNA fragments in the marker are 1,857 bp, 1,058 bp, 929 bp, 383 bp, and 121 bp. Use 20 µL of a 0.075 µg/µL stock solution of this DNA ladder per gel. Other markers or a 100-bp ladder may be substituted.

View and photograph gels as soon as possible after electrophoresis or appropriate staining/destaining. Over time, the small-sized PCR products will diffuse through the gel and the bands they form will lose sharpness.

DNA Sequencing

DNA sequencing of the *rbc*L, *COI* or *ITS* amplicon is required to determine the nucleotide sequence that constitutes the DNA barcode. The forward, the reverse, or both DNA strands of the amplified barcode region may be sequenced. A single, good-quality barcode from the forward strand is sufficient to identify an organism. The majority of database sequences are from the forward strand, so sequencing only the forward strand reduces sequencing cost and simplifies analysis. If you only do a forward read, save the remaining 10 µL of amplicon. If the forward read fails, and time permits, you can send the remainder out to sequence the reverse strand.

However, bi-directional sequencing is important for several reasons. 1) A reverse sequence may provide a readable barcode when the forward sequence fails. 2) Good forward and reverse reads can be combined to produce a consensus sequence that extends the read up to 40 or more nucleotides. This is because the primer itself is not sequenced for either strand, and additional nucleotides downstream from the primer are typically unreadable. Thus, good forward and reverse primers complement these missing sequences, adding most of the primer sequences on either end. 3) One direction may provide a read through a region that is refractory to sequencing in the other direction, such as a homopolymeric region containing a long string of C residues. Thus, the insurance provided by bi-directional sequencing may be worth the added cost, especially if you have need to complete an analysis in a limited time.

Sequencing different barcode regions – *rbc*L, *COI* and *ITS* – and using degenerate and multiplex primers complicate DNA sequencing. Strictly speaking, each different primer would need to be provided for forward and reverse sequencing reactions. As a work-around for this problem, the primers used in this experiment incorporate a universal M13 primer sequence. In addition to a sequence specific to the *rbc*L, *COI* or *ITS* barcode locus, the 5' end of each primer has an identical 17 or 18 nucleotide sequence from the bacteriophage vector M13.

In the traditional approach to genome sequencing, genomic DNA is cloned into an M13 vector. Then a universal M13 primer is used to sequence the genomic insert just downstream from the primer. This same strategy is used in sequencing rbcL, COI and ITS barcodes in this experiment. During the first cycle of PCR, the M13 portion of the primer does not bind to the template DNA. However, the entire primer sequence is covalently linked to the newly-synthesized DNA and is amplified in subsequent rounds of PCR. Thus, the M13 sequence is included in every full-length PCR product. This allows a sequencing center to use universal forward and reverse M13 primers for the PCR-based reactions that prepare any *rbc*L, *COI,* or *ITS* amplicon for sequencing.

The sequence of the M13 forward and reverse primers are:

M13F(-21): TGTAAAACGACGGCCAGT

M13R(-27): CAGGAAACAGCTATGAC

Using Genewiz DNA Sequencing Services

We recommend using GENEWIZ, Inc. for DNA barcode sequencing. GENEWIZ has optimized reaction conditions for producing the barcode sequences in this laboratory and produces excellent quality sequence with rapid turnaround –usually within 48 hours of receipt of samples. GENEWIZ sequences are automatically uploaded to the DNALC's *DNA Subway* website.

Before submitting samples for sequencing, consult the GENEWIZ guide.

Prepare PCR Products

1. Verify that you can see a PCR product of the correct predicted length on an agarose gel. DNA sequence can be obtained from virtually every PCR product that is visible on the gel. (You can take a chance on samples that do not produce bands, as a fair proportion of these will also produce sequences.)
2. Prepare 8-strips of 0.2 mL PCR tubes appropriate for the number of samples you wish to submit. If you will be submitting a large number of samples (≥48), submit the samples on a 96 well plate arranging them vertically (A1 to H1). See the "Tubes and Plates" tab for more details at www.GENEWIZ.com.
3. You must submit 10 µL of PCR product for each sequencing reaction. Forward and reverse sequencing reactions must be submitted in separate tubes.

Register and Submit Samples for Sequencing

1. Go to www.GENEWIZ.com and click "Register" to create a user account.
2. When creating your account enter your institution name followed by "–DNALC" (very important!). The suffix "-DNALC" must be added exactly to your name, otherwise your sequence will not be processed properly or may be delayed.
3. Obtain a valid Purchase Order number from your Purchasing department, or use a valid credit card.
4 .Log in to your user account to place your sequencing order. Under "Place an Order," select "Create Sequencing Order."
5. Under "Service Priority:" select "Standard."
6. Under "Create Order by:" select "Online Form." (Alternatively, you can select "Upload Excel Form," then download the "Custom" GENEWIZ Excel template, fill in the information in Steps 10-18, and upload the file.)
7. Under "Sample Type," select "Custom."
8. When prompted to "Create and online form for," enter the number of samples you will be sending for sequencing. (If you elect to do bi-directional sequencing, you need to count separate forward and reverse reactions for each sample).
9. Click on "Create New Form," and a sample submission form will be displayed.
10. For "DNA Name," enter a name for each sample. This may be a number or initials.
11. For "DNA Type," select "Un-Purified PCR" from the dropdown menu.
12. For "DNA Length (vector + insert in bp)," enter 650 for rbcL or 800 for COI and ITS.
13. Leave "DNA Conc. (ng/uL)" blank. It is best to send in a gel image of representative samples. This will be used by GENEWIZ to calculate the correct amount of clean up reagents to use and the amount of product to use in the sequencing reaction. If a gel image is not supplied, GENEWIZ will use default amounts to set

up the sequencing reactions.

14. Leave "My Primer Name" and "My Primer Conc." blank.
15. For "GENEWIZ Primer," select from the dropdown menu: "M13F(-21)" – to sequence the forward strand. "M13R" – to sequence the reverse strand.
16. Under "Special Request," be sure that "PCR-Clean Up" has been automatically selected. (This is the default when unpurified PCR is selected as the DNA type).
17. In the "Comments" box at the bottom of the form, type "Primer stored at GENEWIZ under DNALC."
18. Click on "Save & Next."
19. Carefully review your form, then click on "Next Step."
20. Enter your payment information, and click on "Next Step."
21. Review your order, then click on "Submit."

Ship Samples to GENEWIZ

1. Print a copy of the order form, and mail it along with your samples.
2. Be sure that the tubes are labeled exactly the same in the gel photo and on the order form. Failure to do so may delay sequencing or make it impossible to complete. Email DNALCSeq@cshl.edu if you need help.
3. Ship your samples via standard overnight delivery service (Federal Express, if possible).
4. Pack your samples in a letter pack or small shipping box, padding samples to prevent too much shifting. Room temperature shipping – with no ice or ice pack – is expected. PCR products are stable at ambient temperature, even if shipped on a Friday for Monday delivery.
5. Address the shipment to GENEWIZ at the following location:

   GENEWIZ, Inc.
   115 Corporate Blvd.
   South Plainfield, NJ 07080
6. You may be able to reduce shipping costs by using a GENEWIZ drop box. Call 1-877-436-3949 to find out if one is available in your area.