LABORATORY

Using DNA Barcodes to Identify and Classify Living Things

OBJECTIVES

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

- Collect and analyze sequence data from plants or animals – or products 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, 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 or mitochondrial 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

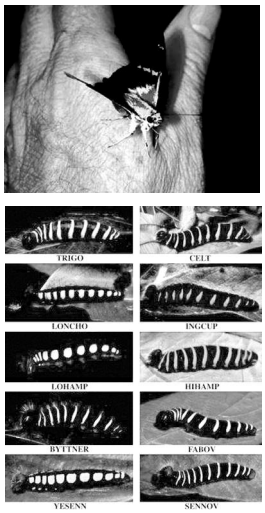
Taxonomy, the science of classifying living things according to shared features, has always been a part of human society. Carl Linneas formalized biological classification with his system of binomial nomenclature that assigns each organism a genus and species name.

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 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 distinguishing 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 and animal groups – including ants, bees, butterflies, fish, birds, mammals, fungi, 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 mitochondrial and chloroplast genomes. 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 photosynthesis. 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 barcode genes are involved in the key reactions of life: storing energy in glucose and releasing it to form ATP.

This laboratory uses DNA barcoding to identify plants 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 sample of muscle are suitable sources. DNA is extracted from the tissue sample, and the barcode portion of the *rbc*L or *COI* gene is amplified by polymerase chain reaction (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 the database at the Barcode of Life Data System (BOLD) (www.boldsystems.org/) at the University of Guelph.

FURTHER READING

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., 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.

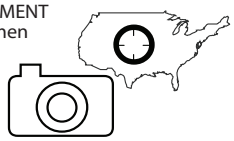
Van Den Berg C., Higgins W.E., Dressler R.L., Whitten W.M., Soto-Arenas M.A., Chase M.W. (2009) A phylogenetic study of laeliinae (*orchidaceae*) based on combined nuclear and plastid DNA sequences. *Annals of Botany* 104(3): 417-30.

OVERVIEW OF EXPERIMENTAL METHODS

I. COLLECT, DOCUMENT, AND IDENTIFY SPECIMENS

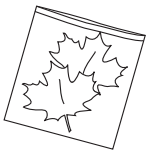
COLLECT specimen

DOCUMENT specimen



IDENTIFY specimen

COLLECT tissue sample

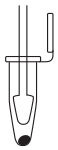


II. ISOLATE DNA FROM PLANT OR ANIMAL TISSUE

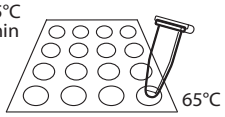
ADD specimen tissue sample

ADD Nuclei Lysis solution

GRIND sample in solution

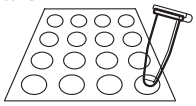


INCUBATE sample at 65°C 15 min



ADD RNAse

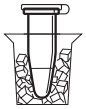
INCUBATE at room tempeture 5 min



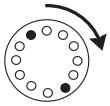
ADD Protein Precipitation solution

VORTEX

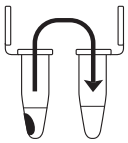
CHILL on ice 5 min



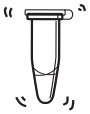
CENTRIFUGE 4 min



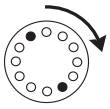
TRANSFER to fresh tube with isopropanol



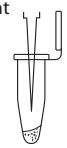
MIX



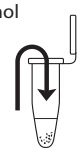
CENTRIFUGE 1 min



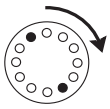
REMOVE supernatant



ADD ethanol



CENTRIFUGE 1 min



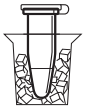
REMOVE ethanol

DRY pellet 10 min

ADD Rehydration solution

REHYDRATE at 65°C 60 min or 4°C overnight

STORE at -20 °C

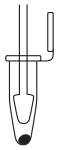


IIa. ISOLATE DNA FROM PLANT TISSUE (ALTERNATE

ADD plant tissue

ADD Edward's buffer

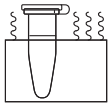
GRIND



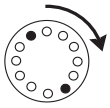
ADD Edward's buffer

VORTEX

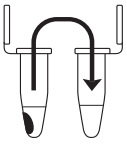
BOIL on heat block 5 min



CENTRIFUGE (2 min)

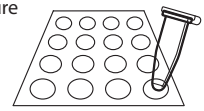


TRANSFER supernatant

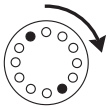


ADD and MIX isopropanol

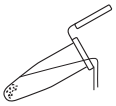
INCUBATE
at room
tempeture
5 min



CENTRIFUGE
5 min

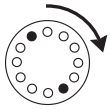


POUR OFF
supernatant

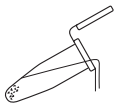


ADD and MIX
ethanol

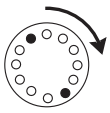
CENTRIFUGE
1 min



POUR OFF
supernatant



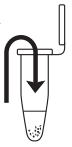
CENTRIFUGE
30 sec.



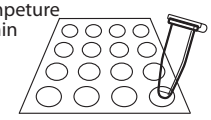
REMOVE
supernatant

DRY
pellet
10 min

ADD
TE/RNAse A



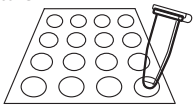
INCUBATE
at room
tempeture
5 min



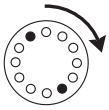
ADD and MIX
water and
sodium
acetate

ADD and MIX
isopropanol

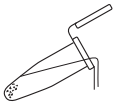
INCUBATE
at room
tempeture
3 min



CENTRIFUGE
(5 min)

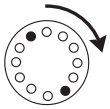


POUR OFF
supernatant

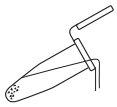


ADD and MIX
ethanol

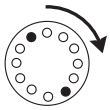
CENTRIFUGE
1 min



POUR OFF
supernatant



CENTRIFUGE
30 sec.



REMOVE
supernatant

DRY
pellet
10 min

ADD
water

STORE
at -20 °C

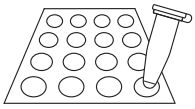
III. AMPLIFY DNA BY PCR

ADD
primer
mix

ADD
DNA

ADD
mineral oil
(if necessary

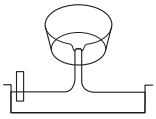
AMPLIFY
in thermal
cycler



STORE
at -20 °C

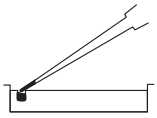
IV. ANALYZE PCR PRODUCTS BY GEL ELECTROPHORESIS

POUR
gel

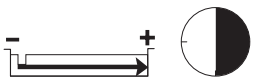


SET
(20 min)

LOAD
gel

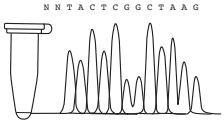


ELECTROPHORESE
130 volts
(30 min)

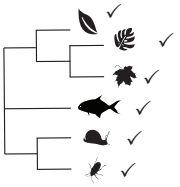


SEQUNCE PCR PRODUCT AND ANALYZE RESULYS

SEND
sample
for
sequencing



ANALYZE
results
using
bioinformatics



PLANNING AND PREPARATION

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

Experiment	Day	Time		Activity
I. Collect, Document, and Identify Specimens	-1	varies	Lab:	Collect tissue or processed material
II. Isolate DNA from Plant or Animal Tissue	1	30-60 min	Pre-lab:	Aliquot nuclei lysis solution
				Aliquot RNAse solution
				Aliquot protein precipitation solution, aliquot ethanol
				Aliquot DNA rehydration solution
		60 min	Lab:	Isolate DNA
IIa. (Alternate) Isolate DNA from Plant orTissue	1	30-60 min	Pre-lab:	Prepare and aliquot Edwards buffer
				Prepare and aliquot TE-RNAse
				Aliquot sodium acetate
				Aliquot isopropanol
				Aliquot ethanol
				Make centrifuge adapters
				Set up student stations
		60 min	Lab:	Isolate DNA
III. Amplify DNA by PCR	2	15 min	Pre-lab:	Prepare and aliquot primer mix
		10 min	Lab	Set up PCR reactions
IV. Analyze PCR Products by Gel Electrophoresis	3	30 min	Prelab:	Dilute TBE electrophoresis buffer
				Prepare agarose gel solution
				Set up student stations
		30 min	Lab:	Cast gels
		45+ min		Load DNA samples into gel
				Electrophorese samples
				Photograph gels

EXPERIMENTAL METHODS

I. Collect, Document, and Identify Specimens

The DNA isolation and amplification methods used in this laboratory work for a variety of plants 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 BOLD.

Do not take more sample than you need. Only a small amount tissue is needed for DNA extraction – a piece of plant leaf about 1/4 inch in diameter or a piece animal 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 well ground. Dormant leaf buds can often be obtained from bushes and trees that have dropped their 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.

Small invertebrate animals, such as insects, can be collected whole and euthanized 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. Bone and skin are difficult. Fresh and frozen samples work equally well.

Other than fish, do not collect vertebrate animals. Use care if collecting from road killed animals, and avoid animal droppings that are possible vectors for disease.

REAGENTS, SUPPLIES, & EQUIPMENT

To Share

Tubes, collection jars, or bags
Tweezers, scalpel, and scissors
(Smart)phone with camera or digital camera with GPS

If you are participating in a collaborative project, you may be asked to follow a specific procedure to document and identify your specimen

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).
2. Use a (smart)phone 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.

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

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 applications (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 your *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 blue pin icon to create a pin, then drag it to the 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" 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 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/):
   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 or Animal Tissue

REAGENTS, SUPPLIES, & EQUIPMENT

For each group

Container with cracked or crushed ice
DNA rehydration solution (250 µL)
70% Ethanol (1.5 mL)
Isopropanol (1.5 mL)
4 microcentrifuge tubes (1.5 mL)
Micropipettes and tips (100-1000 µL)
*Nuclei Lysis solution (1.8 mL)**
Permanent marker
Protein Precipitation solution (0.5 mL)
RNAse solution (7 µL)
2 Plastic pestles
2 Specimen tissue samples (from Part I)

To share
Microcentrifuge
Water bath or heating block at 65°C
Vortexer (optional)

**Store on ice*

This universal DNA extraction method uses a commercial kit. Although it is more expensive than the alternate method for plants using Edward's buffer (see part IIa), it has the advantage of working reproducibly with almost any kind of plant or animal specimen.

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

1. Obtain 2 pieces of plant or animal tissue ~10-20 mg or ¼ inch diameter from two different samples. Be careful not to cross contaminate the specimens. (If you only have one specimen, make a duplicate prep to provide a balance for centrifuge steps.)
2. Place each sample in a clean 1.5 mL tube labeled with an identification number and your group number.
3. Add 100 µL of nuclei lysis solution to each tube.

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

4. Twist a clean plastic pestle against the inner surface of each 1.5 mL tube to *forcefully* grind the tissue for 1 minute. Use a clean pestle for each tube.

Grinding the plant tissue breaks up the cell walls. When fully ground, the sample should be a green liquid.

5. Add 500 µL more nuclei lysis solution to each tube.
6. Incubate the tubes in a water bath or heat block at 65°C for 15 minutes.
7. Add 3 µL of RNAse solution to each tube. Close caps, and mix by rapidly inverting tubes several times.

Step 7 degrades RNA that could interfere with PCR.

8. Incubate the tubes in a water bath or heat block at 37°C for 15 minutes. Then stand at room temperature for 5 minutes.
9. Add 200 µL of protein precipitation solution to each tube. Vortex tubes for 5 seconds by hand or machine (if available).

Step 9 causes many proteins to precipitate out of the solution, leaving DNA in the supernatant.

10. Stand tubes on ice for 5 minutes.

Centrifugation pellets the nucleic acids. The pellet may appear as a tiny teardrop-shaped smear or particles on the bottom side of the tube underneath the hinge. Do not be concerned if you cannot see a pellet. A large or greenish pellet is cellular debris carried over from the first centrifugation.

11. Place your tubes and those of other groups in a balanced configuration in a microcentrifuge, with cap hinges pointing outward. Centrifuge for 4 minutes at maximum speed to pellet protein.
12. Label 2 clean 1.5 mL tubes with your sample and group numbers. Use fresh tips to transfer 600 µl of supernatant for each sample to the appropriate clean tubes. Be careful not to disturb the pelleted debris when transferring the supernatant. Discard old tubes containing the precipitate.
13. Add 600 µL of isopropanol to the supernatant in each tube. Close caps, and mix by rapidly inverting tubes several times
14. Place your tubes and those of other groups in a balanced configuration in a microcentrifuge, with cap hinges pointing outward. Centrifuge for 1 minute at maximum speed to pellet the DNA.
15. Carefully pour off the supernatant from each tube, and add 600 µL of 70% ethanol. Close caps, and flick the bottom of each tube several times to "wash" the pellet.
16. Centrifuge the tubes for 1 minute at maximum speed.
17. Carefully pour off the supernatant. Use a micropipette with fresh tip to remove any remaining ethanol, being careful not to disturb the pellets.

Dry the pellets quickly with a hair dryer. To prevent blowing the pellet away, direct the air across the tube mouth, not into the tube.

18. Air dry the pellets for 10-15 minutes to evaporate remaining ethanol.
19. Add 100 µL of the DNA rehydration solution to each tube, and dissolve the DNA pellet by pipetting in and out several times.
20. Incubate the DNA at 65°C for 45-60 minutes, or overnight at 4°C.
21. Store your sample on ice or at -20°C until you are ready to begin Part III.

In Part III, you will use 2.5 mL of DNA for each PCR. 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.

IIa. Isolate DNA from Plant Tissue (Alternate)

REAGENTS, SUPPLIES, & EQUIPMENT

For each group
Container with cracked or crushed ice
Edward's buffer (2.2 mL)
70% Ethanol (2.2 mL)
Isopropanol (2.2 mL)
4 microcentrifuge tubes (1.5 mL)
Micropipettes and tips (100-1000 µL)
Permanent marker
Plant specimens
2 Plastic pestles
3M Sodium Acetate (150 µL)
Tris/EDTA (TE) buffer with RNase A (250 µL)
dH2O (1.5 mL)

To share
Microcentrifuge
Vortexer (optional)
Water bath or heating block

This method is optimized for plants. Although it takes about 20 minutes longer than the previous method, it uses readily available reagents.

1. Obtain two pieces of plant tissue ~¼ inch diameter from two different speci-

The large end of a 1000-µl pipette tip will punch disks of this size.

mens. Be careful not to cross-contaminate the specimens. (If you only have only one specimen, make a duplicate prep to provide a balance for centrifuge steps.)

2. Place each sample in a clean 1.5 mL tube labeled with an identification number and your group number.
3. Add 100 µL of Edward's buffer to each tube.
4. Grind the tissue for 1 minute by forcefully twisting a clean plastic pestle against the inner surface of each 1.5 mL tube. Use a clean pestle for each different sample.

Detergent in the Edward's buffer, sodium dodecyl sulfate (SDS), dissolves lipids of the cell membranes.

5. Add 900 µL more Edward's buffer to each tube, and grind briefly to remove tissue from the pestle.
6. Vortex the tubes for 5 seconds, by hand or machine (if available).

Step 7 denatures proteins, including enzymes that digest DNA.

7. Boil samples at 100°C for 5 minutes in a water bath or heating block.

Step 8 pellets insoluble material at the bottom of the tube.

8. Place the tubes, along with those from other groups, in a balanced configuration in a microcentrifuge, and centrifuge for 2 minutes to pellet any remaining cell debris. Centrifuge longer if there is still unpelleted debris.

Step 9 precipitates nucleic acids, including DNA.

9. Label 2 clean 1.5 mL tubes with your sample and group numbers. Use fresh tips to transfer 350 µL of supernatant for each sample to the appropriate fresh tubes. Be careful not to disturb the pelleted debris when transferring the supernatant. Discard old tubes containing the precipitate.
10. Add 400 µl of isopropanol to the supernatant in each tube. Close caps and mix by rapidly inverting several times.
11. Stand tubes at room temperature for 3 minutes.

Centrifugation pellets the nucleic acids. The pellet may appear as a tiny teardrop-shaped smear or particles on the bottom side of the tube underneath the hinge. Do not be concerned if you can't see a pellet. A large or greenish pellet is cellular debris carried over from the first centrifugation.)

12. Place your tubes and those of other groups in a balanced configuration in a microcentrifuge, with cap hinges pointing outward. Centrifuge for 5 minutes at maximum speed to pellet the DNA.
13. Carefully pour off the supernatant from each tube, and add 500 µL of 70% ethanol. Close caps, and flick the bottom of each tube several times to "wash" the pellet.
14. Place the tubes in a balanced configuration in a microcentrifuge, and spin for 1 minute. Align tubes in the rotor with the cap hinges pointing outward.

Nucleic acid pellets are not soluble in ethanol and will not dissolve during washing.

15. Carefully pour off the supernatant from each tube. Centrifuge the tubes again for 30 seconds to force any remaining ethanol to the bottom.
16. Use a micropipette to carefully remove the remaining ethanol from each tube. Be careful not to disturb the pellet.

Dry the pellets quickly with a hair dryer. To prevent blowing the pellet away, direct the air across the tube mouth, not into the tube.

17. Air dry the pellets for 10 minutes to evaporate remaining ethanol.
18. Add 100 µL of TE/RNaseA buffer to each tube. Dissolve the nucleic acid pellets by pipetting in and out. Take care to wash down the side of the tubes underneath the hinge, where the pellets formed during centrifugation. Use a fresh tip for each tube.

If needed, you may store DNA in TE/RNase solution at –20°C until ready to continue.

19. Incubate TE/RNaseA solution at room temperature for 5 minutes.
20. Add 400 µL of dH2O to each tube.

21. Add 50 µL of sodium acetate to each tube. Close caps, and mix by rapidly inverting tubes several times.
22. Add 550 µL of isopropanol to each tube to precipitate the DNA. Close caps, and mix by rapidly inverting tubes several times.
23. Stand tubes at room temperature for 3 minutes.

Nucleic acid pellets are not soluble in ethanol and will not dissolve during washing.

24. Place the tubes in a balanced configuration in a microcentrifuge, and spin for 5 minutes. Align tubes in the rotor with the cap hinges pointing outward.
25. Carefully pour off the supernatant from each tube, and add 500 µL of 70% ethanol. Close caps, and flick the bottom of each tube several times to "wash" the pellet.
26. Place the tubes in a balanced configuration in a microcentrifuge, and spin for 1 minute. Align tubes in the rotor with the cap hinges pointing outward.

Dry the pellets quickly with a hair dryer. To prevent blowing the pellet away, direct the air across the tube mouth, not into the tube.

27. Carefully pour off the supernatant from each tube. Centrifuge the tubes again for 30 seconds to force any remaining ethanol to the bottom.
28. Use a micropipette to carefully remove the remaining ethanol from each tube. Be careful not to disturb the pellet.

In Part III, you will use 2.5 mL of DNA for each PCR. 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.

29. Air dry the pellets for 10 minutes to evaporate remaining ethanol.
30. Add 100µl of dH2O to each tube, and dissolve the DNA pellet by pipetting in and out several times.
31. Store your samples on ice or at -20°C until you are ready to begin Part III.

III. Amplify DNA by PCR

REAGENTS, SUPPLIES, & EQUIPMENT

For each group
Container with cracked or crushed ice
*appropriate primer/loading dye mix (25 µL)**
*DNA from specimen (from part II)**
Micropipettes and tips (1-100 µL)
Permanent marker
2 Ready-To-Go PCR Beads in 0.2- or 0.5-mL PCR tubes

To share
Thermal cycler
**Store on ice*

1. Obtain two PCR tubes containing Ready-To-Go PCR Beads. Label the tubes with your identification and group numbers.

The primer/loading dye mix will turn purple as the PCR bead dissolves.

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.

Your teacher will prepare reactions with forward and reverse primers for a single locus

Plants: rbcL primers (rbcLaF / rbcLa rev)

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

Insects: (LepF1_t1/ LepR1_t1)

Other animals: (LepF1_t1/ VF1_t1/ VF1d_t1/ VF1i_t1/ LepR1_t1/ VR1d_t1/ VR1_t1/ VR1i_t1)

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 tips to add 2 µL of your DNA (from Part I) directly into the appropriate primer/loading dye mix. Ensure that no DNA remains in the tip after pipetting.
4. Store your samples on ice until your class is ready to begin thermal cycling.
5. Place your PCR tubes, along with those of the other students, in a thermal cycler that has been programmed for 35 cycles of the following profile:

   Denaturing step: 94°C 15 seconds

   Annealing step: 54°C 15 seconds

   Extending step: 72°C 30 seconds

   The profile may be linked to a 4°C hold program after the 35 cycles have been completed.
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

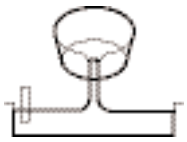
REAGENTS, SUPPLIES, & EQUIPMENT

For each group
2% agarose in 1x TBE (hold at 60°C) (50 mL per gel)
Container with cracked or crushed ice
Gel-casting tray and comb
Gel electrophoresis chamber and power supply
Latex gloves
Masking tape
Microcentrifuge tube rack
2 Microcentrifuge tubes (1.5mL)
Micropipette and tips (1–100 µL)
*pBR322/BstNI marker (20 µL per gel)**
*PCR products from Part II**
SYBR Green DNA stain <!> (6 µL per group)
1x TBE buffer (300 mL per gel)

To share
Digital camera or photodocumentary system
Microwave
UV transilluminator <!> and eye protection
Water bath for agarose solution (60°C)

**Store on ice.*

<!>See Appendix XX for appropriate handling.



Avoid pouring an overly thick gel, which will be more difficult to visualize.

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, 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 open teeth of the comb.
3. Allow the 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 of your PCR products (from part III) to a fresh 1.5mL microcentrifuge tube. Add 2 µL of SYBR Green DNA stain to each tube.

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

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.

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

7. Add 2 µL of SYBR Green DNA stain to 20 µL of pBR/*Bst*NI marker.
8. Orient the gel according to the diagram below, so that 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.
9. Use a micropipette with a fresh tip to load each sample from Step 6 in your assigned wells, according to the following diagram:

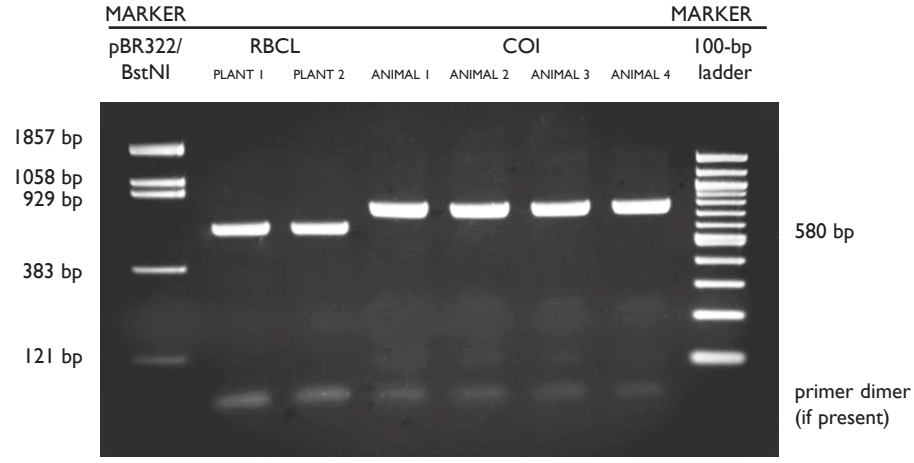
10. Store the remaining 20 µL of your PCR product on ice or at -20°C until you are ready to to submit your samples for sequencing.
11. Run the gel for approximately 30 minutes at 130V. Adequate separation will have occurred when the cresol red dye front has moved at least 50 mm from the wells.
12. View 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 purpose of each of the following steps or reagents used in DNA isolation (Part II(a) of Experimental Methods):
   i. Collecting new leaves or leaf buds.
   ii. Using only a small amount of tissue.
   iii. Grinding tissue with pestle.
   iv. Nuclei lysis buffer or Edward's buffer.
   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 above 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 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 PCR products, do the bands all appear to be the same bp size and inensity?
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 METHODS

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

1. Perform a BLAST search as follows:
   i. Do an Internet search for "ncbi blast."
   ii. Click on 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).
   ii. Under the heading "Basic BLAST," click on "nucleotide blast."
   iii. Enter the primer set you used into the search window. These are the query sequences.
   iv. Omit any non-nucleotide characters from the window because they will not be recognized by the BLAST algorithm.

The following primer set was used in this experiment:

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

Animal *COI* gene

*Lep*F1 5'- ATTCAACCAATCATAAAGATATTGG -3' (forward primer)
*Lep*R1 5'- TAAACTTCTGGATGTCCAAAAAATCA-3'(reverse primer)

VF1F 5'- TCTCAACCAACCACAAAGACATTGG-3' (forward primer)
VF1R 5'- TAGACTTCTGGGTGGCCAAAGAATCA-3' (reverse primer)

v. Under "Choose Search Set," select "NCBI Genomes (chromosome)".

vi. Under "Program Selection," optimize for "Somewhat similar sequences (blastn)."

vii. Click on "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 1 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:

i. First, a *Graphic Summary* illustrates how significant matches, or "hits," align with the query sequence. Why are some alignments longer than others?

ii. This is followed by *Descriptions of sequences producing significant alignments*, a table with links to database reports.

a. 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.

b. Note the scores in the "e" 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.

c. 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?

d. What do the descriptions of significant hits have in common?

iii. 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*). Notice that hits have matches to one or both of the primers:

	Forward Primer	Reverse Primer
*rbc*L	nucleotides 1-26	nucleotides 27-46
Lep or *VF*	nucleotide 1-25	nucleotides 26-53

Which of the hits would be amplified, in vitro, in a PCR reaction using the two primers? Why?

3. Predict the length of the product that the primer set would amplify in a PCR reaction (in vitro).

i. In the *Alignments* section, select a hit that matches both primer sequences.

ii. Which nucleotide positions do the primers match in the subject sequence?

iii. 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.

iv. However, the PCR product includes both ends, so add 1 nucleotide to the result that you obtained in Step 3.iii to determine the exact length of the fragment amplified by the two primers.

v. What value do you get if you calculate the fragment size for other species that have matches to the forward and reverse primer? Why is this so?

4. Determine the type of DNA sequence amplified by the primer set:

i. Click on the accession link (beginning with "*ref*") to open the data sheet for the hit used in Question 3 above.

ii. The data sheet has three parts:

a. 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.

b. The bottom section lists the nucleotide sequence.

c. 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).

iii. Identify the feature(s) located between the nucleotide positions identified by the primers, as determined in 3.ii above.

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

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.

1. Create a *DNA Subway* Project and Upload DNA Sequences

i. Log into *DNA Subway* at www.dnasubway.org. If you do not have an account, you will need to register first to be able save and share your work.

ii. Select "Determine Sequence Relationships" (Blue Line) to begin a project.

iii. Select "*rbc*L" or "COI" from the "Select Project Type" section. (*rbc*L (plant) sequences must be analyzed separately COI (animal) sequences.)

iv. Select "Sequence Source" provides several ways to obtain sequences for barcode analysis:

a. *Upload sequence(s) in ab1* (files ending with .ab1) *or FASTA format*. Click

To select multiple consecutive files, click on the first file you want, then hold down the shift key and then click on the last file in the sequence.

"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".

b. Enter a sequence in FASTA format:

>sequence name

atcgccccttaatattgcctt…..

c. Import a sequence/trace from the DNALC. Click on your tracking number. Select one or more files from the list. Click to "Add" selected files.

d. Select a sample sequence.

v. Provide a title in the *Name Your Project* section.

vi. Write a short description of your project in the *Description* section (optional).

vii. Click on "Continue."

2. View and Build Sequences

The electropherogram view is only available for ab1 files that include a trace file. FASTA files do not include trace information needed to output an electropherogram.

You can increase peak height by clicking the + button for the Y axis.

i. On the *View & Build Sequences* branch line, click on "Sequence Viewer." Click on a sequence name to view an electropherogram with quality scores for each nucleotide.

a. 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.

b. 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.

c. 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%

d. The electropherogram viewer represents each Phred score as a blue bar. A Phred score of 20, represented by the horizontal line, is considered the cut-

off for high-quality sequence. What is the error rate and accuracy associated with a Pred score of 20?

e. 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.

f. 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"?

If you only have a single read – sequence from only a forward or reverse primer – skip to Step 4.

ii. Click on "Sequence Trimmer" to automatically remove Ns from the 5' and 3' ends of selected sequences. Click again to view the trimmed sequences. Why is it important to remove excess N's from the ends of the sequences?

3. Pair Forward and Reverse Reads

If you have two reads with one primer, you can also build a consensus for these reads. Ensure that both sequences are oriented correctly: for a forward primer, F should be displayed for both primers, while for a reverse primer, both sequences should display R.

i. If you have good quality forward and reverse reads for any sample, click on "Pair Builder" to associate a forward read with its corresponding reverse read.

ii. Check the boxes for two sequences you wish to pair, and confirm your selection in the pop-up.

iii. Click on the "F" to the right of the reverse sequence. The entry will change to *R*, indicating that the sequence has been transformed into its reverse complement.

Remember that the DNA molecule is composed of two antiparallel strands, which "read" in opposite orientations. The reverse complement makes the reverse strand sequence equivalent to the forward strand by: 1) reversing the sequence order, so that it reads in the same direction and 2) complementing each nucleotide (A>T, T>A, G>C, C>G) so that the sequence reads on the same strand.

iv. Click on "Save" to save your pair assignments.

v. Click on "Consensus Builder" to align the paired forward and reverse reads. Then, click on a forward-reverse pair to view its consensus sequence. How does the consensus sequence optimize the amount of sequence information available for analysis? Why does this occur?

A consensus sequence is the best agreement between multiple sequences – in this case, between a forward and reverse read. In nature, the forward strand and its reverse complement are a perfect match. However, the sequencing process is not perfect, so there are often differences between forward and reverse reads. When there is a discrepancy at a nucleotide position in two or more reads, the consensus software selects the nucleotide with the highest quality (Pfred) score.)

vi. Positions highlighted in yellow mark differences in nucleotide calls between the forward and reverse reads. Do differences tend to occur in certain areas of the sequence? Why?

vii. 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.

a. Return to *Pair Builder* to check your pairs and reverse complements.

b. Click on the red "x"to redo a pairing, and toggle "F" and "R" settings, as needed.

viii. 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:

a. Examine the distribution of Ns to see if they are mainly confined to one of the two sequences.

b. Examine the electropherograms to see if one of the two sequences is of good quality.

c. If one of the sequences seems of good quality, return to Pair Builder, and click the red x undo the pairing.

d. Continue on to Step 4.

ix. 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:

a. Click on a highlighted mismatch to see the electropherograms and Pfred scores for each read.

Changing the consensus sequence arbitrarily is likely to create a change in the sequence that does not represent the sequence in the organism.

b. Click on 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.

c. 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.

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

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

a. 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).

b. Organism and sequence description or gene name of the hit. Click on 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).

c. 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?

iii. Add BLAST sequence data to your phylogenetic analysis by checking the box(es) above 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

If you have a good idea of the taxonomy of your sample, you may want to select Reference Data from a narrow range of plants or animals including the putative family your sample is from. If you have little idea of the taxonomy of your sample, include a very broad selection of Reference Data.

i. Click on "Upload Data" to include additional data. Either upload data in ab1 or FASTA format or import data from other sources.

ii. Click on "Reference Data" to select data that will let you compare your barcode sequence in an appropriate phylogenetic context.

6. Analyze Sequences: Select and Align

Many unknown species can be rapidly identified by a BLAST search. In this case, a phylogentic 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.

i. Click on "Select Data." Then check boxes to select any or all of the sequences you have uploaded from your own sequencing projects, from BLAST searches, and from reference data sets. Click on Save.

MUSCLE is a multiple sequence alignment program, like CLUSTALW, which aligns two or more sequences in a manner that produces the fewest gaps. Jalview is a Java utility for viewing and editing the alignments produced by Muscle. Jalview also calculates and displays phylogenetic trees.

ii. Click on "MUSCLE" to align your sequences. When the program is finished, click again to view the alignment in Jalview.

a. 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.

b. 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?

c. 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 on "MUSCLE" to recalculate.

iii. Trim unaligned ends of the sequences

a. Identify the leftmost point at which all or most sequences show corresponding nucleotide color bars. (There should be few or no gaps in the vertical column of nucleotides at this point.)

b. Click in the nucleotide coordinate bar directly above this nucleotide in the first sequence. This will activate a red cursor and a pop-up menu.

c. Click on "Remove left" to trim the leftmost sequences to this nucleotide

position.

d. Repeat Steps a-b, and click Remove right to trim the rightmost sequences.

e. You can return to Select Data (Step ii.C. above) to remove any sequence that has large sequence gaps. Why is it important to remove sequence gaps and unaligned ends?

f. Click Submit trimmed alignment

Tree-building algorithms attempt to reconstruct the order in which sequence mutations accumulated as different lineages diverged from a common ancestor. A number of plausible trees can be constructed from any set of sequences, so an algorithm presents what it determines to be the optimal one. The maximum likelihood algorithm evaluates possible trees and determines which is mostly likely to have been produced by the observed data. Because it fits mutations to a tree, the maximum likelihood method produces the most parsimonious tree – one that accounts for the data with the shortest branch lengths.

The tree visualization software may assign a numerical value to each branch, which is proportional to its length.

7. Analyze Sequences: Create a Phylogenetic Tree

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

ii. 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.

iii. Look at your tree.

a. 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.

b. 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, those with longer branches are more distantly related.

c. 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.

d. 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?

e. 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?

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

a. If your sequence is closely related to any of the reference or uploaded sequences, it will share a single node with those species.

b. If your sequence is identical to another sequence, the two will diverge directly from the node *without branches.*

c. 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.

d. To identify the smallest clade that includes your sequence, click on the node that is directly connected to your sequence. The sequences that are

highlighted are the closest relatives of your sequence in the tree.

e. 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.

v. 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?

vi. 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 most order or family clade that contained your sequence. Then repeat Steps 6-7 to select, align, and generate trees from your refined data set.

The neighbor-joining algorithm builds a tree from the bottom up by comparing the evolutionary distance between pairs of DNA sequences. Sequences with best matching sequences are linked as "neighbors" that share common nodes in the tree. Because the branch distances are produced in a pairwise manner, neighbor joining does not optimize branch length and tree parsimony. The chief advantage of neighbor joining – that it is less computational intensive than maximum likelihood – has become less important as the processing power of computers has increased.

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(a) of Experimental Methods):**

   i. **Collecting new leaves or leaf buds.**

   New leaves and buds have about the same number of cells mature leaves, so they contain about the same amount of DNA in a smaller volume of tissue. The cell walls are thinner than in mature plant materials, making them easier to break during mechanical grinding used in this protocol

   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) and 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 which are more easily lysed to release DNA.

   iv. **Nuclei lysis buffer or Edward's buffer.**

   Nuclei lysis buffer and Edward's buffer contain a detergent that dissolves lipids in the cell membrane and membrane bound organelles (nucleus, mitochondria, chloroplast, etc.). Edward's buffer also includes salts that aid in precipitating the DNA.

   v. **Heating or boiling**

   Heating to 65°C (with the nuclei lysis buffer) or boiling (with Edward's buffer) helps to break down the cell and nuclear membranes, and also denatures enzymes that can degrade the purified DNA.

II. Interpret Your Gel and Think About the Experiment

3. **Looking across the gel at PCR products, do the bands all appear to be the same bp size and intensity? (REPLICATE THIS QUESTION ON PAGE 15.)**

   Each barcode primer set is optimized to amplify the same region across a range of species. Although the size of products can vary, the majority of PCR products will be of similar basepair size and, therefore, will migrate to the same position on the gel. (Of course, barcodes amplified using different primer sets – for example, *rbc*L vs. *COI* – will produce differently sized products that will migrate to different positions on the gel.) However, the intensity of staining (thickness of bands) will vary between reactions. This is related to the mass of product produced by the PCR reaction and the volume of reaction loaded that is successfully loaded in a 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) 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.i. **What do you notice about the lengths (and colors) of the matches (bars) as you look from the top to the bottom?**

The lengths and colors give you information about 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. There may also be a color difference between the forward and reverse primer matches. Typically, most of the significant alignments will have complete matches to the forward and reverse primers.

2.ii. **What is the E-value of the most significant hit and what does it mean?**

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! More precisely, 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.

2.iii. **Some of the matching subjects (accession numbers) may only match the forward primer, and where there is a match to the forward and reverse primer, the match may be poorer (as indicated by the color coding). Looking at the sequence for the reverse primer, why do you think this is?**

The reverse primer for *rbc*L is shorter in sequence so any BLAST match is necessarily lower in significance. The reverse primer also contains an ambiguous nucleotide "R" which means (A or G).

3. I. **What type of sequences are your BLAST hits?**

For the *rbc*L primers, the source of the sequences should all be chloroplast genomes. For *COI* primers, the hits should all be mitochondrial genomes.

4. ii. **To which positions do the primers match in the subject sequence?**

The answers will vary for each hit and primer set. For *Phoenix dactylifera*

(NC_013991.2), the *rbc*L primers match 56930-56955 and 57509-57528, respectively. For *Spathius agrili* (NC_014278.1), the *Lep* primers match 2035-2060 and 2718-2743, respectively. For *Rattus tunneyi* (NC_014861.1), the *VF1* primers match 5339-5363 and 6022-6048 respectively.

4.iii. **Subtract the lowest nucleotide position in the subject sequence from the highest nucleotide position in the subject sequence. What is the difference between the coordinates? If you calculate this value for one or two other species that have matches to the forward and reverse primer, do you get the same number?**

For the *rbc*L primers, using *Phoenix dactylifera* (NC_013991.2) as an example gives 56955-57509 = 554 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 *Lep* primers, using *Spathius agrili* (NC_014278.1) as an example yields 2743-2035 = 708 nucleotides. The range of possible lengths for these primers is between 670 and 750 nucleotides.

For the *VF1* primers, *Rattus tunneyi* (NC_014861.1) as an example gives 6048-5339 = 709 nucleotides. These primers usually have hits ranging from 700 to 720 nucleotides.

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

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

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

2.i.f. **What do you notice about the electropherogram peaks and quality scores at nucleotide positions labeled "N"?**

At "N" positions, peaks for different channels of similar amplitude often overlap, or no prominent peak rises above low-amplitude background ("noise"). Quality scores are less than 2.

2.ii. **Why is it important to remove excess N's 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.v. **How does the consensus sequence optimize the amount of sequence information available for analysis? Why does this occur?**

Consensus sequence extends the ends, producing a longer contiguous sequence. The 5' sequence immediately following the primer has many sequencing errors and is trimmed. So, the reverse read extends this low-quality region at the 5' end of the sequence, and the forward read extends the low-quality region at the 3' end. Also, the sequence quality often drops as the distance increases from the primer. With some sequences, this will result in regions where only the forward or reverse sequence will be of high quality. By selecting the best sequence for these regions, the overall quality of the consensus will be better than either for-

ward or reverse sequences.

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

Differences clustering at 5' and 3' ends are typically a combination of low-quality sequence near the primers and loss of signal as the distance increases from the primer, both of which cause the ends of the consensus to have the poorest sequence quality.

4.2.c. **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 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.ii.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.iii.e. **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.iii.d. **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 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.

7.iii.e. **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.s

7.v. **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.