December 2021

Species Diversity Barcoding

By Lisa Yoneda, ABE San Diego

ABE Master Teacher Fellowship Program

AMGEN® Biotech Experience

Scientific Discovery for the Classroom

The curriculum projects designed by the 2020–21 ABE Master Teacher Fellows are a compilation of curricula and materials that are aligned with the Amgen Biotech Experience (ABE) and prepare students further in their biotechnology education. These projects were created over the course of a 1-year Fellowship in an area of each Fellow's own interest. Each is unique and can be adapted to fit the needs of your individual classroom. Objectives and goals are provided, along with expected outcomes. Projects can be used in conjunction with your current ABE curriculum or as an extension.

As a condition of the Fellowship, these classroom resources may be downloaded and used by other teachers for free. The projects are not edited or revised by the ABE Program Office (for content, clarity, or language) except to ensure safety protocols have been clearly included where appropriate. We are grateful to the ABE Master Teacher Fellows for sharing their work with the ABE community.

If you have questions about any of the curriculum pieces, please reach out to us at ABEInfo@edc.org. We will be happy to connect you with the author and provide any assistance needed.

Species Diversity Barcoding Project Overview & Scope

Project Purpose: To expose and increase understanding of all levels of students to biodiversity, experimental design, and data analysis through doing an authentic science experiment

All students will design an experiment to study their local ecosystem, collect data, and analyze their initial (collected) and final data. One aspect of data collection will have students choose a few samples to do species barcoding on and collect and document the specimens for this purpose. The collected specimens will then be sequenced, and barcoding/species data will be returned to the students/class. The class can then do a final analysis to see if the returned data support or doesn't support their initial conclusions. Classes can choose to do different projects each year or do an ongoing research project, but in both cases, the barcoding/species data would be published and be available in scientific databanks such as Genbank.

Barcoding Prior Knowledge:
This project depends on at least a basic understanding of what DNA barcoding is, potential uses of barcoding, and techniques needed to generate barcodes. I have put together a DNA barcoding background document that goes over some basics. It is not by any means comprehensive but goes through some of the essentials for understanding barcoding needed to be successful with doing a student driven research project.

Project Overview:
Through this project, students will use modern methods of species identification to answer a student- or class-determined question about biodiversity. The project is scalable to any classroom or age group but is designed for full implementation in a high school biotechnology or other advanced high school science class. At any level, the project will give students an authentic citizen-science biodiversity experiment in which they can publish their results to be used by the scientific community. At any level, the emphasis is on scientific design and data analysis. The project is tiered so biotechnology students do an internship (ScienceBridge) that supports the learning of other classrooms. This allows younger classrooms and/or classrooms without modern biology equipment to still participate in the species biodiversity barcoding project and do an authentic science experiment. The following is a list of suggested project components at different class levels.

Project Component	Implementation Needs/Level	Support**
Part 1: Introduction and Sample Collection		
Scientific method/experimental design	All	
Data collection	All	
Initial data analysis	All	
Evolution/Cladogram creation	High school	** Practice for Part 3
Part 2: Molecular Genetics Laboratory Steps		
DNA extraction	All, but low yields can impact ability to sequence	ScienceBridge Program (pilot) or Kit
PCR & gel electrophoresis	Biotechnology/Advanced science course	ScienceBridge Program (pilot) or Kit
Amplification clean-up	Biotechnology/Advanced science course	ScienceBridge Program (pilot) or Kit
Sequencing	Ship for now	ScienceBridge Program (pilot) or Kit

Project Component	Implementation Needs/Level	Support**
Part 3: Species and Experiment Analysis		
Sequencing cleanup/initial analysis	Biotechnology/Advanced science course/teacher	ScienceBridge Program (pilot) or Kit
Species identification	all	
Speciation/Cladogram analysis	High school	
Final experiment analysis	all	

*****Support:** If you do not have the appropriate modern biology tools to do these steps within your classroom, then you will need to purchase a kit or work with another class or local researcher that can do these steps for you.*

Currently, Carolina has a barcoding kit that supports the collection and preparation of 25 samples for $228: Item 211386P (does not look like sequencing is included but should be about $6/read + shipping). Plant, fish, and insect/mammal primers are included. Note: For high-quality sequencing and analysis, you need to do two reads per sample: forward and reverse.

Project Component Breakdown and Resources:
The following breakdowns on each step of the project can help you choose which parts of the barcoding project you wish to implement. Most of these steps can be completed in a single hour, but it can be expanded to several lessons depending on your goals and student interest.

Part 1: Introduction and Sample Collection

This section contains lessons that can be used with all levels of classes and students. These lessons will help students to better understand the scientific method and allow them to conduct authentic research on a question that interests them. It is important to note that there are limitations to barcoding technology, legal restrictions to obtaining samples, and ethical considerations. Thus, in general, most K–12 schools would typically work with plants, vertebrate fish, insects, soil invertebrates, or fungi if collecting from the wild. Whichever you choose, you will need to have primers to match your sample, so it is important to know ahead of time which primers you will have access to before designing the experiment.

Overall, this project is designed to help students understand how the scientific method works, in particular, a realization that science is not a single experiment with a perfect answer at the end. It is instead a continual process, where scientists take data and constantly adjust their thinking, make experimental changes, and ask more questions, all in a way to constantly reassess and refine their understanding of the initial question/problem. Having students study and use the interactive diagram Understanding Science, put together by UC Berkeley, can be a useful teaching tool to help students understand that science is not a straight path.

Scientific Method/Experimental Design:
This is the heart of the authentic research. Students/classes will determine a research project and design a way to test their ideas. The overall purpose of this project is on better understanding and tracking species biodiversity. Thus, the experiment should relate to the local ecosystem and finding/identifying the species found there. Encourage students to compare two sites that vary in some way. (Here are some archived DNLC student projects.) The complexity of the project should match the science standards and abilities of students. However, all projects should include at least identifying a question/problem, designing an experiment that would have both a control and experimental test sites,

deciding on what observational data you want to collect, and what actual samples you will collect for the species barcoding. Have students write a hypothesis before going onto the data collection.

Class Time: 1–3 hours

Data Collection:

Students conduct their initial research and collect data to answer their question or problem. Depending on the scope and level of science ability, the observational data can be simple or complex but should include multiple ways of collecting evidence to compare the two test areas. A typical ecological survey might include an insect trap and/or a 2-meter transect. Making some type of observational recording form on which students can collect data can be helpful for getting them to collect comparable data and keeping track of it. Here is a sample recording form from a bumblebee community science project in the UK by Moors for the Future Partnership. Have students consider recording these additional factors: GPS location, humidity, soil moisture levels, and shade cover. With plants, overall size, new growth, flowers, or browning can also be helpful. Using *Seek* or *iNaturalist* can also help students identify species they are looking at. Depending on age and district requirements, teachers can have students have their own accounts or create a class account that students can use. *Seek* and *iNaturalist* is discussed as a research tool in another section. During this phase, students collect samples to barcode. For plants, a single leaf is sufficient. Insects and other small invertebrates can be put into a freezer to humanely kill and preserve. Typically, you would want good pictures of anything you want to barcode to help with the final identification verification process, but the actual amount of tissue needed to get a good barcode is typically the size of a grain of rice. For a typical class, to keep the cost reasonable, it is good to have students select up to six samples to do barcoding analysis on.

Class Time 1–3 hours (but can set up a transect to allow for ongoing data collection: 15–60 min)

Initial Data Analysis:

Students should use the data collected from their recording forms to graph differences and similarities between the sites. Typically, the *Y* axis would be abundance, and the *X* axis would be what they were recording at the different sites. A bar graph with the sites grouped for each *X*-axis variable is usually suitable for identifying significant differences between the sites. More advanced classes could do % abundance graphs or even calculate diversity indexes to make the numbers more meaningful to compare. Sample data graphs are attached. The idea here is to have students use the data they have to make initial conclusions to their question or problem. Then when they get the barcoding results back, they can go back to their initial conclusions and see if the additional information helps clarify, adds to, or even disagrees with their initial conclusion. Having students take additional transect data at different intervals, seasons, weather conditions, or just over time can add data to this project as well.

Class Time: 1–3 hours

Evolution Tree/Cladogram Creation:
This section is not part of the barcoding project directly. However, if students have never built and/or read cladograms before, then they will need some practice if you plan to do the full barcoding analysis. This supplemental lesson can be implemented at any time to prepare students to be able to make and read cladograms. If you are not doing Part 2 yourself but sending it out to be done, then this can be a good time to implement these lessons. Depending on the level of the class and the depth you wish to go, you can have students do all these lessons or just some of them.

- NOVA's Evolution Lab is a great way for students to learn about cladograms/phylogenetic trees and to learn how to interpret data to build them.
- BLAST introduction: Students learn to do BLAST searches.
- Sample blue line modules

Class Time: 1–2 hours for each lesson

PART 2: MOLECULAR GENETICS LABORATORY STEPS

If your school does not have access to gel electrophoresis and thermocyclers, then it will be necessary for you to partner with someone to help you with the completion of these steps.[1] If you do have access to gel electrophoresis and thermocyclers, Carolina offers a kit to help you make samples that are suitable for barcoding.

DNA Extraction:
There are many ways to do this, and most will work for DNA barcoding. One of the easiest ways in the classroom is to use Chelex beads. Students will need to mash up the sample before heating to try to help release DNA. This is particularly important for plant cells and insects. It is also important to realize that too much DNA can actually prevent a good PCR reaction from occurring. Students will want to use large pieces of the specimen, but usually a piece the size of a rice grain is plenty. Using a picodrop to determine DNA concentration can be helpful but is not required.

Class Time: 30–60 min

PCR & Gel Electrophoresis:
Using PCR beads is extremely helpful as they can be stored at room temperature and don't need to be mixed with anything other than the primer and the specimen DNA before use. Remember, you will need to use the correct primers with each sample that you run to get any PCR products. If you are ordering your own primers, DNA learning center has a good list of current barcoding primers. It is important to note that you only want to run 5 uL of your product in a gel as barcoding uses 10 uL for each sequence, and PCR reactions are almost always 25 uL. Depending on your budget and how accurate you want your

[1] If you are in San Diego, CA, USA, then currently ScienceBridge (Mira Mesa Tech Site) is running a pilot program that may be able to assist.

results, you will want to run either the forward sequence (10 uL) or both the forward and reverse sequences (20 uL) for barcoding analysis.

Class time: PCR 30–60 min; gel electrophoresis 60 min

Amplification Clean-up:
At this time, this step is not required (nor does it affect the cost) of samples that are sent to Genewiz for sequencing (see Optional Step).

Optional Step

Sequencing:
Most classrooms are going to need to send their samples out for sequencing. Genewiz, for instance, offers discounts for educational classrooms but is not the only company out there that can do sequencing. If you plan on doing both the forward and reverse sequences to increase the reliability of your results, you will need to separate each sample into 10 uL aliquots before sending.

Class time: 30 min

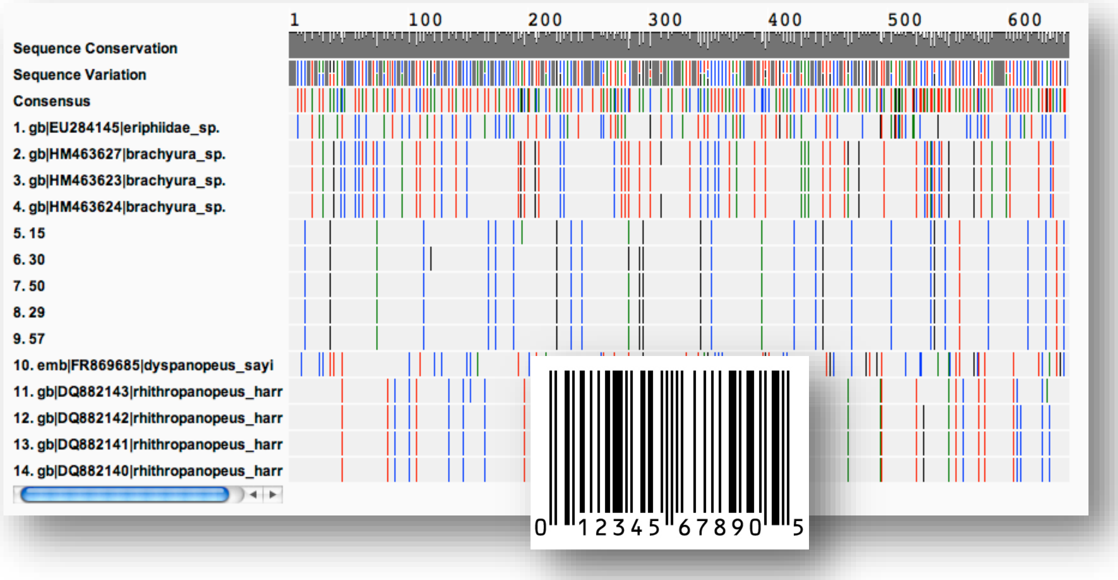
Part 3: Species and Experiment Analysis

For each sample that has a successful sequence reaction, you will get one or two sequences returned to you. Analyzing it and determining the most likely species will take a little bit of computer work. Depending on the level of your class, the teacher or students can do the computer work. All levels of students should be able to interpret the final sequencing alignment and cladogram and then use this to go back to their initial research question and see if it supports or raises questions about their initial conclusions.

The DNA Subway program is a free self-contained program that houses many of the analysis tools used by researchers in a single program. This is great because it remembers the specific project details that your students are doing and allows them to make changes as they go, without having to reload information. In addition, if you use Genewiz for your sequencing, the data will go directly to DNA Subway without your having to upload it. (Other companies may also do this.) The DNA Subway Sequence Analysis Directions will help you use the tools in the DNA Subway to analyze your data. Once you've analyzed them, you will have both a species sequence comparison diagram and a cladogram to use for your final project analysis. Depending on your project, you might have all the data from your project in a single diagram, or you might have separate ones for different questions. Either way, the students should be able to take these final pieces to add to their research conclusions.

Class time: 1–3 hours

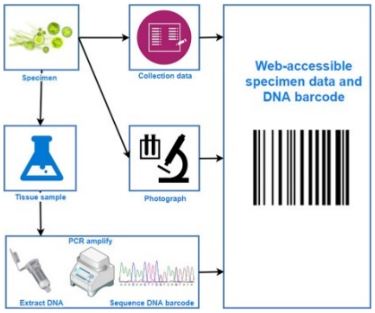
DNA Barcoding Background: What is it and how does it work?


1
100
200
300
400
500
600
Sequence Conservation
Sequence Variation
Consensus
1. gb|EU284145|eriphiidae_sp.
2. gb|HM463627|brachyura_sp.
3. gb|HM463623|brachyura_sp.
4. gb|HM463624|brachyura_sp.
5. 15
6. 30
7. 50
8. 29
9. 57
10. emb|FR869685|dyspanopeus_sayi
11. gb|DQ882143|rhithropanopeus_harr
12. gb|DQ882142|rhithropanopeus_harr
13. gb|DQ882141|rhithropanopeus_harr
14. gb|DQ882140|rhithropanopeus_harr
0 12345 67890 5

Barcode Image: Cold Spring Harbor Laboratory, DNA Learning Center

DNA barcoding is increasingly used in research to identify a specific species. This is important, as many species are difficult, if not impossible, to identify by sight alone. This leads to potentially inaccurate and/or inconclusive results if a species is misidentified. In addition, more and more attention is being focused on microbiology which can take an immense amount of time and resources to identify with more traditional methods of identification, if even possible. Finally, although science has identified over 2 million species, it is estimated there is more than 8 million species that haven't even been identified and classified yet. Better methods of quickly and accurately identifying species are greatly needed not only for current research and understanding, but to help combat biodiversity loss. DNA barcoding helps to fill this need. The basic concept behind DNA barcoding is to be able to identify a specific DNA sequence that is unique to that species, much like each item at a store has a unique barcode that a scanner can read.

Recent improvements and cost decreases in both PCR and sequencing make the idea both accessible to any reasonably equipped lab and financially viable. The final step is creating a database of species barcodes to compare so that any species barcode generated can be easily identified. This is the current status of species barcoding, a complete inventory of comparison is not yet available. So when an individual species barcode is made from a sample, it may or may not match to a sample in the database. Typically, some organisms such as fish are highly represented in the known database and are likely to come back with a match. Other organisms, such as less common insects and native plants have much larger known gaps in the data base and so are less likely to be 100% positively identified by doing a species barcode. However, new barcodes can then be authenticated through a peer review submission process to make sure that there is sufficient documentation on the donor species and sequence is of high quality. Once accepted, new barcodes add to the database of known species and make future matches more likely.

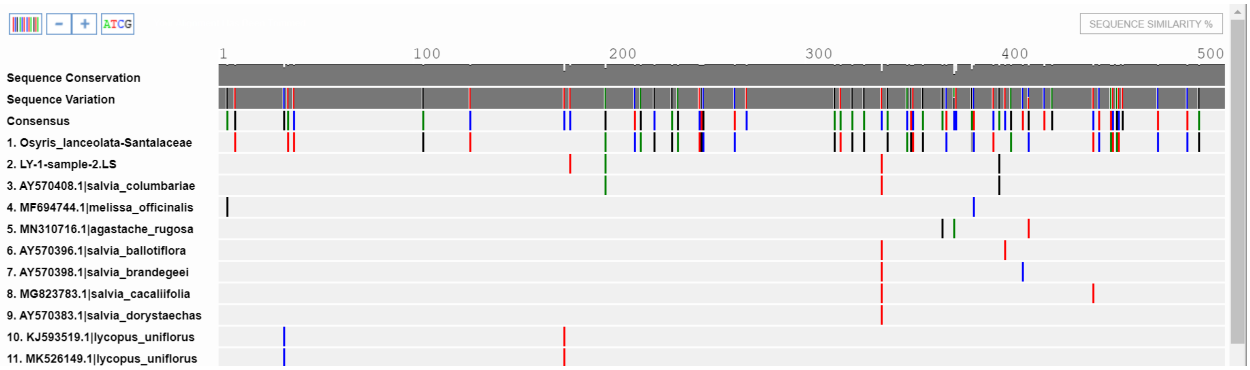


Science Direct image of the steps for creating a DNA barcode

Molecular Genetics Background

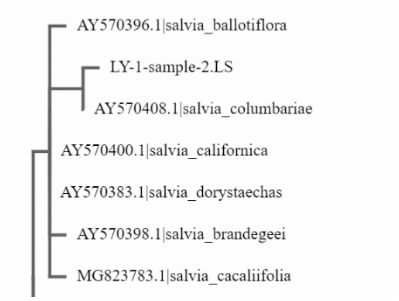
Barcoding relies on two different molecular biology techniques, Polymerase Chain Reaction (PCR) and Sequencing. PCR is the same technique used to generate modern DNA fingerprints that are put into the CODIS system and used by law enforcement. In this case, PCR and DNA fingerprinting is used to identify unique differences between humans. Since all humans share 99.9% of the same DNA sequence, we simply target areas that are highly variable from person to person. If you target enough of these variable sections of the DNA, you can statistically eliminate any random person having a DNA fingerprint match. Primers are what we use to target where we want to copy a DNA sequence during a PCR reaction. In PCR, primers match sequences much like the control F function in a document. With primers you can then find the beginning and end of a DNA sequence that you want to copy. During the PCR reaction you make billions of copies of the target sequence. Once you've made your copies you can then use gel electrophoresis to determine the fragment size. For a DNA fingerprint this is plenty to identify an individual since we target 13 or more different locations in the human DNA genome. The combination of different lengths of DNA outcomes can then be used to match someone's DNA to the sample.

For DNA barcoding, we can't use the exact same technique we use for DNA fingerprinting in humans. One initial problem is that we want an outcome that is identical for all individuals of the same species and different for each other species. This means that the area we target has to be found in all species genomes but have variation. Unfortunately, no one location has been identified to meet both requirements. However, since organisms in the same kingdoms tend to have common conserved sequences, primers have been identified that can be used with most organisms in each taxonomic group. Typically, these primers target regions of the chloroplast *rbcL* gene, mitochondrial COI gene, and nuclear ribosomal ITS region. The primers used bind to a highly conserved area that bracket a highly variable area of the genome. This gives a unique PCR product that can be confirmed with gel electrophoresis. However, much of the variable region is not different in length from species to species, but different in sequence. Gel electrophoresis, which is used in DNA fingerprinting, can only tell differences in length, thus the final PCR product must be sequenced to determine the DNA barcode. Sequencing the PCR product will give the ATCG sequence of the targeted region, when that sequence is compared to other species a computer can easily highlight the parts of the sequence that have high variability. These differences are usually highlighted for each species in the comparison with a color for the sequence change (A = green, T = red, G = black, C = blue).


ATCG
SEQUENCE SIMILARITY %
1
100
200
300
400
500
Sequence Conservation
Sequence Variation
Consensus
1. Osyris_lanceolata-Santalaceae
2. LY-1-sample-2.LS
3. AY570408.1|salvia_columbariae
4. MF694744.1|melissa_officinalis
5. MN310716.1|agastache_rugosa
6. AY570396.1|salvia_ballotiflora
7. AY570398.1|salvia_brandegeei
8. MG823783.1|salvia_cacaliifolia
9. AY570383.1|salvia_dorystaechas
10. KJ593519.1|lycopus_uniflorus
11. MK526149.1|lycopus_uniflorus

In this sequence alignment or barcode you can see that number 10 & 11 have the identical pattern and when you look at the species information, they are 2 different samples from the same species. This is how barcoding should work, each species should have a unique pattern that is the same within the species, but different from others. #2 is the barcode generated by the researcher and nothing is an exact match. This highlights the concept that the species barcode database is not yet complete, but for common species it is fairly well fleshed out.

However, my sample #2 shares 3 out of the 4 differences it has with #3, so it is probably closely related. These barcodes can be used to generate cladograms or phylogenetic trees. In the cladogram I generated from these barcodes, you can see that my sample (#2) and #3 are in the same clade. In addition, many of the salvia genus that also share the same T mutation in the ~330bp are just one clade out, but species 4, 5, 10/11 are further out (not shown) showing a larger evolutionary gap even though 10/11 is the same genus. Thus, barcodes are not just useful for quickly identifying species, but also for better understanding of evolutionary relationships.


AY570396.1|salvia_ballotiflora
LY-1-sample-2.LS
AY570408.1|salvia_columbariae
AY570400.1|salvia_californica
AY570383.1|salvia_dorystaechas
AY570398.1|salvia_brandegeei
MG823783.1|salvia_cacaliifolia

Resources:

- iBOL: international Barcode of Life This is a 3 part global plan to fully implement barcoding all the species on earth. Has some interesting information on how barcoding can be used and explains where we currently are in barcoding species worldwide.
- DNA learning Center: Barcoding 101 This has an amazing amount of resources for implementing barcoding in your curriculum as well as current and past student projects. Finally, it has a few citizen science barcoding that you can potentially join with your classes. Note, most of these are out of New York, but not all.
- Carolina: Background information for barcoding and information on the Carolina Barcoding kit
- Agricultural uses of barcoding video
- Sanger Sequencing Video This goes through how Sanger sequencing works and how you get the trace files.

Name:

Analyzing Amino-Acid Sequences to Determine Evolutionary Relationships

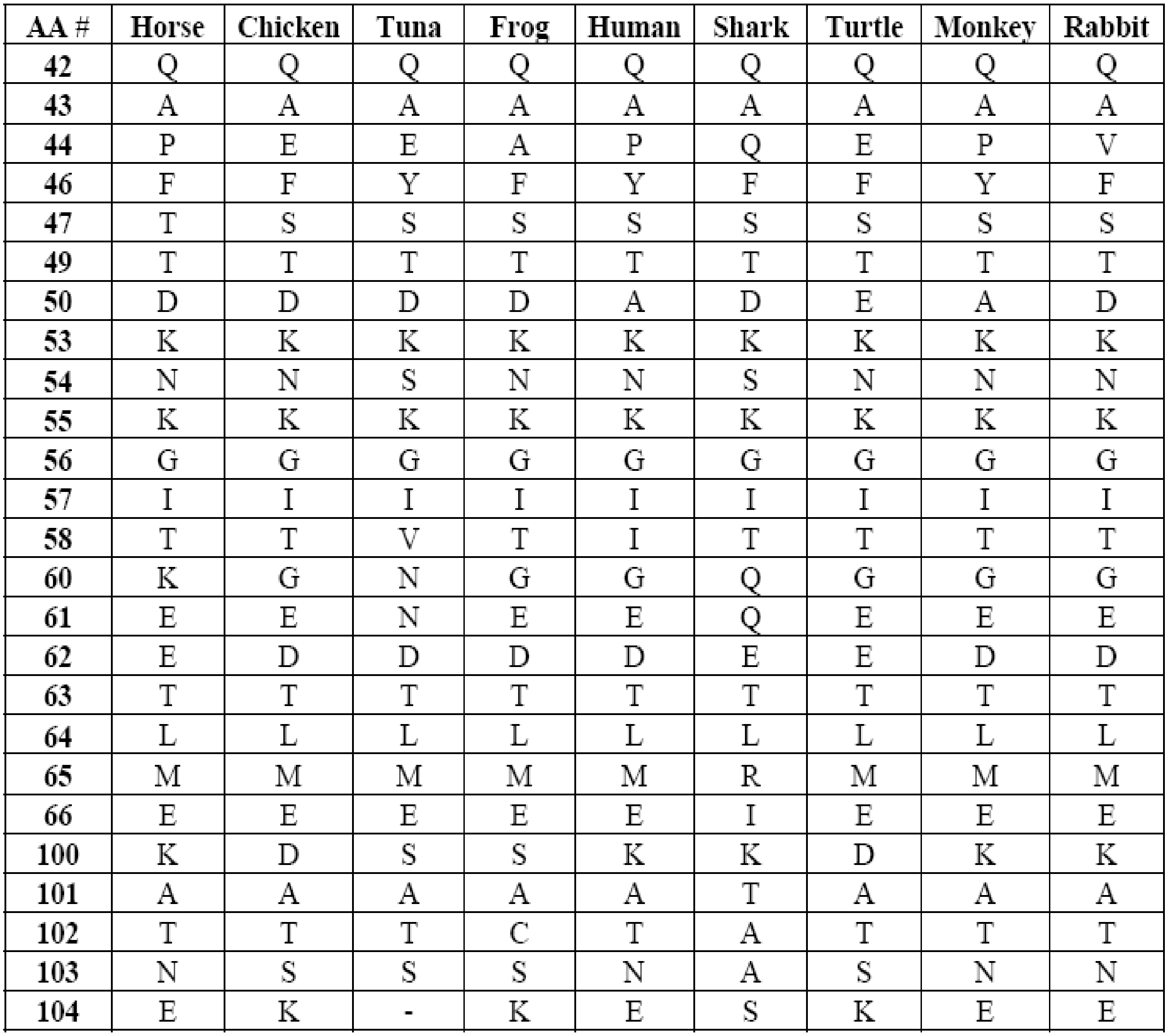
Background Information:
The biochemical comparison of proteins is a technique used to determine evolutionary relationships among organisms. Proteins consist of chains of amino acids. The sequence, or order, of the amino acids in a protein determines the type and nature of the protein. In turn, the sequence of amino acids in a protein is determined by the nucleotide sequence in a gene. A change in the DNA nucleotide sequence (mutation) of a gene that codes for a protein may result in a change in the amino acids sequence of the protein. Biochemical evidence of evolution compares favorably with the structural evidence of evolution. Even organisms that appear to have few physical similarities may have similar sequences of amino acids in their proteins and be closely related through evolution. Researchers believe that the greater the similarity in the amino acid sequences of the two organisms, the more closely related they are in the evolutionary sense. Conversely, the greater the time that organisms have been diverging from a common ancestor, the greater the differences that can expected in the amino-acid sequences of their proteins.

Two proteins are commonly studied in attempting to deduce evolutionary relationships from differences in amino-acid sequences. One is cytochrome-c and the other is hemoglobin. Cytochrome-c is a protein used in the electron transport chain of cellular respiration and is found in the mitochondria of many organisms. Hemoglobin is the oxygen-carrying molecule found in the red blood cells.

Procedure:

Part 1 – Cytochrome-c

A cytochrome-c molecule consists of 104 amino acids. The chart below shows the amino acids sequence in corresponding parts of the cytochrome-c molecules of nine vertebrates. The numbers along the side of the chart refer to the position of these sequences in the chain. The letters identify the specific amino acids in the chain.



AA #	Horse	Chicken	Tuna	Frog	Human	Shark	Turtle	Monkey	Rabbit
42	Q	Q	Q	Q	Q	Q	Q	Q	Q
43	A	A	A	A	A	A	A	A	A
44	P	E	E	A	P	Q	E	P	V
46	F	F	Y	F	Y	F	F	Y	F
47	T	S	S	S	S	S	S	S	S
49	T	T	T	T	T	T	T	T	T
50	D	D	D	D	A	D	E	A	D
53	K	K	K	K	K	K	K	K	K
54	N	N	S	N	N	S	N	N	N
55	K	K	K	K	K	K	K	K	K
56	G	G	G	G	G	G	G	G	G
57	I	I	I	I	I	I	I	I	I
58	T	T	V	T	I	T	T	T	T
60	K	G	N	G	G	Q	G	G	G
61	E	E	N	E	E	Q	E	E	E
62	E	D	D	D	D	E	E	D	D
63	T	T	T	T	T	T	T	T	T
64	L	L	L	L	L	L	L	L	L
65	M	M	M	M	M	R	M	M	M
66	E	E	E	E	E	I	E	E	E
100	K	D	S	S	K	K	D	K	K
101	A	A	A	A	A	T	A	A	A
102	T	T	T	C	T	A	T	T	T
103	N	S	S	S	N	A	S	N	N
104	E	K	-	K	E	S	K	E	E

Compare the amino-acid sequence of human cytochrome-c with that of each of the other eight vertebrates. For each vertebrate's sequence, count the number of amino acids that differ from those in the human sequence. List the eight vertebrates in order from the fewest differences to the most differences in the table below.

Cytochrome-c Amino-Acid Sequence Differences
Between Humans and Other Vertebrate Species

Species	Number of Differences from Human Cytochrome-C

- According to this evidence, which organism is the most closely related to humans? Which is least closely related to humans?

- Frog and turtle cytochrome-c molecules have the same number of differences from human cytochrome-c. Which vertebrate, frog or turtle, would you put higher on the list? Make an educated guess, and explain your answer.

- The values listed for the chicken and the horse differ only by one. Can you conclude from this that the chicken and the horse are very closely related to each other? Why or why not?

- Now look at the amino acid sequences and see if you can build a basic cladogram. You will not have a full one, just see if you can find, the outgroup, the closest relation(s) to humans, and any other organisms that seem to be fairly similar to each other and maybe are in the same clade. (See if you can get 3-5 clades) Draw it below and give a short annotation for each clade on what separated them (and kept them together if there are multiples in the clade).

- Now do a BLAST comparing the chicken and the horse. What did you find when you did this? What does this indicate about their relationship? Does this match with what you got when you just compared number of differences from humans? Why or why not?

- Now try a BLAST comparing all the sequences to humans. Look at the Distance Tree of Results and compare this to your cladogram that you have above. Draw the cladogram that the BLAST suggests and annotate why this is different from your cladogram above.

BLAST Introduction (Uses NCBI protein match rather than DNA subway)

Materials Needed:

- **Analyzing Amino-Acid Sequences to Determine Evolutionary Relationships (modified from unknown original source)**
- **Internet access to use NCBI website**
- **Cytochrome C BLAST sequences**

Completion of this assignment will allow you to complete BLAST part 2: BLAST nucleotide matches

In this introduction activity, you will be learning how to do BLAST searches using protein sequences.

You're going to start with a fairly simple amino acid sequence comparison of the cytochrome C in your cladogram assignment. The NCBI is the US national center for biotechnology information and houses all the known sequences of all species. These sequences are both genetic (nucleic acid) and protein sequences. Researchers constantly submit new sequences to be included in the database. This allows researchers to go to one place to scan for sequences and to use comparisons to help them understand more about what they are researching.

Our first search is going to be an amino acid comparison search using the BLAST program (the link is on the right hand side of the NCBI page.

We'll do our first search together during class, but here are some helpful reminders/hints:

- Query Sequence is the one you want to compare to either other sequences you put in or the database at NCBI
- If you want to compare a query to specific sequences, check the align two or more sequences box and then enter the comparison sequence(s) into the enter subject sequence box that appears (if you have more than 1 sequence here, you need to format it correctly:
    - line 1: >name
    - line 2: sequence
    - repeat for each additional sequence

- Click BLAST
    - In the description box the following will help you analyze the value of your results
    - Query Cover: % of query sequence that aligned
    - E value: ideal smaller than 1e-50, but 0.01 considered good homology match
    - Percent Identify: % of residues that match
    - Accession length: # residues
    - Accession number: what did you match the query to?
        - if database- then gives you database ID of matches
        - if you input, then gives query # for each input... needed for understanding alignment matches.
        - Distance tree of results will give a suggested cladogram
    - In the alignments tab you can see how the sequences line up (pairwise is default)
        - Identities: identical matches (same letter)
        - Positives: different amino acid, same property (+)
        - Gaps: insertions/deletions or missing info

        - changing alignment view to Query-anchored with dots for identities shows all sequence changes compared to the query in one picture
- Compare the cladogram here with the one you drew, what similarities and differences do you see?

BLAST Sequences for Cytochrome C

>Horse

QAPFTTDKNKGITKEETLMEKATNE

>Chicken

QAEFSTDKNKGITGEDTLMEDATSK

>Tuna

QAEYSTDKSKGIVNNDTLMESATS

>Frog

QAAFSTDKNKGITGEDTLMESACSK

>Human

QAPYSTAKNKGIIGEDTLMEKATNE

>Shark

QAQFSTDKSKGITQQETLRIKTAAS

>Turtle

QAFSTEKNKGITGEETLMEDATSK

>Monkey

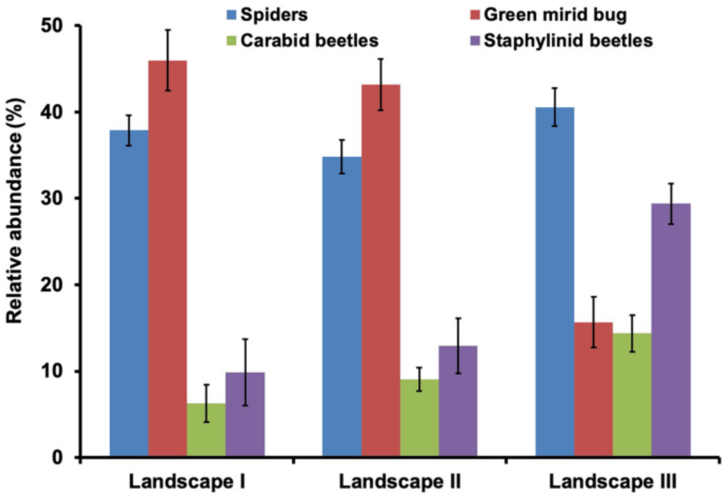
QAPYSTAKNKGITGEDTLMEKATNE

>Rabbit

QAVFSTDKNKGITGEDTLMEKATNE

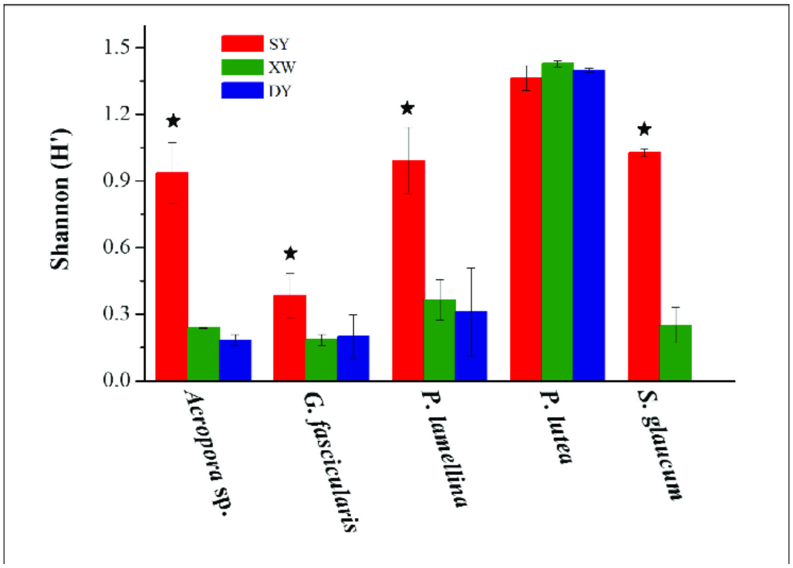
Data Analysis Sample Graphs

Comparison of different insect abundance between sites



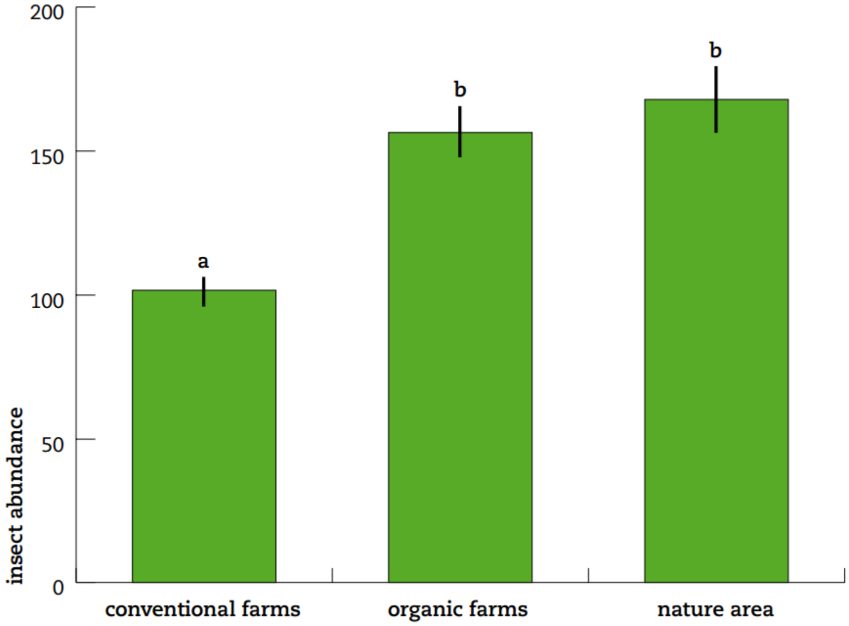
https://rdcu.be/cx1Gj

Effect of locations on coral species diversity



Gong, Sanqiang & Chai, Guangjun & Xiao, Yilin & Xu, Lijia & Yu, Kefu & Li, Jinlong & Liu, Fang & Cheng, Hao & Zhang, Fengli & Liao, Baolin & Li, Zhi-Yong. (2018). Flexible symbiotic associations of symbiodinium with five typical coral species in tropical and subtropical reef regions of the Northern South China Sea. *Frontiers in Microbiology. 9.* 10.3389/fmicb.2018.02485.

Comparison of overall insect abundance at different land use sites



4. Mean insect abundances per cow dung pat for the three farming systems. One standard error is given around the mean. A significant difference was found between a and b (P = 0.039).

https://reducing-suffering.org/how-cattle-grazing-affects-insect-populations/

BUMBLEBEE TRANSECT RECORDING FORM

CommunityScience

Lead surveyor		**Total number of surveyors**	

Transect Name	***Snake Summit S1*** – PW south from gate	**Start time**	
Date		**Finish time**	

Temperature	This can be taken from a weather forecast or car temperature sensor		
Wind (please circle)	Calm	Light breeze	Strong breeze
Sun (please circle)	Sunny	Sunny intervals	Cloudy

Section (see map & PTO)	**Bilberry bumblebee**	**Tree bumblebee**	**Red-tailed bumblebee**	**Other (record species name if known)**
1 135 metres)				
2 308 metres				
3 120 metres				
4 233 metres				
5 117 metres				
6 200 metres				

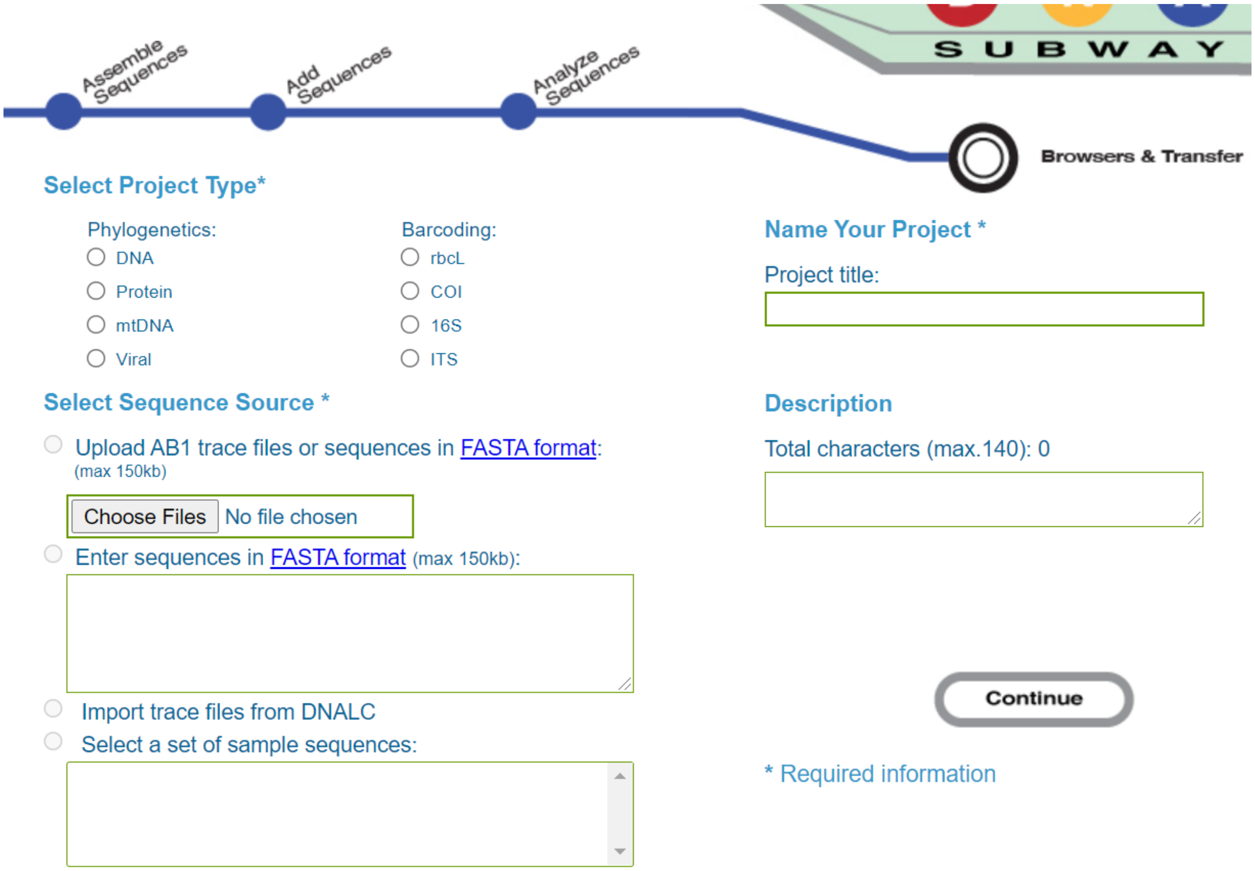
Did you note any interesting or unusual behaviour? Have there been any significant changes in habitat or management since your last visit or from the transect map?

DNA Subway: Analyzing Sequences Using the Blue Line

Login to your DNA Subway account: https://dnasubway.cyverse.org/ (free accounts just sign up)

Once you're in, click on my projects

- To continue a prior project: click on the project title you want to work on
- To start a new project, click on the blue square labeled: Determine Sequence Relationships. This will bring up the following project page:



Select Project Type*

Phylogenetics:
○ DNA
○ Protein
○ mtDNA
○ Viral

Barcoding:
○ rbcL
○ COI
○ 16S
○ ITS

Name Your Project *

Project title:

Select Sequence Source *

○ Upload AB1 trace files or sequences in FASTA format:
(max 150kb)

Choose Files No file chosen

○ Enter sequences in FASTA format (max 150kb):

○ Import trace files from DNALC

○ Select a set of sample sequences:

Description

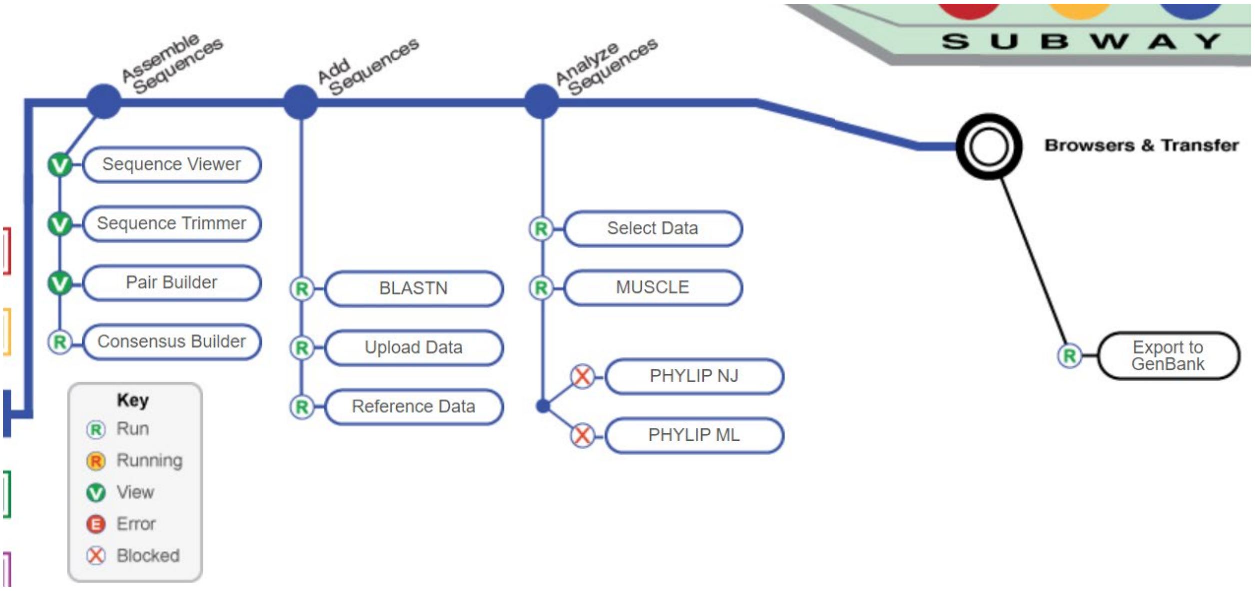
Total characters (max.140): 0

Continue

* Required information

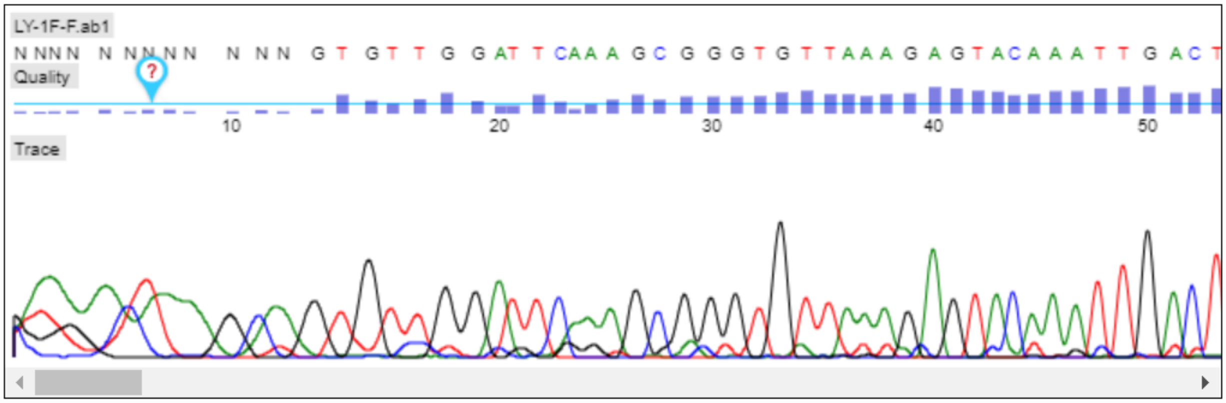
- Click on DNA in phylogentics
- Click on the correct barcoding sequence (This will depend on what primer you used)
  - rbcL =
- Depending on how you received your sequences you will need to upload a trace file, copy and paste the sequence, or import trace files from DNALC
  - Note: if you are just practicing, the select a set of sample sequences and you will be given sample sequences to choose from
  - If you used Genewiz for your sequencing, then you will select the import trace files from DNALC and search by your tracking # on your order receipt (sent in with your sample and on your Genewiz account).

Once you have uploaded the sequences, you can now start working on your project (when you continue you automatically go to the screen below):



All of these steps will start with an R or an X, meaning you haven't done them yet. As you complete the steps they will change from an R to a V. Once they are at a V, you can still go back and edit, but you can also go onto next steps.

Your first job is to Assemble Sequences. This is where you will check the quality of your sequence read that you get. You can do this with either only a forward read, or with both a forward and backward read of your PCR product.


LY-1F-F.ab1
N NNN N N NNNN N N N G T G T T G G AT T CAA A G C G G G T G T T A A A G A G T A C A A A T T G A C T
Quality
?
10
20
30
40
50
Trace

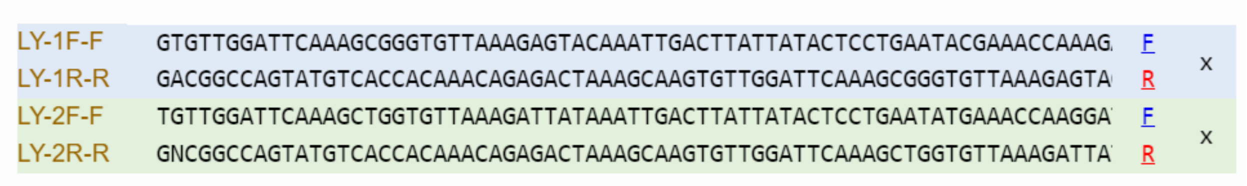
It’s normal to have the beginning to be N’s with basically 0 Phret scores and then the first section of know bases to have low Phret scores as seen above. The reverse will be true for the end of the read. This will be dealt with in the next step. At the bottom of the trace, you will see the actual florescence read that the computer used to determine the sequence. In general, a high peak means high consistency and leads to a higher Phret score. You can override the consensus sequence at the top, but generally you want to do that once you’ve compared the forward and reverse sequences together and then analyzed any inconsistencies. This will be step 4

Step 2: Sequence trimmer

This is just to cut off the ends of the reads that don’t have good reliability. Generally, just hit auto trim.

Step 3: Pair Builder

This is where you pair the forward and reverse reads. If you click try auto pairing, you can see if it can automatically pair your forward and reverse reads. When you’re done it should look like the following for each pair

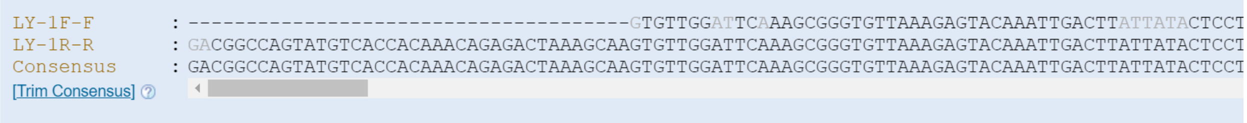


```
LY-1F-F   GTGTTGGATTCAAAGCGGGTGTTAAAGAGTACAAATTGACTTATTATACTCCTGAATACGAAACCAAAG   F
                                                                                      x
LY-1R-R   GACGGCCAGTATGTCACCACAAACAGAGACTAAAGCAAGTGTTGGATTCAAAGCGGGTGTTAAAGAGTA   R
LY-2F-F   TGTTGGATTCAAAGCTGGTGTTAAAGATTATAAATTGACTTATTATACTCCTGAATATGAAACCAAGGA   F
                                                                                      x
LY-2R-R   GNCGGCCAGTATGTCACCACAAACAGAGACTAAAGCAAGTGTTGGATTCAAAGCTGGTGTTAAAGATTA   R
```

Step 4: Consensus Editor

The first thing you should do here is edit the name. Groups of students can work on the same project, so each group should find their pair and edit the name so others don’t try to work on it. If all the pairs are for the same group, then each one should be labeled with the actual sample name so it is easier to track the data.

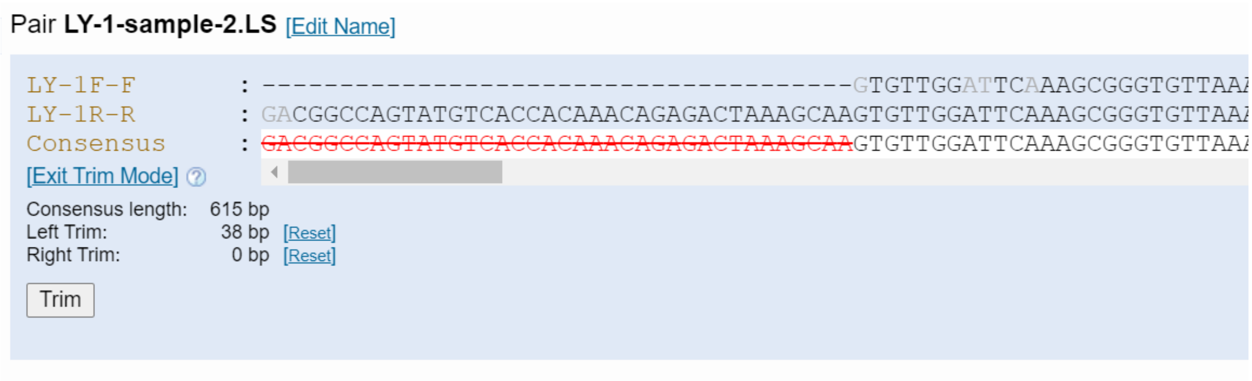
There should be 3 lines of data. It is important to note, that for simplicities sake the reverse one is NOT identical to the trace, but the opposite base pair. This makes it easier to stop differences between the forward and the reverse read, since by doing this the top two lines should be identical rather than base pair matches.



```
LY-1F-F      : --------------------------------------GTGTTGGATTCAAAGCGGGTGTTAAAGAGTACAAATTGACTTATTATACTCCT
LY-1R-R      : GACGGCCAGTATGTCACCACAAACAGAGACTAAAGCAAGTGTTGGATTCAAAGCGGGTGTTAAAGAGTACAAATTGACTTATTATACTCCT
Consensus    : GACGGCCAGTATGTCACCACAAACAGAGACTAAAGCAAGTGTTGGATTCAAAGCGGGTGTTAAAGAGTACAAATTGACTTATTATACTCCT
[Trim Consensus] ?
```

Grey letters show where the trace had a Phret score below 20 and so is not reliable. However, the consensus might not show an N if the reverse had a good read. The beginning and end will have one of the reads with - - - instead of a base. This is often due to where the primer bound and or the length of the target sequence. Sanger sequencing has a max read length of about 800 bp and is typically only reliable up to the 500-600 bp range. As PCR products for barcoding typically gets in the 600 bp range this will affect the ends of your sequence consensus.

You will need to trim off the ends of your consensus, so that you only have the sections where the forward and reverse reads match. Click on trim consensus above. Then click on the last base at the beginning of your consensus read that you want to remove. Below I clicked on the last A that is unmatched and it shows everything from the beginning to that A as crossed off.



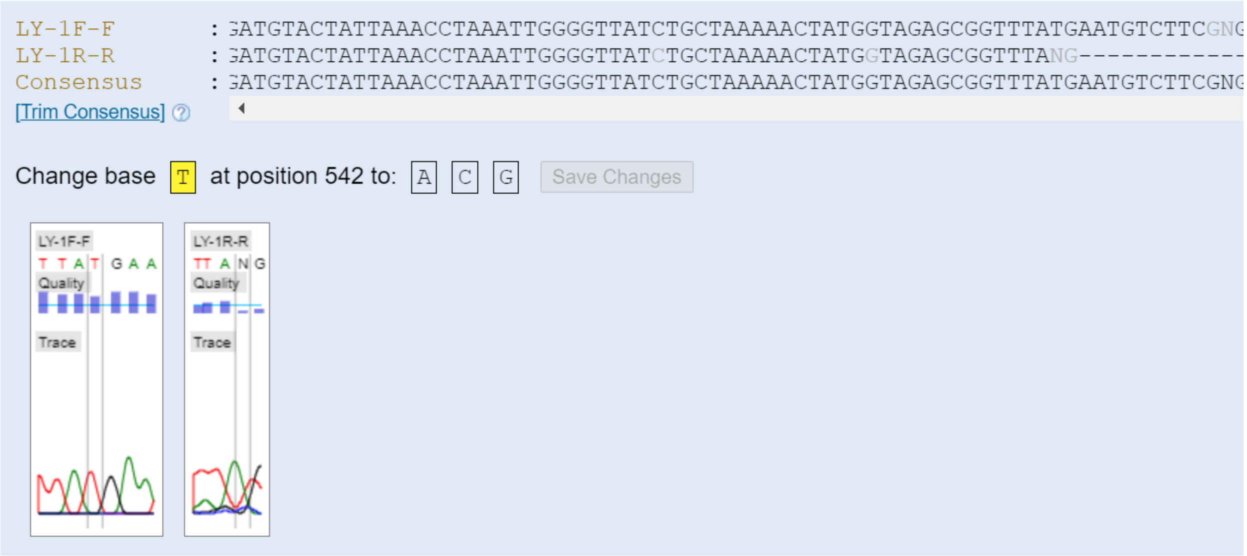
Pair **LY-1-sample-2.LS** [Edit Name]

```
LY-1F-F      : --------------------------------------GTGTTGGATTCAAAGCGGGTGTTAAA
LY-1R-R      : GACGGCCAGTATGTCACCACAAACAGAGACTAAAGCAAGTGTTGGATTCAAAGCGGGTGTTAAA
Consensus    : GACGGCCAGTATGTCACCACAAACAGAGACTAAAGCAAGTGTTGGATTCAAAGCGGGTGTTAAA
```

[Exit Trim Mode]

Consensus length: 615 bp
Left Trim: 38 bp [Reset]
Right Trim: 0 bp [Reset]

In general, remove any single red and any grey or other ones near the ends. If you want to check on the trace for any area, click on the base on the consensus read and it will bring up the trace for both reads for you to compare. Then scroll to the end and do the same for the end. See below where I had an N in the reverse read, but a T for the consensus. In this case, I'm going to cut the grey TG and the single read since it's right next to the single read.



```
LY-1F-F      : GATGTACTATTAAACCTAAATTGGGGTTATCTGCTAAAAACTATGGTAGAGCGGTTTATGAATGTCTTCGNG
LY-1R-R      : GATGTACTATTAAACCTAAATTGGGGTTATCTGCTAAAAACTATGGTAGAGCGGTTTANG-----------
Consensus    : GATGTACTATTAAACCTAAATTGGGGTTATCTGCTAAAAACTATGGTAGAGCGGTTTATGAATGTCTTCGNG
```

[Trim Consensus]

Change base T at position 542 to: A C G

You should look at all highlighted areas and potentially any grey areas to determine if you want to change any of the sequence. If in doubt, go with the computer suggested one. When done, click trim. Repeat for each sample pair you have. You are then done with the assemble sequences and are ready to do a BLASTN.

Add Sequences: This is where you will compare to the NCBI database to pick close and likely matches. Plus pick your outgroups to use for your cladogram.

BLASTN: view results for each sample you have. In general, you can pick anything and or everything, but here are some rules that can help you eliminate matches that are unlikely to give good return.

- Only pick it if it has the full species name

- Has more than 400 bp (longer is better)
- Bit score over 50
- E value of 0 (likelihood of match by random accident)
- Mismatches less than 2% of the length, but this could be high because yours isn't in the database.
- NOTE: this step allows you to choose them to compare, you can always remove them at later stages.
- Finally, be sure to click on add to your project at the top or bottom of this page.

Click on the 3rd one down: Reference Data. If you know what you want to use as an outgroup, pick that, if you're not sure you might want to pick several.

Analyze Sequences: For this first section I'm going to go through on how to do a species identification and build a cladogram from sequences in the database. If you have a project where you have many comparable samples then you might also include those in the same cladogram. An example of that might be comparing ants in an area to see if they're all the same or different species.

Step 1: Select Data. If you know your outgroups, then just pick a couple of species you want to use. If you're not sure I like to only select the outgroup for this first round and see what seems to be a close match to my sample.

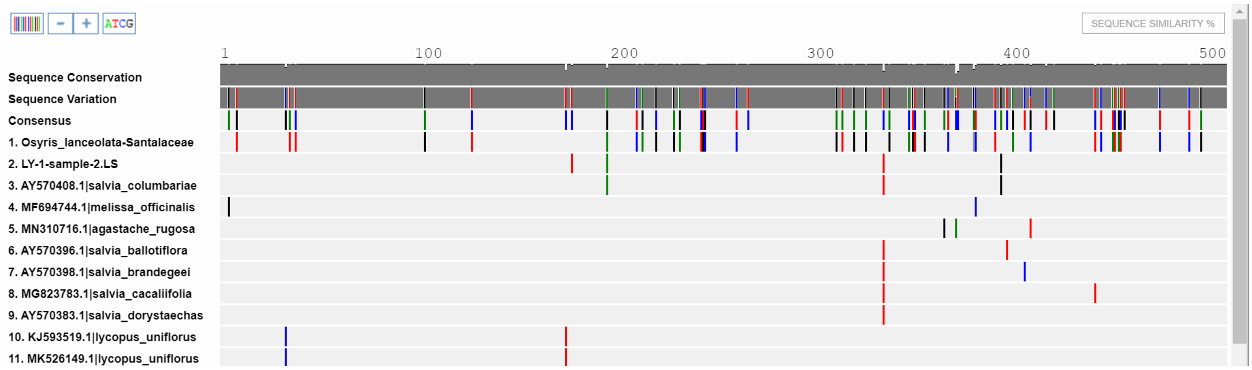
Step 2: MUSCLE. For now, I would just run this and go to the cladogram

Step 3: Build Cladogram. There are two choices here. PHYLIP ML is generally considered to be a better algorithm for building cladograms from sequences, but either can be used and compared. PHYLIP NJ (neighbor joining) gives you percents on likelihood it would split that way based on overall differences and is typically faster to run. PHYLIP ML (Maximum Likelihood) considers factors such as likelihood of one mutation occurring before another.

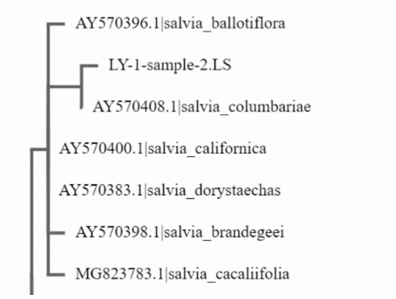
Since I'm not good with my plant species relatedness, I tend to work back and forth a bit on building my cladogram. In general, I kind of do the following:

- Only select outgroups to compare to my sample
    - Pick 1 that's kind of close and maybe keep 1-2 more that are really close. Set outgroup
- Go back to data select and remove all but the outgroup and any other species I want to compare with. Add in blast hits that have species data.
- Note, you must redo the muscle step each time you go back and change the select data or your cladogram will not change.
- Compare actual sequence alignment: after you run the muscle, click on it
    - Trim alignment to get rid of sections that you didn't sequence.
    - Essentially the colors you see are where that sample has a variation in the sequence that is different from the conserved sequence. At the very top you can see white notches in the grey, the white areas are where there is more variability in sequence.
    - + button will zoom in so you can see the alignment better.
    - If your unique sequence matches exactly to another it is likely the same species, but not a guarantee. This depends on many factors including: quality of your sequence, the length of your sequence, the number of species uploaded into the NCBI database.

- I find this is a good guide to determine which species can potentially be eliminated from my cladogram as you want to keep similar ones, but ones that have lots of differences maybe at most keep just one representative in a group to reduce the visual complexity of your cladogram.
- If your sequence is not an exact match, it may be a new species that is not in the database yet.
- I find using google searches on the species name helps me to figure out if some of these are likely to be good hits or not.


ATCG
SEQUENCE SIMILARITY %
1
100
200
300
400
500
Sequence Conservation
Sequence Variation
Consensus
1. Osyris_lanceolata-Santalaceae
2. LY-1-sample-2.LS
3. AY570408.1|salvia_columbariae
4. MF694744.1|melissa_officinalis
5. MN310716.1|agastache_rugosa
6. AY570396.1|salvia_ballotiflora
7. AY570398.1|salvia_brandegeei
8. MG823783.1|salvia_cacaliifolia
9. AY570383.1|salvia_dorystaechas
10. KJ593519.1|lycopus_uniflorus
11. MK526149.1|lycopus_uniflorus

This sequence alignment above shows that in the overlapping 500bp of sequence alignment of my sequence (#2 in the picture) with those that are close matches in the database, there are 4 differences from the conserved sequence. You can see the outgroup (#1) has the most bars or differences. #3 is the most similar to my sequence as three of the differences are the same, but mine has an additional T mutation in the ~180bp. If I click on sequence similarity, I can see that these two are 99.82% identical. This is good, because I had tentatively identified it as a Black Sage (*Salvia mellifera*) and this is the same genus. When I check the cladogram, I see many of these sages in close clades as you can see here.


AY570396.1|salvia_ballotiflora
LY-1-sample-2.LS
AY570408.1|salvia_columbariae
AY570400.1|salvia_californica
AY570383.1|salvia_dorystaechas
AY570398.1|salvia_brandegeei
MG823783.1|salvia_cacaliifolia

- At this point I will go back to the select data and start dropping off species that are repetitive, perhaps start dropping species that are in heavily populated clades that are kind of far from my species. I will typically also start googling species that are in the nearby/same clades to see if I think this seems reasonable.
- Once you finish eliminating species that are distracting or just too different to be helpful you will have both the sequence alignment in the muscle and a phylogenetic tree or cladogram of your barcoded species. Both of these can be used to give additional information to your original research question.
- If the student has good quality data and documentation of the original specimen, they can submit their sequence to be included in the NCBI database. This is the last step of the blue line