

From Bugs to Barcodes: Using Molecular Tools to Study Biodiversity

Madeline Butler, Heather Henter and Stephanie Mel

University of California San Diego, Division of Biological Sciences, 9500 Gilman Dr., La Jolla CA 92093 USA

(mabutler@ucsd.edu; hhenter@ucsd.edu; smel@ucsd.edu)

Biodiversity refers to the variety of species within an ecosystem. Most species on earth have yet to be identified, and there is an ongoing international effort to build a complete species inventory. Many students are aware of the importance of biodiversity but few realize that they can play an important role in addressing this biodiversity knowledge gap. DNA barcoding is a technique that uses a specific region of DNA as a genetic marker to identify species. In this module, students learn how to document biodiversity using ecological, molecular, and bioinformatics tools while generating novel data.

Keywords: Biodiversity, barcoding, PCR, DNA sequencing, bioinformatics

Introduction

Biodiversity, usually defined as the number of species in a specific ecosystem or area, is important for numerous reasons. A diversity of organisms support ecosystem services such as purification of air, climate control, water purification, food production, pollination, and erosion prevention. Many people feel that biodiversity is important for aesthetic, ethical, and cultural reasons as well. But biodiversity is at risk. Habitat destruction is probably the most serious threat, although over-exploitation of natural resources and invasive species play a role, and climate change will become increasingly important in the future.

However, we cannot assess the impact of any of these threats if we do not know what is being threatened. Current estimates of the total number of eukaryote species on Earth range from five to ten million (May, 2010). Of this, less than two million species have been named or described, and this lack of information is not distributed evenly among taxa despite 250 years of modern taxonomy. Although groups such as mammals and birds are fairly well known, it is estimated that 70% of arthropod species have yet to be discovered (Hamilton et al., 2010). This is particularly troubling at a time when human activities are impacting virtually every organism on the planet. This knowledge gap is a huge obstacle for conservation efforts. It is critical that we develop a better understanding of what organisms exist if we want to conserve species, or even know the effect of our conservation strategies. With DNA barcoding, students can help in this effort.

Species are usually identified by their morphology. It is possible for students or other non-experts to identify large organisms such as birds or mammals this way, but identifying smaller organisms such as invertebrates can be very difficult. Morphological differences are almost always so subtle it takes an expert to distinguish species. Also, as is the case with cryptic species, there may not be any morphological differences even when the genetic evidence suggests that the organisms do not interbreed, and thus by definition are different species.

Molecular taxonomists proposed using a specific DNA sequence called the barcode as an identifier for species (Herbert et al., 2003). There has been debate about reliability of the barcode sequence to detect taxonomic subtleties, but many taxonomists have embraced the use of barcode sequences as an additional tool. The Consortium for the Barcode of Life (CBOL) is an international collaboration of experts in genomics, taxonomy, and computer science whose mission is to create a reference library of DNA barcodes in the form of the Barcode of Life Database (BOLD, <http://www.barcodeoflife.org/>).

Students at any academic institution can become involved in barcoding projects and even contribute novel sequences to the BOLD database. Barcoding is presently being used in a variety of educational settings as a means of involving students in discovery-based science (Santschiet al., 2013; and <http://www.urbanbarcodeproject.org/>). Barcoding can also be used to integrate concepts and provide hands-

one exposure to techniques in a variety of different disciplines such as ecology, taxonomy, genetics, molecular biology, and bioinformatics.

As part of our effort to bring authentic research into large undergraduate biology labs, we have initiated a DNA barcoding project in which students are documenting biodiversity at the UC San Diego Scripps Coastal Reserve. We have used barcoding both to discover the diversity of species in a particular habitat as well as to test specific hypotheses. For example, one quarter we documented the vegetation-inhabiting spiders at our reserve and another quarter we looked at the intraspecies diversity in the honeybee *Apis mellifera*, which has a number of subspecies. Our hypothesis-testing projects have explored whether polychaete worms from different intertidal zones of the sandy beach are different species and whether flower-inhabiting thrip species specialize on different species of plant hosts. Other groups have used barcoding for similar studies, both basic and applied. For example, students have done barcoding to detect whether fish being sold in stores are the actual advertised species (Stockle et al., 2010).

In order to do barcoding, students collect specimens, extract DNA, then amplify the DNA barcoding region using Polymerase Chain Reaction (PCR). After running a gel to verify that they have a PCR product of the correct size, students then purify the PCR product and send it for Sanger sequencing. The DNA sequences are then analyzed using several free bioinformatics programs. The methods are all straightforward and require only basic specimen collection and molecular biology lab equipment. We typically spread the experiments out over four lab periods of about 3 hours each, and the schedule can be adjusted to fit lab periods of shorter or longer times. The experimental protocols in this paper have been designed for use with insects but they can be used with other invertebrates as well. However, the DNA extraction methods and primers should be tested with the type of animals to be studied in your class before actual implementation of the module.

Summary of Steps in Barcoding Experiment

1. Collect specimens (1 hour or less)
2. Extract DNA from insect legs (2.5 hours to overnight)
3. Set up PCR reactions (30 minutes)
4. Run gel to verify that PCR worked and purify PCR product (1.5 to 2 hours)
5. Send purified PCR product for sequencing (done by outside company)
6. Analyze sequences (2 to 4 hours or more)

Student Outline

In order to barcode an organism, DNA is first extracted from an organism and then the barcode sequence is amplified using PCR. Because we want to simply amplify and sequence the DNA without having to clone it, it is important to use a haploid gene. Mitochondrial DNA is only inherited from the mother, and thus all the genes on the mitochondrial DNA are haploid. Mitochondrial genes also have a low level of intraspecies diversity and a high level of interspecies diversity, which makes it useful for differentiating species based on DNA sequence differences. Also, there are many copies of DNA per mitochondria, and there are many mitochondria per cell so the copy number of mitochondrial genes is higher than nuclear genes.

We will be amplifying DNA using part of the cytochrome c oxidase (CO1) gene located on mitochondrial DNA. This gene has been accepted by scientists as the standard gene to be used for all animal barcoding studies (Herbert et al., 2003). For the barcode PCR, we use what are known as “universal primers” which are designed to recognize conserved areas in the CO1 gene in many invertebrate species. Because the primer sequences will not be an exact match to the CO1 target sequence in all invertebrate species, the PCR reaction is performed at a low annealing temperature. This should allow primers that are not an exact match to still anneal well enough to form a stable duplex for the PCR reaction.

After running the PCR, some of the PCR product is run on a gel to make sure it is the expected size which is about 660-680 base pairs. The remainder of the PCR sample is then cleaned up to remove the free nucleotides, primers, and enzyme and it is then sent for Sanger sequencing. The sequencing results are analyzed using several free bioinformatics programs and databases.

1. Collecting Insects in the Field

DNA barcoding can be used to document species in a particular area. We will be working with insects because they are they are highly diverse and easy to collect, and it is estimated that 70% of arthropod species have yet to be discovered by scientists (Hamilton et al., 2010). Insects and other arthropods are virtually everywhere and can be collected a number of ways. Your instructor will demonstrate the equipment you can use, but here are a few ideas to get you started.

Places to Find Insects

Flowers– Pollinators (bees, etc.) are very common at flowers. Tap a flower over a white tray, white paper plate, or white pad of paper to find thrips, spiders, and other species that lurk therein.

Leaves– Look on the underside of leaves for plant sucking insects. Look for evidence of leaf chewing, which suggests that herbivorous insects might be active. Galls are oddly shaped plant growths caused by the immature insect developing inside. Break open the gall to see if the insect is still there.

Underneath logs – Turn over any object that creates damp, protected conditions - stones, logs, old lumber, or trash. You will be sure to find earwigs and other moisture-loving insects. Ants, termites, roaches, beetles, and bristletails are common.

Lights– Lots of insects are attracted to lights at night, especially moths and lacewings. Look around your porch light at night.

Water– Look under stones in running streams for immature mayflies, stoneflies, and the cases of caddisflies. Water striders are common walking on water, and look in the shallows along the edges of ponds for various aquatic beetles and immature dragonflies, midges, and mosquitoes.

Basements– Look in old books and newspapers for silverfish and booklice which are primarily feeding on the mold that grows in humid conditions.

Traps– Put out fruit, rotting or otherwise, as baits to attract insects. Try meat, cookies, or a soda. Create habitat by putting out pieces of wood.

Pitfall traps – These are traps used to catch ground dwelling insects. Sink any sort of plastic jar or vial (such as a 50 ml Falcon tube) into the dirt so that the top is level with the ground. Fill with a preservative, such as 70% ethanol and leave overnight. Ants, bristletails, beetles, and others will be trapped.

Bee bowls – Fill any sort of small disposable cup or bowl with soapy water (one squirt per gallon of water, Dawn dish soap is usually used). The soap breaks the surface tension of the water so that the insects sink. Paint the cups bright yellow or blue, or try just white and place the cups in the open. The day must be sunny and warm for bees to be active, and when conditions are right you will get bees within seconds. However, some species of bees are attracted to bee bowls, others are not.

Labeling and Preserving Insects

For the best DNA preservation, store invertebrates in 95% ethanol in cool conditions (a refrigerator or -20°C freezer). The most important aspect of collecting insects is labeling – a specimen is worthless if it is not labeled with the location, date, and collector. Each specimen or each vial needs an internal label. Use a small piece of paper (about 1 cm x 2 cm) inside the vial. Do not use tape (it falls off). For storage in alcohol, write the label in pencil rather than pen as pencil will be more permanent. If possible, also record the latitude and longitude. See Fig. 1 for an example.

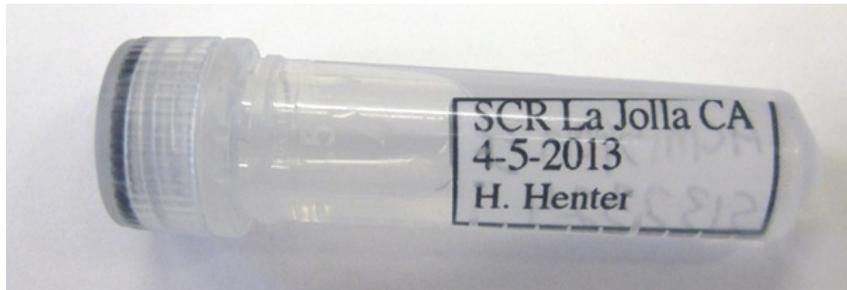


Figure 1. Example of a collection vial with internal label.

Photographs

To create a record of the specimen, take a photo. You can use any digital camera – even the camera on a phone. It is a good idea to photograph the specimen before preserving it in ethanol.

2. DNA Isolation Using Qiagen DNeasy Blood and Tissue Kit

You now need to isolate DNA from your insect. If you collected a large enough insect, you will use one or two legs. If you collected a small insect, you will use the entire specimen.

Protocol

- a. For large insects, carefully cut off one of the back legs with scissors and save the rest of the specimen. You should have about 3 to 4 mm of leg so if the specimen legs are very small, use more than one. Try to cut as close to the body as possible. Do not simply pull off the leg.
- b. Place the leg (or the entire specimen if your bug was very small) in a blue microfuge tube. Place the material on the side of the tube and grind it with the blue pestle. The idea is to break the leg or small insect into smaller pieces. If this too hard, pull the tissue out of the tube and use the scissors to cut it up into smaller pieces. Then try grinding those.
- c. Add 180 µl of ATL buffer to the tube and use the pestle to further grind up the insect tissue in the ATL buffer.
- d. Add 20 µl of proteinase K to the microfuge tube. Let incubate for at least 2 hours at 56°C. (Overnight incubation is good if you want to skip this grinding step or if your lab is not long enough to do the 2 hour incubation and subsequent extraction in the same period.)
- e. Vortex for 20 seconds, and then add 200 µl of AL buffer to the tube. Vortex again.
- f. Add 200 µl of ethanol to the tube and mix again by vortexing.
- g. Place a column in a collection tube. Now pipette all the liquid from your ground up insect tissue onto this column. Centrifuge at 8,000 rpm for 1 minute. After the spin, discard both the flow-through (this is the liquid that you spun through the column) and the collection tube.
- h. Place the column in a new collection tube. Add 500 µl of AW1 buffer, and then centrifuge at 8,000 rpm for 1 minute.
- i. Again, discard the flow-through and collection tube and place the column in a new collection tube. Add 500 µl AW2 buffer and centrifuge for 3 minutes at maximum speed. Once again, discard the column and the flow-through.
- j. Now place the column in a microfuge tube and add 50 µl of AE buffer. Let sit for 1 minute, then centrifuge at 8,000 rpm for 1 minute. **SAVE THE FLOW-THROUGH – THIS HAS THE DNA.** You will use this DNA to set up a PCR reaction.

NOTE: If a Nanodrop is available, students can measure how much DNA they got from their specimen by reading the A260 and A280.

3. Setting Up the PCR Reaction

The primers you will be using today are designed for use with invertebrate samples – the sequences are in Table 1.

Table 1. Cytochrome c oxidase invertebrate barcoding primers.

Primer	Name	Sequence
Forward	LCO1490	5'-GGTCAACAAATCATAAAGATATTGG-3'
Reverse	HCO2198	5'-TAAACTTCAGGGTGACCAAAAAATCA-3'

PCR Master mix

Make up the master mix for a single PCR reaction by adding the components in the respective volumes listed in Table 2 into a PCR tube. The total reaction volume will be 50 μ l with the insect DNA added.

Table 2. Components of and respective volumes for a single PCR reaction.

Volume (μ l)	Component	Stock Concentration (μ M)
25	GoTaq green	na
2.5	forward primer	10
2.5	reverse primer	10
17.5	sterile water	na

PCR Reaction

Add 2.5 μ l of the purified insect DNA to the PCR master mix. Ensure all DNA is added to the master mix. Mix by inverting the PCR tube a few times. Make sure you label the top of the PCR tube with your group number and then centrifuge briefly to spin reagents down.

The PCR conditions are as follows. An initial denaturation step at 94°C for 3 minutes. Then 35 cycles of 95°C for 45 seconds, 42°C for 45 seconds, and 72°C for 60 seconds. The PCR reaction finishes with a final extension step at 72°C for 7 minutes.

4. Run Gel to Verify that PCR Worked and Clean-Up PCR Product

You now need to run a gel to verify that your PCR reaction worked and that you have a product of approximately 660-680 base pairs. You will run some of your PCR product on the gel and then clean up the rest if your PCR worked.

Run agarose gel

- Each group only needs to run 5 μ l of their sample on a gel so to save agarose, four groups will share one gel. Also, please note that there is no need to add any sample dye into the PCR samples because the Go Taq Green solution contains a dye.
- Put 0.75 g of agarose in a flask and add 50 μ l of TAE buffer. Microwave for about 1 minute to melt the agarose. Once the agarose has cooled a bit, add 5 μ l SYBR Safe.
- Set up your gel rig with a comb with ten teeth or more. Pour the agarose into the gel rig and let solidify.
- Once the gel has hardened, remove the dams and comb and flood the gel with TAE buffer. Please pay close attention to how the gel should be loaded – it is very important that you do not mix-up your sample with the other groups on the gel! Assuming your comb has ten teeth, load 10 μ l of the ladder into one lane. Then have four groups load 5 μ l of their PCR product into a lane, skipping lanes between samples. After loading the gel, run at 150 mV until the green dye front is about halfway down the gel. At that point, turn off power.

- e. Take a picture of your gel and determine if you got a PCR product of the correct size. If you did, clean-up the remaining 45 μ l of your PCR sample.

Clean-Up PCR Product

The PCR product you hopefully generated will still contain salts, primers, and enzyme that all need to be removed before you can send it out for sequencing. You will use a PCR purification kit to purify your PCR product.

GeneJet kit protocol

- a. Add 45 μ l of binding buffer to the 45 μ l of the PCR reaction you have left after running the gel.
- b. Apply the sample to a column in a collection tube and spin for 1 minute at max speed.
- c. Discard the flow-through; apply 700 μ l wash buffer to the column and centrifuge for 1 minute at 10,000 rpm.
- d. Discard flow-through – centrifuge empty column for 1 minute until it is dry.
- e. Place column in new, labeled 1.5 ml tube. To elute DNA, apply 20 μ l elution buffer to center of column, let sit 1 minute, and spin for 1 minute.

If available, use a Nanodrop to determine the concentration of your purified DNA. Most companies require between 5 and 35 ng/ μ l of PCR product for sequencing.

5. Sequencing the PCR Product

In order to sequence your PCR product, you need a primer from which the DNA polymerase can extend. Since you know your PCR product has incorporated the forward and reverse primers, we can use the same primers to start the sequencing reaction. Thus, along with your cleaned up PCR product, you will also send some of your forward and/or reverse primer in separate tubes at a concentration of 5 μ M to the facility that will do the sequencing. Make sure you label the top of your tube clearly with your specimen number.

6. Bioinformatics Analysis

Part 1: Assessing the Quality of the Sequence and Doing a BLAST

(Note: this section would be adapted to use with the sequences generated by your class.)

- a. The first thing you must do is look at your sequence chromatogram and determine if it is good enough to use in the subsequent analyses. Although most of the time we get PCR product, it may or may not have sequenced well. You will first look at some examples of good and bad sequencing runs, and then analyze your own sequences.
- b. First find the “good_sequence_Apis.AB1” file containing a chromatogram of a Sanger sequencing reaction. (All of the files necessary for doing the bioinformatics exercises can be found in Dropbox using the following link <https://>

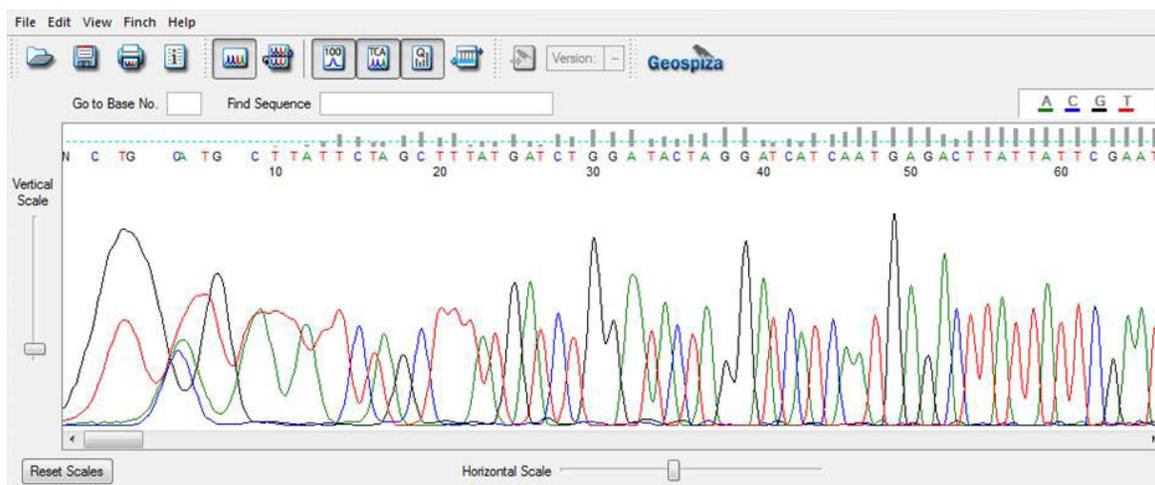


Figure 2. Example of a chromatogram from a good sequencing reaction.

www.dropbox.com/sh/1unf2ozfmbnu5yv/TWM4A2yCfd.) Open the chromatogram in the program Finch TV, which you must first download onto your computer (<http://www.geospiza.com/Products/finchtv.shtml>). The chromatogram should look like the one below in Finch TV.

The peaks in the chromatogram represent the actual sequence of the PCR product (for a good animation of Sanger sequencing, see <http://www.dnalc.org/resources/animations/cycseq.html>). Note that there are four colors, each representing a different base. Also note how the first 25 peaks or so do not look very sharp, but from about peak 25 on, the peaks are well resolved and there is no background. This is a good sequencing reaction. You can also use the gray bars above each base to tell how good the sequence is at that particular point. You can see there is a horizontal green dotted line and then perpendicular gray bars above each base. The higher the bar, the more certain the computer program was about “calling” or identifying that base. For a good sequence, the height of all the grey bars after the first 20 or so should be above the dotted line.

Now open the “Bad_sequence.AB1” file and look at the chromatogram (Fig. 3). Note how the peaks all overlap and very few of the gray bars are above the dotted green line.

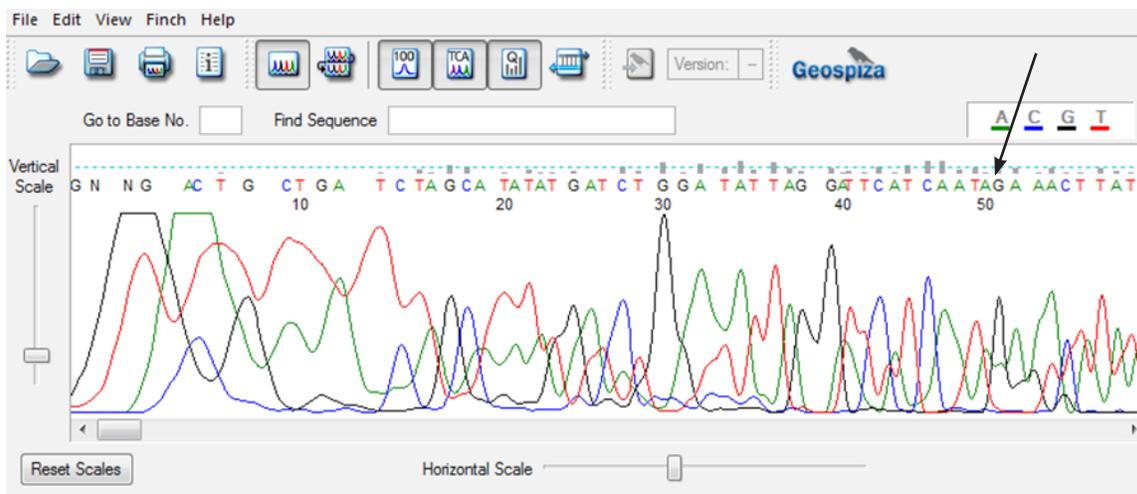


Figure 3. Example of a chromatogram from a bad sequencing reaction. The above sequence would not be usable for the subsequent bioinformatics analyses.

- c. As mentioned above, even in a good sequencing run, the first 25 to 50 base-calls are unreliable and you need to delete that sequence from your analysis. First, open the “Good_sequence_Apis” file again in Finch TV. In order to make sure that everyone’s sequence is trimmed the same, please find the sequence GGATC around position 40 and highlight all the sequence to the left of the C – do not include the C. This is illustrated in Fig. 4. Then select Delete under the Edit menu. The actual peaks will not disappear but all the letters above the peaks will.

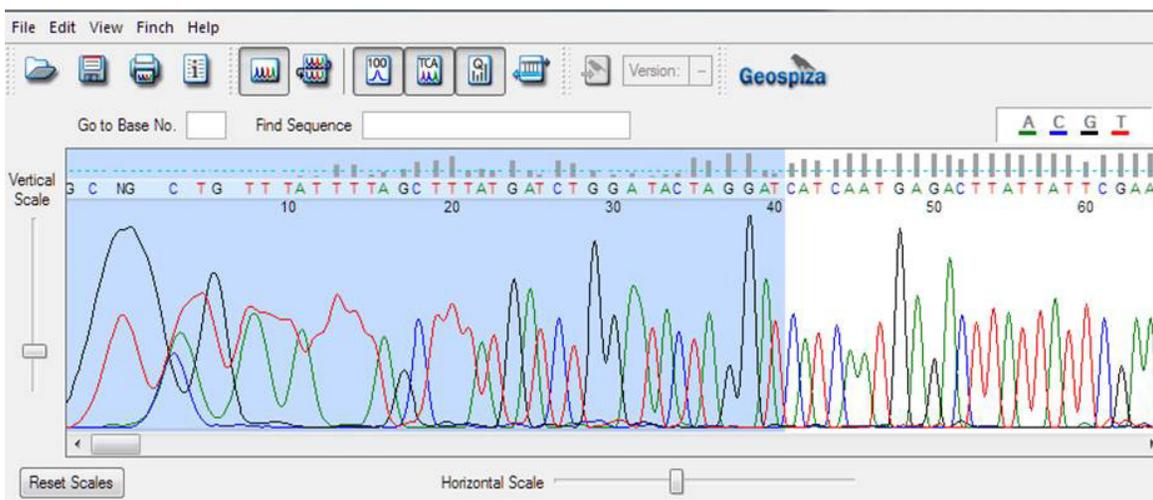


Figure 4. Trimming the beginning of the chromatogram.

Now do the same for the end of the sequence. Look for the sequence TGATTTTT and highlight the sequence to the right of the T (do not include the T) as shown in Fig. 5. Delete this sequence.

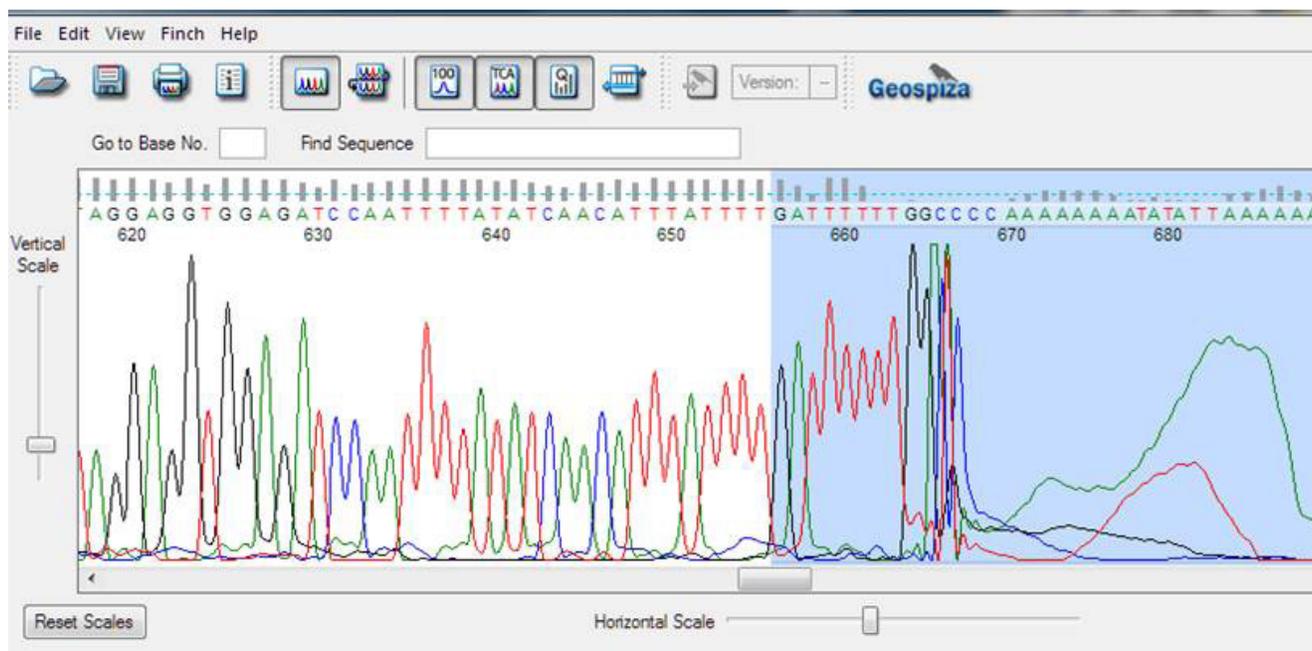


Figure 5. Trimming the end of the chromatogram.

- d. Once the sequence has been cleaned up by trimming the ends, you are going to export the sequence. Go to File on the toolbar, and then select “Export – DNA sequence FASTA” in Finch TV. Save that file to your desktop, but also keep the chromatogram open. Then open the file, using Notepad, Textedit, or Word. This is the text version of the chromatogram file and represents the sequence of the PCR product. This is known as a FASTA file – note how the sequence is preceded by a name and a “>”. Many bioinformatics programs require this type of file format. The exported sequence should look something like that shown in Fig. 6.

```
>Good_sequence_Apis.AB1
GATCTGGATACTAGGATCATCAATGAGACTTATTATTCGAATAGAATTAA
GATCCCCAGGATCATGAATTAATAATGATCAAATTTATAATACAATTGTT
ACTAGCCATGCATTCCTAATAATTTTTTTTATAGTTATACCATTTTTAAT
TGGAGGATTTGGAAATTGGCTTATTCCTTTAATACTAGGATCACCTGATA
TAGCATTTCCCCGAATAAATAATATTAGATTTTGATTACTTCCTCCCTCA
TTATTTACACTTTTATTAAGAAATTTATTTTACCCAAGACCAGGAACCTGG
ATGAACAGTATATCCACCATTATCAGCATATTTATATCATTCCTCACCTT
CAGTAGATTTTGCAATTTTTCTCTTCATATATCAGGAATTCCTCAATT
ATAGGATCATTAACTTAATAGTTACAATTATAATAAAAAAATTTTTTC
TATAAATTATGACCAAATTCATTATTTCCATGATCAGTTTTTATTACAG
CAATTTTATTAATTATATCACTACCTGTATTAGCTGGAGCAATTAATA
CTATTATTTGATCGAAATTTAATACATCATTTTTTCGATCCTATAGGAGG
TGGAGATCCAATTTTATATCAACATTTATTTT
```

Figure 6. Sample of a FASTA DNA sequence file.

- e. Next you are going to see if there are any similar sequences to yours in the GenBank database. We will do this by using the NCBI BLAST tool.

Go to NCBI BLAST (<http://BLAST.ncbi.nlm.nih.gov/BLAST.cgi>) and select “nucleotide BLAST”. Then under “Chose Search Set”, use the pull down menu to select “nucleotide collection” (note that the default is for Human sequence). Paste your FASTA sequence into the entry box, and then click on the BLAST button.

- f. When you do the BLAST, you will get a list of entries in GenBank that are closest in match to your sequence, with the most similar match at the top of the list. (Note that GenBank is constantly updated and the results may differ from what is shown in Fig. 7). Scroll through the list and look at the names – often there are several entries for the same Genus and species. Note that for this exercise, we used a barcode sequence from the honeybee, *Apis mellifera*, which has been well studied so there are many entries in GenBank for this species.

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
HQ978595.1	Apis mellifera voucher BIOUG<CAN>:TDWG-0140 cytochrome	1065	1065	99%	0.0	99%
FJ582089.1	Apis mellifera voucher shefft-84 cytochrome oxidase subunit	1065	1065	99%	0.0	99%
M23409.1	Apis mellifera tRNA-Trp gene, complete sequence; cytochrom	1063	1063	100%	0.0	98%
L06178.1	Apis mellifera ligustica complete mitochondrial genome	1063	1063	100%	0.0	98%
JQ350734.1	Apis mellifera isolate KOR-H2 cytochrome oxidase subunit I (C	1061	1061	99%	0.0	98%
JF700128.1	Apis mellifera voucher AS37MT01 cytochrome oxidase subunit	1061	1061	99%	0.0	98%
FJ582092.1	Apis mellifera voucher shefft-86 cytochrome oxidase subunit	1061	1061	99%	0.0	98%
FJ582091.1	Apis mellifera voucher shefft-85 cytochrome oxidase subunit	1061	1061	99%	0.0	98%
FJ582090.1	Apis mellifera voucher shefft-83 cytochrome oxidase subunit	1056	1056	99%	0.0	98%
HQ978594.1	Apis mellifera voucher BIOUG<CAN>:TDWG-0139 cytochrome	1023	1023	95%	0.0	99%
FJ582088.1	Apis mellifera voucher shefft-87 cytochrome oxidase subunit	1020	1020	95%	0.0	99%

Figure 7. List of BLAST matches.

Now scroll down below the list to see the actual alignment of your sequence to the top match in NCBI – it should look something like that shown in Fig. 8.

Apis mellifera voucher BIOUG:TDWG-0140 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial
 Sequence ID: [gb|HQ978595.1](#) Length: 658 Number of Matches: 1
[▶ See 3 more title\(s\)](#)

Range 1: 45 to 658 [GenBank](#) [Graphics](#) [▼ Next Match](#) [▲ Previous Match](#)

Score	Expect	Identities	Gaps	Strand
1072 bits(1188)	0.0	606/614(99%)	0/614(0%)	Plus/Plus
Query 1	CATCAATGAGACTTATTATT	CGAATAGAATTAAAGATCC	CCAGGATCAIGAATTAATA	ATG 60
Sbjct 45	CATCAATGAGACTTATTATT	CGAATAGAATTAAAGATCC	CCAGGATCAIGAATTAACA	ATG 104
Query 61	ATCAAATTTATAATAACA	ATTGTTACTAGCCATGCAT	TCCCTAATAAATTTTTAT	AGTTA 120
Sbjct 105	ATCAAATTTATAATAACA	ATTGTTACTAGTCATGCAT	TCCCTAATAAATTTTTAT	AGTTA 164
Query 121	TACCAATTTTAAATGGAG	GATTTGGAAATGGCTTAT	TCTTTAATACTAGGATCAC	CTG 180
Sbjct 165	TACCAATTTTAAATGGAG	GATTTGGAAATGGCTTAT	TCTTTAATACTAGGATCAC	CTG 224
Query 181	ATATAGCATTCCCCGA	AATAAATAATATTAGAT	TTTGATTACTTCCCTCC	CTCATTATTTA 240
Sbjct 225	ATATAGCATTCCCCGA	AATAAATAATATTAGAT	TTTGATTACTTCCCTCC	CTCATTATTTA 284
Query 241	CACITTTTAAAGAAAT	TATTTTACCCAGACCAG	GAACCTGGATGAACAG	TATATCCAC 300
Sbjct 285	TACTTTTAAAGAAAT	TATTTTATCCAGACCAG	GAACCTGGATGAACAG	TATATCCAC 344
Query 301	CATTATCAGCAATTTAT	ATCATTCTTCACCTTCAG	TAGATTTTGCATTTTT	CTCTTC 360
Sbjct 345	CATTATCAGCAATTTAT	ATCATTCTTCACCTTCAG	TAGATTTTGCATTTTT	CTCTTC 404

Figure 8. Sequence alignment from BLAST.

The Query sequence is the sequence you submitted to GenBank. The Subject sequence is the sequence of the closest match to your sequence in GenBank. Look first at the “Identities” value in Fig. 8 – this tells you how similar your sequence is to the sequence in GenBank. In this case, the sequence was 99% identical to an entry for *Apis mellifera*. In the alignment shown in Fig. 8, the numbers “604/614” tell you that 604 out of 614 bases were the same in the submitted sequence and the top match found in GenBank. The value in parentheses (99%) tells you how similar this is on a percentage basis. The Expect value (e-value) of the hit is the match you might expect by chance, given the size of the database. A smaller e-value indicates a more meaningful match.

- g. For now, we are most interested in the Identities value. Although there is no hard and fast number, assume for this exercise that any match that is 97% or above is probably the correct match for your specimen – that is, assume your specimen is the same as the species identified in GenBank. If the closest match to your sequence is less than 97% to any entry in GenBank, the barcode sequence for your specimen is probably not in GenBank.

To illustrate this, open the “Good_sequence_Spider.AB1” file. Go through the directions above, trimming the sequence and saving the FASTA file, and then doing a BLAST. In this case, the Identities value for the top match is 92%. Therefore, there is no match at the species level for this CO1 sequence in GenBank. With a 92% match, though, we could reasonably guess that it is in the same genus.

Part 2. Using ClustalW to Align Sequences and Look for Polymorphisms

Barcoding sequences can be used to illustrate the concept of DNA polymorphisms. It is necessary to sequence the CO1 gene in several specimens to do this type of analysis.

- Go to the file called “Apis file for ClustalW”. This file contains the barcode sequences for four different *Apis mellifera* specimens. They were generated by trimming four chromatograms just as you did above. Copy all the files.
- Now go to <http://www.genome.jp/tools/clustalw/>. This brings you to the ClustalW alignment tool. ClustalW can be used to align several sequences in order to compare the sequences to each other.
- Paste the Apis sequences into the entry box. Then make sure to click DNA (protein is the default), and then click on “Execute multiple alignment”.
- You should get something that looks like Fig. 9, only longer and with four entries instead of just two:

Note that wherever the sequences match exactly, an asterisk appears at the bottom of the alignment at that position in the sequence. If there is a space, the sequences differ at that point.

```

Bee81      GATCTTGTATATTATTCTAGCTTTATGATCTGGAATACTAGGATCATCAATGAGACTTAT 60
Bee82      GATCTTGTATATTATTCTAGCTTTATGATCTGGAATACTAGGATCATCAATGAGACTTAT 60
*****

Bee81      TATTCGAATAGAATTAAGATCCCCAGGATCATGAATTAATAATGATCAAATTTATAATAC 120
Bee82      TATTCGAATAGAATTAAGATCCCCAGGATCATGAATTAATAATGATCAAATTTATAATAC 120
*****

Bee81      AATTGTTACTAGCCATGCATTTCCTAATAATTTTTTTTATAGTTATAACCATTTTTAATTGG 180
Bee82      AATTGTTACTAGCCATGCATTTCCTAATAATTTTTTTTATAGTTATAACCATTTTTAATTGG 180
*****

```

Figure 9. Sample results of multiple sequences aligned using ClustalW.

- How many places in the *Apis* barcode sequences are there differences among the sequences? In this example, all the polymorphisms are SNPs, or single nucleotide polymorphisms. SNPs can be either transitions or transversions (Transitions are interchanges of two-ring purines (A to G) or of one-ring pyrimidines (C to T); they therefore involve bases of similar shape. Transversions are interchanges of purine for pyrimidine bases, which therefore involve exchange of one-ring and two-ring structures.) The most common mutations are reportedly C-T transitions – do these data support that?
- ClustalW can also be used to align sequences from different species and build evolutionary trees. To try this, open the “Spider file for ClustalW” and copy and paste the sequences into ClustalW. These are sequences from different species of spiders, and you can see that there are many sequence differences among them. One of the reasons the CO1 gene is used is because the level of intraspecies sequence diversity is low, and the interspecies diversity is high so it is easier to define species boundaries. To use ClustalW to draw evolutionary trees, go to the top of the page, and select “Rooted tree with branch length”.

Part 3. Using ClustalW to Find Synonymous versus Non-synonymous SNPs

Since the sequence you amplified from the CO1 gene codes for a protein, you can determine if any of the SNPs that you observed in doing the DNA alignment in Part 2 changes the protein by doing an alignment of the translated amino acid sequences. Within DNA sequences that code for proteins, a synonymous SNP is one that does not lead to a change in the amino acid and a non-synonymous SNP is one that does lead to a change in the amino acid.

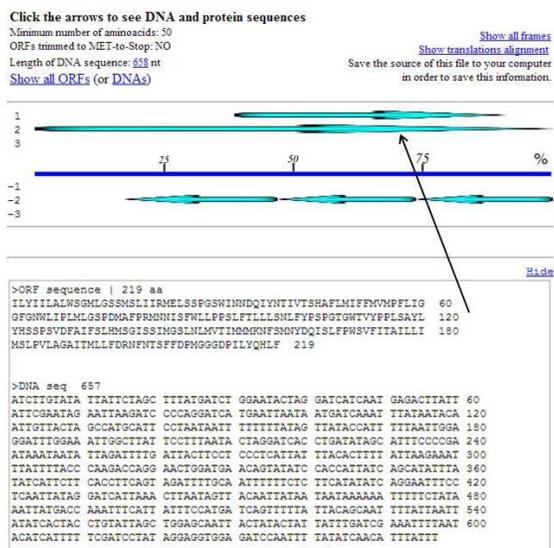


Figure 10. Sequence of longest open reading frame from barcode sequence. Arrow indicates longest open reading frame.

- Open the same "ClustalW Apis" file you used for the nucleotide alignment in Part 2.
- First, you need to translate each of the DNA sequences into amino acid sequence. Do this by going to <http://insilico.edu.es/translate/>. Copy and paste one of the sequences into the box making sure that you use only the DNA sequence and not the header. Select invertebrate mitochondrial DNA, and click on "Translate to protein". On the new page, click on the longest blue arrow, which represents the longest open reading frame (Fig. 10).

Note that the translated amino acid sequence should be 219 amino acids long.

- Copy the amino acid sequence into a new word file, adding back the ">" and a file name in the line immediately above the amino acid sequence so that your sequence is once again in the FASTA format.
- Repeat for the three other sequences, and paste into the same file leaving no spaces in between the sequences. (So it should look like the Apis file you first opened but with amino acid sequence instead of DNA.)
- Now go back and do a ClustalW alignment of the amino sequences. Are any of the SNPs non-synonymous?

Part 4. The BOLD Database

As mentioned above, the BOLD database is a database specifically for barcode sequences, and anyone can access published barcode data. You can use BOLD to compare your sequence against the BOLD database.

- To simply compare your sequences to those in the BOLD database, go to the BOLD Student Portal http://www.bold-systems.org/index.php/SDP_Home. Then select the "Identification" tab on the upper toolbar.
- Copy and paste the "Good_sequence_Spider" FASTA sequence you saved in step g in Part 1, and enter it into the search box. Make sure you click "All barcode records" for your search and hit submit.
- You are now looking at the list of the matches in BOLD to your sequence. Remember the BLAST results for this sequence? The top match was 92% similar. In BOLD, note there are sequences that are 99% similar – but the submitters knew only the order of the spider, Araneae, and in one case, the genus, *Cheiracanthium*. This is because BOLD allows investigators to submit sequences for organisms even if the genus and species is unknown. Note that the next closest match is for 96%—this is probably a different species.
- Click on the little blue arrow mark next to the Published note, and it will bring you to the record page in BOLD for the top match. It shows a picture of the spider from which the barcode was obtained. Note that the page contains a lot of information about the spider— where and when it was collected, by whom, and a picture of the actual specimen.

Materials

For 24 students (twelve groups of two).

Insect Collection

- Twelve kill jars
- Twelve aspirator setups (<http://www.bioquip.com/#1135A>)
- Twelve nets (<http://www.bioquip.com/#7612NA>)
- Tubes with 70% ethanol
- Labels (<http://www.bioquip.com/#1213>)
- Insect pins and boxes (<http://www.bioquip.com/#1208B2> and [1009](http://www.bioquip.com/#1009))

DNA Extraction

- Twelve disposable blue microfuge tubes with pestles (<http://www.fishersci.com/#K749520>) or plain Eppendorf tubes without pestles if overnight incubation
- Twelve small scissors and tweezers
- Six dissecting scopes (or magnifying glasses if scopes not available)
- DNA extraction kit for twelve samples (<http://www.qiagen.com/#69504> or similar)
- Vortexers
- Minicentrifuges
- Twelve sets of micropipettes
- Water bath at 56°C

PCR

- 700 µl GoTaq Green (<http://www.promega.com/#M712B>)
- Thermal cycler
- PCR tubes
- Primers from IDT (<http://www.idtdna.com/site>)

Agarose Gel

- Three agarose gel rigs
- Agarose and Sybr Safe
- DNA ladder – lambda HindIII is good
- TAE buffer
- UV light box
- Camera

PCR Clean-up

- GeneJET PCR Purification Kit (<http://www.thermoscientificbio.com/#K0701> or similar)

Sequence Analysis

- Computers with Finch TV
- Internet connection

Notes for the Instructor

As mentioned in the introduction, barcoding projects may be discovery based (i.e., “what types of insects live on the campus?”) or hypothesis driven (“Do different species of thrips prefer different species of plants?”). Although this module was written with insects as the target specimens, the methods can be easily adapted for other invertebrates and essentially the same protocols will work for almost any animal tissue. They will not, however, work for plants which require different primer sets and involve more difficult DNA extraction protocols.

This set of experiments from specimen collection to sequence analysis can be performed in three to four lab periods of two to three hours each.

(Note that in order to submit sequence data to the BOLD database, there are specific requirements for specimen collection and storage vouchering. This is not necessary if sequence data is being used only for in-class exercises.)

1. Collecting Specimens for Barcoding

Insects and other arthropods are virtually everywhere and can be collected a number of ways. Insects are easiest to find in the warmer months so this activity is best early in fall semester or late in spring semester. You can collect without any special equipment beyond glass jars or Tupperware, but a few inexpensive items will help. See Bland et al. (2010) for more information on collecting, identifying, and preserving insects.

Equipment to capture insects

Nets – Use insect nets to catch flying insects. No special technique is required – just do whatever works! Fold the net over the rim to prevent escape. Use nets to “sweep” soft vegetation, like long grass. An amazing number of small insects will be caught. Transferring insects from the net to the kill jar is the trickiest part. Again, do whatever works but be aware that most insects will fly upwards if allowed – so hold the net upside down to work your kill jar inside.

Kill jars – Insects are most easily killed by simply putting them in a freezer for 24 hours. But if you want to collect a lot in the field, a kill jar is useful. Use any wide-mouthed glass jar (12-16 ounces) with a tight-fitting lid. Pour a thick mixture of plaster of Paris into the bottom, enough for about one to two inches. When the plaster of Paris is completely dry, add ethyl acetate which acts as a fumigant to kill the insects. Use enough to saturate the plaster of Paris but not so much that there is standing liquid. As the ethyl acetate evaporates you will need to re-charge your jar (after several hours of collecting).

Aspirators – These are plastic vials with a cork and two tubes. One tube has a screen on the end, the other does not. You suck through the tube with the screen (to create suction without inhaling the specimen) and use the other tube to

capture the insect. This works extremely well with smaller specimens.

Aquarium nets – The type of small nets you buy for your home fish tank work well to catch aquatic insects in the shallows at the edge of a pond.

Fine paint brushes – Are useful for moving very small insects around. Wet the tip with alcohol and the insects will stick.

Forceps and blank labels – Forceps are handy for manipulating the dead specimens, labels are needed to identify the date, location, and collector for each individual specimen.

Berlese funnels – Are used to extract insects and other arthropods from soil or leaf litter. Insects that live in these environments avoid heat and light. The Berlese funnel is made of a piece of screen or hardware cloth on which you place your soil sample, a light bulb that hangs overhead, a funnel below, and a jar or preservative, such as 70% ethanol at the bottom. As the insects move away from the light, they fall through the funnel and into the beaker. You will catch many Collembola this way, as well as other arthropods not commonly seen. Soil rich in organic matter works best.

Identifying Insects

Identifying insects to species is really tough. Identifying to order is pretty easy, and is enough information for BOLD. There may be published field guides to your area but there are also numerous websites that are very useful. Some suggestions include: <http://www.cals.ncsu.edu/course/ent425/> (go to *Resource Library*, then *Spot ID*); <http://bugguide.net/node/view/15740>; <http://tolweb.org/Insecta>; <http://biokeys.berkeley.edu> (key to orders, including wingless specimens).

Preserving Insects

Any glass or plastic vials or microcentrifuge tubes are useful to store individual specimens. We find a combination of two dram glass vials (any style) and 2.0 ml microcentrifuge tubes (with screw caps) fit all the arthropod specimens we collect. Avoid tubes smaller than 2.0 ml because it is impossible to fit labels inside the tube.

2. DNA Isolation

There are several different kits that can be used to do the DNA isolation. The easiest is the prepGem Insect kit from ZyGEM (<http://www.zygem.com/Products/Products-PG-Insect.html>), which only takes about 20 minutes. However, the DNA obtained is not very clean and does not seem to store well. If you want to keep the DNA, a better option is the Qiagen DNeasy Blood and Tissue kit (see Materials section), and the directions for that kit are in the student protocols. You can also make your own reagents, which is the

least expensive option. For a review of methods, see Ball and Armstrong (2008) and *Protocols for High Volume DNA Barcode Analysis* (http://barcoding.si.edu/PDF/Protocols_for_High_Volume_DNA_Barcode_Analysis.pdf).

If using the Qiagen DNeasy kit, overnight incubation is good if you want to skip the grinding step or if your lab period is not long enough to do the two hour incubation and subsequent extraction. We find that both ethanol-preserved and dry tissue work equally well. With large insects, we take just one leg; with small insects, we may use two or three legs, or even the whole insect. Although we will grind up the legs before digesting the tissues today, this is not really necessary – you can skip the grinding step and just incubate the tissue overnight in the ATL buffer and proteinase K solution. If you have a small organism and want to save it for vouchering purposes, you can just soak the whole insect overnight, spin down the carcass and save it, and use the supernatant for the rest of the procedure. Luckily PCR does not require much DNA and we usually get some product. We have tried digesting overnight at 56°C and then saving the digest at 4°C until the next class – this works fine but the samples should be warmed to room temperature before proceeding with the extraction.

3. Setting up PCR reaction

The Barcode of Life website has lists of primer sets that have been used successfully with various types of organisms. We use the Folmer primers for invertebrates and they work quite well with most insects (Folmer et al., 1994). However, there are more specific primers for certain insects that can be used if the results with the Folmer primers are unsatisfactory.

We order the primers from IDT (<http://www.idtdna.com/site>) and then store the stocks at 100 µM in Tris-EDTA buffer. For the PCR reaction, we use 2X GoTaq Green mix from Promega. It is relatively inexpensive and already has a gel loading dye in the master mix.

The sequences of the Folmer primers LCO1490 and HCO2198, which amplify a 660-680 bp fragment of the COI gene in a wide range of invertebrate taxa, are in Table 1.

4. Run Gel to Verify that PCR Worked and to Clean-up PCR product

You can use whichever procedure you routinely use for agarose gels. The gel is simply a diagnostic tool to see if the PCR reaction worked or not. There should be one clear band at about 660-680 base-pairs and no other bands. We do use a fairly high concentration of primer in these reactions, so often there is a primer band at the very bottom of the gel, but this can be easily identified by running a no-template control sample.

Before sending the PCR samples out to be sequenced, they have to be cleaned up. Many sequencing companies will tell you that you do not have to clean up the PCR samples, but we find that the sequences are dirty if we do not clean them up. We use a Thermo Gene Jet, Qiagen or Lambda

PCR clean up kit to do this step, but we always run a gel first to make sure that we have product and that it is clean, i.e., that we only have one band at 660-680 base pairs. Because we do the annealing step at such a low temperature, we do sometimes see bands in the PCR reaction besides our desired product. Thus it is sometimes necessary to gel-purify our PCR products and then send them out for sequencing. Any gel purification kit works for this – we have used the Qiagen, Invitrogen, and Thermo kits successfully.

5. Sending DNA Samples for Sequencing

Since we want as long a read as possible for analysis, Sanger sequencing is still the best method to use for barcoding work. We send our samples out to a company to do the sequencing. We use Eton (<http://www.etonbio.com>) because we have negotiated an educational discount with them – they charge us \$5 a sample and they actually pick up samples from our campus. Genewiz (<http://www.genewiz.com/>) is the company used by the Urban Barcode Project and we understand that they have low rates for academic purposes as well.

When you send out the samples, you must also include some of your primer at 5 μ M – they need this to start the sequencing reaction. If you want to also sequence in the reverse direction, you must also send the reverse primer

Once the sequences come back from Eton, we go through them to see which have worked and which have not. For those that work, we then ask Eton to sequence the DNA using the reverse primer – this is required of BOLD in order to submit samples, but it is not necessary if you are just doing the barcoding for class projects.

6. Bioinformatics

All the programs used for the bioinformatics analysis in this paper are free and PC and Mac compatible. There are a variety of other programs that can be used to do analyses of your sequences. For example, you can have students calculate how much sequence diversity there is among members of the same species using a program such as Mega (<http://www.megasoftware.net/>). The BOLD student portal also has a variety of tools but you must submit both forward and reverse sequences to BOLD to be able to use these tools.

Note that the amplified CO1 sequence is a coding region of the DNA, and thus should be a continuous open reading frame. When translating the sequence as shown in Part 4 of the bioinformatics section, the longest open reading frame was +2. Your data may differ, however, since it is dependent on how the chromatogram is trimmed.

7. Submitting Sequences to BOLD

In order to use BOLD as a place to store and analyze student work, you must register for an account, and then register your course (http://www.boldsystems.org/index.php/SDP_Home). This information will not be available to the public, unless you ask BOLD to import it into their public database. You can then upload your students' names and emails,

and they will be sent a login code. If the data your students collect is complete and of high quality and you vouch for the specimen, you can also submit your work into the BOLD database. This is all explained on the BOLD site, under Quick Start Guide, Instructor Interface and User Guidelines.

Acknowledgements

Much of this work was funded by NSF TUES grant 1140640.

Literature Cited

- Ball, S. L., and K. F. Armstrong. 2008. Rapid, one-step DNA extraction for insect pest identification by using DNA barcodes. *Journal of Economic Entomology*, 101:523-532.
- Bland, R.G., and H.E. Jaques. 2010. *How to know the Insects*. Third edition. Waveland Press Inc., Long Grove, Illinois, 418 pages
- Folmer, O., M. Black, W. Hoeh, R. Lutz, and R. Vrijenhoek. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, 3: 294-297.
- Hamilton, J., Y. Basset, K.K. Benke, P. S. Grimbacher, S.E. Miller, V. Novotny, G. A. Weiblen, and J.D. L. Yen. 2010. Quantifying uncertainty in estimation of tropical arthropod species richness. *The American Naturalist*, 176: 90-95.
- Hebert, P. D. N., A. Cywinska, S.L. Ball, and J. R. deWaard. 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society*, 270: 313-321.
- May, R.M. 2010. Tropical arthropod species, more or less? *Science*, 329: 41-42.
- Santschi, L., R. H. Hanner, S. Ratnasingham, M. Riconsciente, and R. Imondi. 2013. Barcoding Life's Matrix: Translating Biodiversity Genomics into High School Settings to Enhance Life Science Education. *PLOS Biology*, 11:1-8.
- Stoeckle M. Y., and P. D. N. Hebert. 2008. Bar Code of Life: DNA Tags Help Classify Animals. *Scientific American*, 299:66-71.

About the Authors

Madeline Butler received a PhD in Neuroscience from the University of Rochester School of Medicine and Dentistry, and she has since worked in both industry and academia. Since 2002, Mandy has been the Academic Coordinator for the Undergraduate Laboratories in the Division of Biological Sciences at UCSD. Mandy teaches undergraduate laboratory courses and oversees the operation of the undergraduate teaching laboratories. Mandy is also involved in the San Diego Biodiversity Project which seeks to involve undergraduates in novel research experiences related to biodiversity.

Heather Henter received her Ph.D. from the Entomology Department at Cornell University. In addition to teaching she has been involved in both applied and basic ecological research and science writing for a non-science audience. She is currently the Academic Coordinator for the UC San Diego Natural Reserve System, a network of land preserves set aside

for education, conservation and research. Her primary goal is to use the campus reserves as an outdoor laboratory to increase the number of undergraduate students that are able to participate in original research.

Stephanie Mel received her Ph.D. from the University of California, San Francisco where her work focused on protein biochemistry and biophysics. As a post-doctoral research fellow at Harvard Medical School she studied cholera. Stephanie moved to a teaching position in the Division of Biological Sciences at UC San Diego, where she has worked as a Lecturer for the last 16 years. She has taught a wide variety of both lab and lecture courses in fields including Structural Biochemistry, Molecular Biology, Physiology, and Genomics. One of Stephanie's main interests has been to introduce research into the undergraduate curriculum at UCSD.

Mission, Review Process & Disclaimer

The Association for Biology Laboratory Education (ABLE) was founded in 1979 to promote information exchange among university and college educators actively concerned with teaching biology in a laboratory setting. The focus of ABLE is to improve the undergraduate biology laboratory experience by promoting the development and dissemination of interesting, innovative, and reliable laboratory exercises. For more information about ABLE, please visit <http://www.ableweb.org/>.

Papers published in *Tested Studies for Laboratory Teaching: Peer-Reviewed Proceedings of the Conference of the Association for Biology Laboratory Education* are evaluated and selected by a committee prior to presentation at the conference, peer-reviewed by participants at the conference, and edited by members of the ABLE Editorial Board.

Citing This Article

Butler, M., H. Henter and S. Mel. 2014. From Bugs to Barcodes: Using Molecular Tools to Study Biodiversity. Pages 41-55 in *Tested Studies for Laboratory Teaching*, Volume 35 (K. McMahon, Editor). Proceedings of the 35th Conference of the Association for Biology Laboratory Education (ABLE), 477 pages. <http://www.ableweb.org/volumes/vol-35/?art=3>

Compilation © 2014 by the Association for Biology Laboratory Education, ISBN 1-890444-17-0. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the copyright owner.

ABLE strongly encourages individuals to use the exercises in this proceedings volume in their teaching program. If this exercise is used solely at one's own institution with no intent for profit, it is excluded from the preceding copyright restriction, unless otherwise noted on the copyright notice of the individual chapter in this volume. Proper credit to this publication must be included in your laboratory outline for each use; a sample citation is given above.