

Using DNA Subway to Analyze DNA Barcoding Sequences

INTRODUCTION

DNA Subway is an online educational bioinformatics platform. It bundles research-grade bioinformatics tools, high-performance computing, and databases into workflows with an easy-to-use interface.

In this exercise, we will show you how to determine an organism's taxonomic identity based on DNA sequence, using a collection of mosquito DNA sequences generated by students at James Madison University.

Mosquito-borne diseases

Mosquitos are one of the world's most deadly animals. It is the pathogens that mosquitoes carry which make them so deadly. According to the World Health Organization [1], mosquitos are responsible for spreading pathogens linked to millions of deaths every year, with 438,000 deaths worldwide caused by mosquitoes spreading malaria in 2015 alone. The most important vector (pathogen-carrying) mosquitos come from three taxonomic groups: *Aedes*, *Anopheles*, and *Culex*. Identification of a mosquito's genus is important because different mosquitoes are capable of carrying different diseases (see list below):

***Aedes*:**

- Chikungunya
- Dengue fever
- Lymphatic filariasis
- Rift Valley fever
- Yellow fever
- Zika

***Anopheles*:**

- Malaria
- Lymphatic filariasis

***Culex*:**

- Japanese encephalitis
- Lymphatic filariasis
- West Nile fever

Although some mosquitoes are most common in tropical and sub-tropical countries, as the climate changes, we see mosquitos that may have previously been reported in one region able to survive in new regions.

Like most insects, mosquitoes undergo a metamorphosis in which their adult form and juvenile form have very different appearances. While the adult forms are rather distinct, the larval forms look very

similar, and require some training to distinguish (Figure 1). However, since DNA barcodes do not change throughout insect life stages, we can use DNA barcoding [<https://www.dnalc.org/view/16979-DNA-Barcoding.html>] to identify samples taken from mosquito eggs, larvae, or adults to without the need for extensive training in mosquito morphology. In this lab, you will be provided with DNA sequences isolated from mosquito larvae collected from Virginia's Shenandoah Valley, and use DNA Subway tools to determine the taxonomic identity of these mosquito larvae. This knowledge will help determine whether members of disease-carrying groups of mosquitos (*Aedes*, *Anopheles*, or *Culex*) are present in the area.

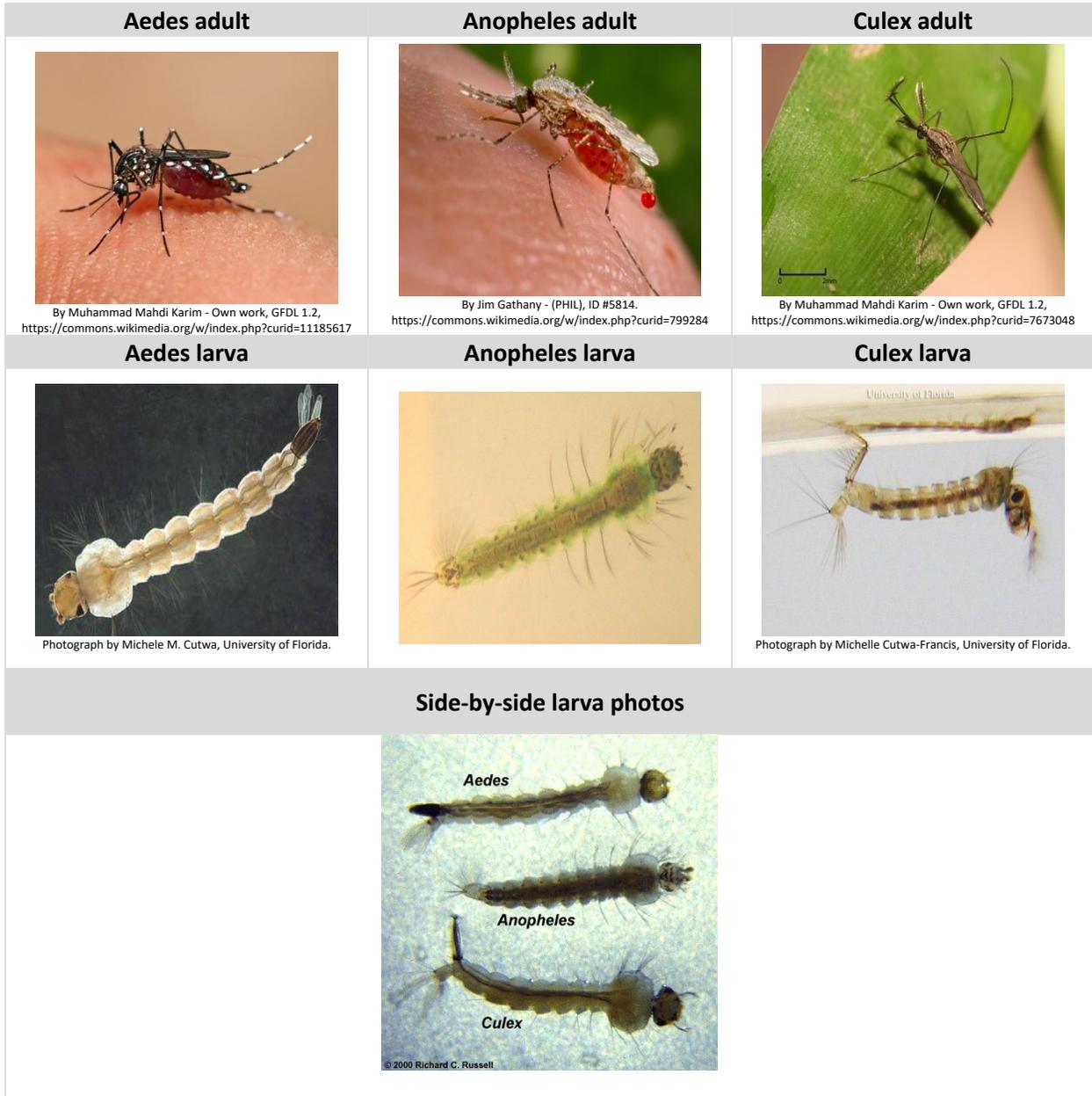


Figure 1. Adult and larval photos of the *Aedes*, *Anopheles*, and *Culex* mosquitoes.

To generate these samples, students collected mosquito larvae from ponds in Harrisonburg, VA public parks. DNA was extracted from these larvae, and a small region of each organism's DNA, called a barcode region, was copied millions of times using a process called **p**olymerase **c**hain **r**eaction (PCR) [<https://www.dnalc.org/view/15924-Making-many-copies-of-DNA.html>]. Finally, those millions of copies were used to determine the order of the DNA nucleotides in this barcode region through a process called DNA sequencing. Using these raw DNA barcode sequences from each sample, we will use DNA Subway to perform the following steps to generate a taxonomic identification for each mosquito larva:

- **Clean and assemble sequence results:** We will examine the quality of the DNA sequence from each sample and use two DNA sequences (forward and reverse) taken from the same sample to verify accurate sequencing results.
- **Identify similar sequences:** Using BLAST, we will search for similar (related) DNA sequences from a database of known sequences.
- **Determine sequence relationships:** We will add our DNA sequences to a multiple sequence alignment and then create a tree to visualize the relationships of our unknown organisms to each other as well as known organisms.

LABORATORY

Create a DNA Subway Blue Line Project: Determine Sequence Relationships

In this step, we will setup a new barcoding project. Since the mosquito samples are all animal material, we select a COI barcoding project type. COI (Cytochrome oxidase C subunit I) is the gene sequence commonly used for animal DNA barcoding. Scientists have selected COI as a barcoding gene because it is present in most animals. Additionally, COI has a relatively high rate of mutation and as a result, animals in different genera will likely have a unique COI sequence [2].

1. Log-in to [DNA Subway](#); unregistered users may ‘Enter as Guest’ but your work will not be saved.
2. Click the blue square (“Determine Sequence Relationships”) on the left side of the screen to start a Blue Line project.
3. Under “Select Project Type.” Select **COI** from the “Barcoding” column.
4. Under “Select Sequence Source” choose **Intro to Barcoding Bioinformatics: Mosquitos** from “Select a set of sample sequences”.
5. Give the project a title and, if desired a description. Click ‘Continue’ to complete the project setup.

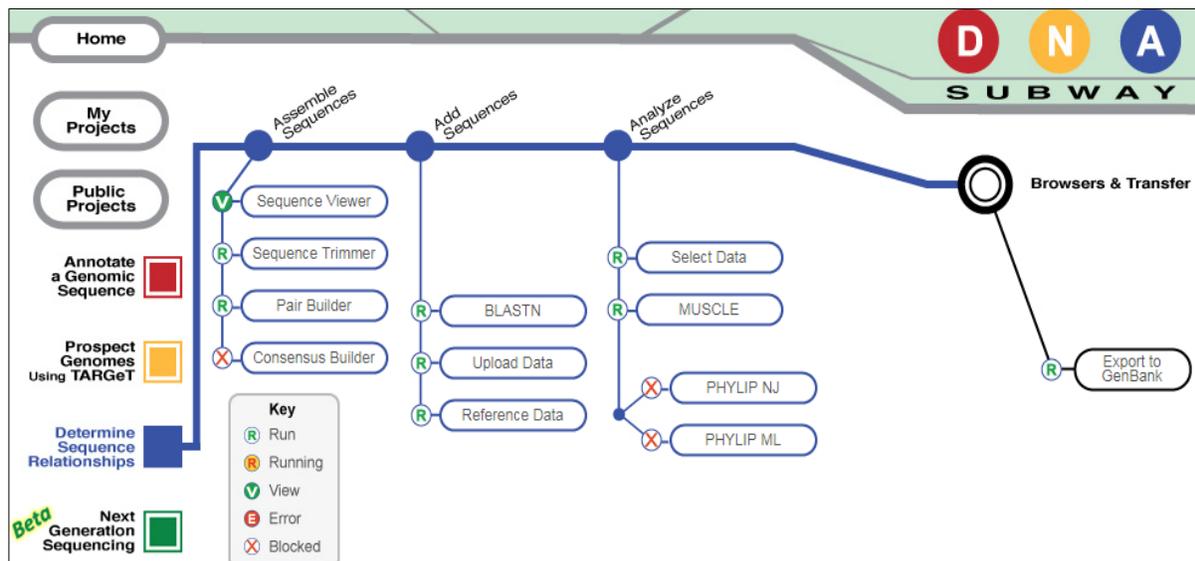


Figure 2. DNA Subway Blue Line. Note that DNA Subway gives us an outline of the procedure we will follow. In this laboratory exercise, Roman numerals will correspond to each Subway ‘Stop’ on the main Blue line, while letters will correspond to stops on the side lines shown. Take a moment to look at the figure for an overview of the process.

I. Assemble Sequences: Clean and assemble sequence results

Results obtained from DNA sequencing [<https://www.dnalc.org/resources/animations/cycseq.html>] are not perfect. If they were, we could skip this step and use DNA sequencing data directly in our analysis. In reality, DNA sequencing can include error at several steps:

- **Human error:** Samples we send to a sequencing facility may have been mislabeled, or they may be mishandled or mixed at the sequencing facility.
- **Systematic error:** DNA sequences will often be unreliable at the very beginning or end of the sequence read.
- **Signal to Noise:** At any given position in a DNA sequence the amount of signal (fluorescently labeled nucleotide) may be too low to distinguish from background, leading to one or more “miscalled” bases.
- **Electrophoresis:** In modern Sanger sequencing machines, electrophoresis is used to move individual bands of DNA through a capillary. These fragments of DNA migrate based on their size. However, several factors can cause the flow of DNA to be non-uniform. As a result, bands may not be well separated, leading to one or more miscalled bases.

All experimental observations contain error. In this section, we will determine the quality of the sequence data we obtained. Additionally, we have sequenced each sample twice – once in the forward direction (capturing the 5'-3' strand) and in the reverse direction (capturing the 3'-5' strand). These complementary sequences should be identical, so differences between the two sequences indicate some type of error. We will build a consensus sequence from each of the two-sequence pairs to use in our analysis.

A. Sequence Viewer

In this section, we will view the DNA sequences, both in text format (i.e. a sequence of nucleotides) and as an electropherogram.

- The DNA sequencing software measures the fluorescence emitted in each of four channels – A, T, C, G – and records these as a trace, or electropherogram (see Fig. 3 below). In a good sequencing reaction, the nucleotide at a given position will be fluorescently labeled far in excess of background (random) labeling of the other three nucleotides, producing a “peak” at that position in the trace. Thus, peaks in the electropherogram correlate to nucleotide positions in the DNA sequence.
- A software program called Phred analyzes the sequence file and “calls” a nucleotide (A, T, C, G) for each peak. If two or more nucleotides have relatively strong signals at the same position, the software calls an “N” (i.e. aNy) for an undetermined nucleotide.
- Phred also examines the peaks around each call and assigns a quality score for each nucleotide. The quality scores correspond to a logarithmic error probability that the nucleotide call is wrong, or, conversely, to the accuracy of the call (Table 1).

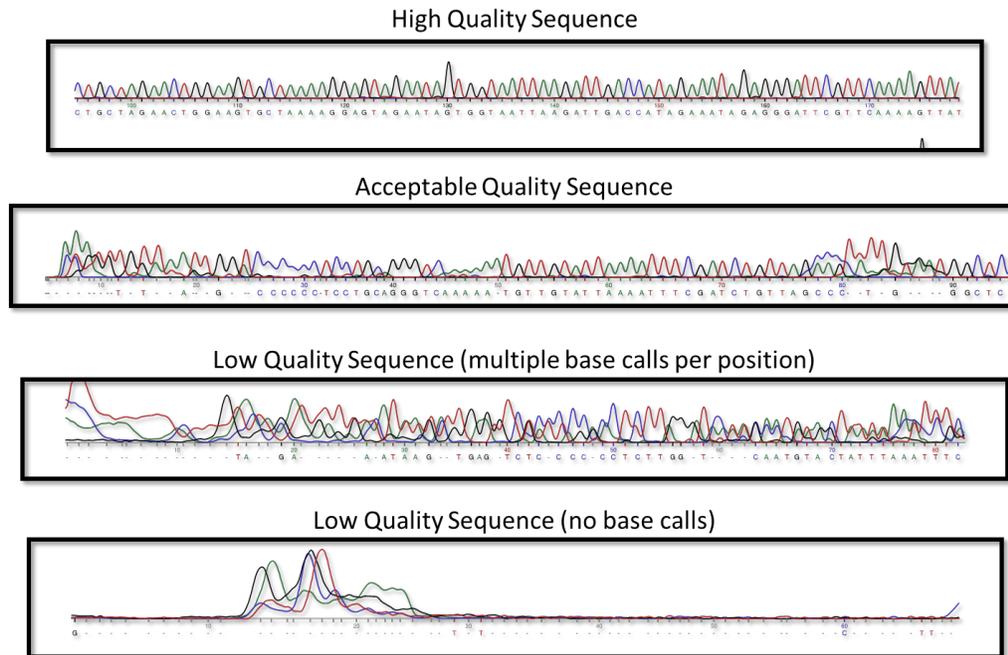


Figure 3. Example trace files: A high quality sequence is characterized by distinct, non-overlapping peaks. A poor quality sequence may contain lots of overlapping sequences, low or no distinct peaks, or both of these issues.

Table 1. Phred Scores

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

1. Click "Sequence Viewer" to display the sequences you have input in the project creation section. As you scroll (left to right) you will see the DNA sequence obtained for each student sample.
2. Click on a sequence name (e.g. mosquito-1F). This will display the electropherogram for that DNA sequence (Fig. 4)

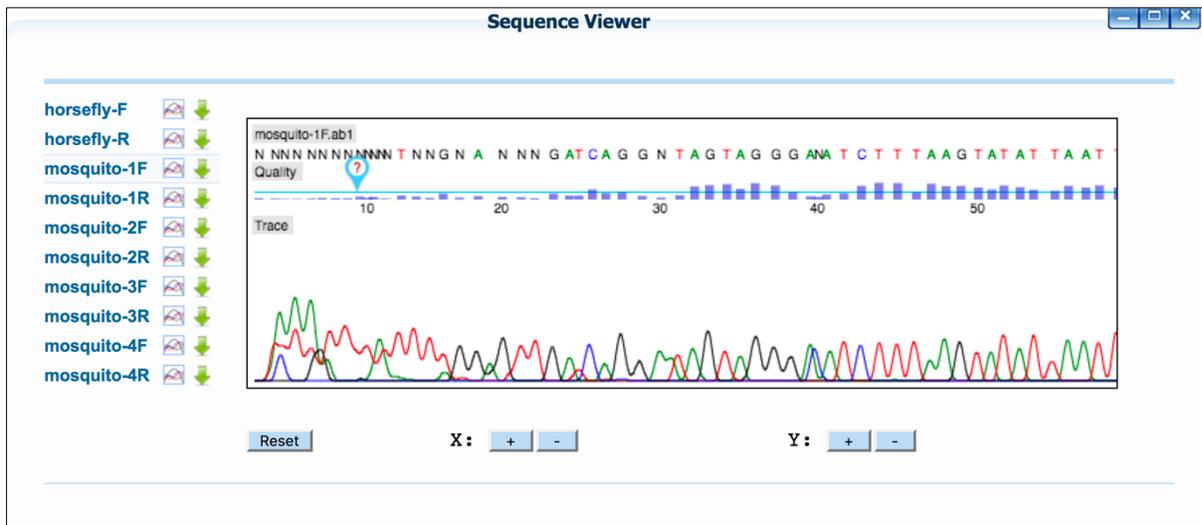


Figure 4. DNA Subway electropherogram. DNA Sequence (as measured by fluorescence) is shown as colored peaks at the bottom of the readout. The quality (corresponding to Phred score) is shown as a histogram (blue bars at the top). The thin blue line in the histogram represents a Phred score of 20. Note the N's at the beginning of the sequence. In this example, although the software calls most of the bases after base 30, we may want to trim off bases up to where the Phred score stabilizes at base 43.

- The electropherogram viewer represents each Phred score as a blue bar. The horizontal line equals a Phred score of 20, which is generally the cut-off for acceptable sequence; any bar at or above the line (30 or higher) is considered a high-quality read.
- Even high-quality sequence reads will begin with Ns. In "clean" sequences, where experimental conditions were near optimal, the initial Ns will end within the first 25 nucleotides. The remaining sequence will have very few, if any, internal Ns. Then, at the end of the read the sequence will abruptly change over to Ns as the end of the DNA molecule being sequenced is reached.
- Large numbers of Ns scattered throughout the sequence indicate poor quality sequence. Sequences with average Phred scores below 20 will be flagged with a "Low Quality Score Alert." You will need to be careful when drawing conclusions from analyses made with poor quality sequence.
- **Note:** An exclamation icon (!) next to the sequence name indicates poor quality sequence.

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

Examine the quality of the sequence(s), scrolling all the way to the end of each one. Any sequence that has the warning icon should not be used in your analysis. Bases called incorrectly in poor quality sequence could appear to be novel mutations, which could result in faulty taxonomic identifications. Make a note of any poor-quality sequences. After you have examined your sequences, close the "Sequence Viewer" window.

B. Sequence Trimmer and C. Pair Builder

In this section, we will remove leading and trailing “N” calls from the sequence data. We will also pair complementary sequences taken from the same sample. Because the sequence ends will not overlap completely, it may be difficult to see the match until you get to the next step (D. Consensus Builder)'.

1. Click “Sequence Trimmer”; when the run is completed, a green “V” will appear. Click the “Sequence Trimmer” again to view the results as in section A above.
 - “Sequence Trimmer” takes the average Phred score of several nucleotides at a time, moving along the sequence from left to right. When the number of N’s drops, the leading sequence is trimmed. This removes Ns at the beginning of the sequence but does not remove Ns that appear in the middle or end of the sequence.
2. Click “Pair Builder.’
3. Click the “Try Auto-pairing” link to pair sequences from the same DNA sample. The mosquito-1**F** read corresponds to the forward read of the mosquito-1 sample and mosquito-1**R read** corresponds to the reverse read of the same mosquito-1 sample.
 - By default, *DNA Subway* assumes that all reads are in the forward orientation and displays an "F" to the right of the sequence. If any sequence is not in that orientation, click the "F" to reverse complement the sequence. The sequence will display an "R" to indicate the change.
 - Reverse complementing involves reversing the order of the reverse read and then changing the bases to their complementary bases. In this way, the two sequences should be identical, and should (mostly) overlap.
4. Click “Save” to save the newly created pairs.

Section Questions

- **Why is it important to remove excess Ns from the end of the sequences?**

- **Will a sequence and its reverse complement sequence completely overlap?**

D. Consensus Builder

In this section, we will combine the “cleaned” data from the forward and reverse reads. We will also remove poor quality areas at the 5’ and/or 3’ ends of the consensus sequence.

1. Click “Consensus Builder”; when the run is completed, a green “V” will appear. Click the “Consensus Builder” again to view the results. One sequence will now be displayed for each of the pairs.
2. Click on a sequence name to view the two sequences that were used to generate the consensus sequence.
 - The consensus sequence is the “sum” of both the forward and reverse read. Where the reads overlap, the highest quality read is selected.
 - A large number (> 4 or 5) of mismatches in properly paired and reverse complemented sequences indicate that one or both sequences is of poor quality. Often, one of the sequencing reactions produces a high quality read (30 or greater) that can be used on its own. Recall that clicking on a nucleotide shows the electropherogram at that position, a blue line indicates the Phred score of 30 as a cutoff.

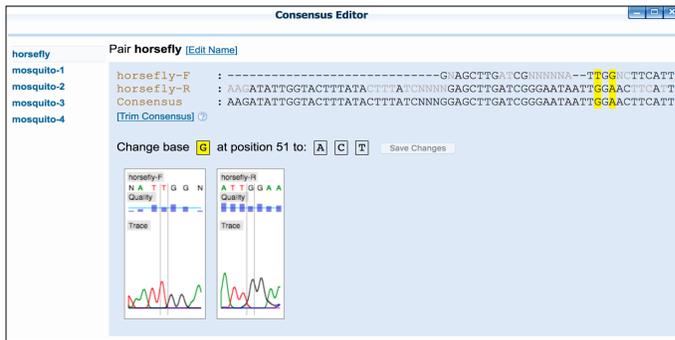


Figure 5. DNA Subway Consensus Editor. Where reads conflict, the nucleotides will be highlighted in yellow; clicking on a mismatched nucleotide will also display the electropherogram at that location.

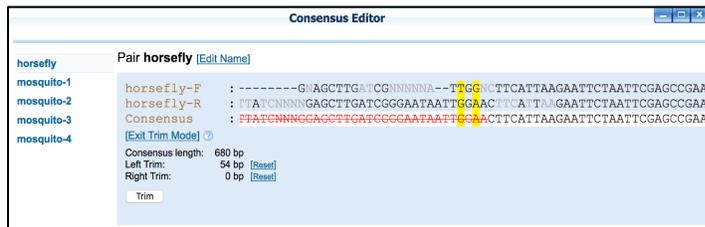


Figure 6. DNA Subway Consensus Editor in Trim Mode. A red line will indicate the sequences to be trimmed. Lower quality sequences will appear in grey font. Although they may not appear as a mismatch, you may wish to trim these as well, especially since “N” bases will not be flagged as mismatches.

3. A consensus sequence should be trimmed so that it includes only the region of overlap between the forward and reverse read. To trim, click the “Trim Consensus” link beneath the sequence names. On the consensus sequence, click on the last nucleotide you wish to remove from the beginning of the sequence. Repeat this procedure on the end of the sequence. Also note, when viewing a sequence in the Consensus Editor, you can click on a sequence and click the ‘Edit Name’ link to rename a sequence if desired.
4. After you have selected the regions to trim, click the “Trim” button. Repeat this procedure for all your sequences, filling out Table 2 below as you proceed. Try to be consistent with your trimming -- don’t adopt a conservative policy for some sequences, and a liberal policy for others.
 - A dash (–) is used to represent a gap in the data. In our consensus editor, the dash is used to “pad” the alignment between the forward and reverse sequences. A dash is a useful feature in an alignment because one of the possible mutations that could differentiate two related sequences is an insertion or deletion. In our case, misalignments between a forward and reverse read from the same sample are due to sequencing error. Since they are sequences from the same sample, they should be identical.
 - One recommendation on trimming at the beginning of the sequence is to trim up to the last position where one sequence has an ‘N’ or a ‘–’ within the first 50 or so bases. Starting from the right, you can also trim the sequence starting at the first ‘N’ or ‘–’ you find 75 or so base pairs from the end of the read. These recommendations are only rules of thumb. You will have to choose how strictly you wish to trim. If a trim on either end is more than 100bp, you may have to consider the effects of discarding large amounts of sequence. Trimming 200 bp in total represents 1/3 of our approximately 600 bp of sequence.

Section Questions

- For each of the consensus sequences, complete the following chart to indicate how you trimmed the sequence. The horsefly is included to act as an 'outgroup' – a species that is less closely related than the other sequences (see 'More on Phylogenetic Trees' below).

Table 2: Sequence trim information

Sequence name	Left trim bp removed	Right trim bp removed	Consensus length (after trimming)	Notes
<i>horsefly</i>				
<i>mosquito-1</i>				
<i>mosquito-2</i>				
<i>mosquito-3</i>				
<i>mosquito-4</i>				

II. Add Sequences: Search for matching sequences in a database

A. BLAST DNA Sequences

A BLAST search can quickly identify any close matches to your sequence in sequence databases. In this way, you can often quickly identify an unknown sample to the genus or species level. BLAST also provides a means to add samples to DNA subway for a phylogenetic analysis. In this exercise, we will use a version of BLAST called BLASTN, which allows us to blast a DNA nucleotide sequence against an online database of DNA nucleotide sequence.

1. Click “BLASTN”; next to the first DNA sequence, click the BLAST link. When the BLASTN search is complete (this may take up to a minute) you will be presented with a results table.
2. The BLASTN results table contains detailed information about sequences (matches) that are similar to your DNA sequence (query). Click on the species name* for the first match in the list to learn more information about the species. Links to Wikipedia/Encyclopedia of Life may also contain photos of this species.
3. In the Accession column next to each match, there is a checkbox; click to select the top match, then click the “Add BLAST hits to project” button. Add the species names of these matches to Table 3 below. You may see multiple species that match your query; read the ‘More About BLAST’ box for information about how to proceed in this case.
4. Repeat the BLASTN procedure for each DNA sequence and answer the questions below.

*Note: In some cases, your BLASTN will not return a species name, but a genus name, or an indication that a result is a member of a genus such as (*Homo Sp.*). In this case, click on any available link to learn more, and record results for the genus rather than a species.

Section Questions

- For each of the consensus sequences, complete the following chart

Table 3. Hit names

Sequence name	Genus/species name of top hit	No. of Mismatches	Alignment (ALN) length (bp)	% of Mismatches (No mismatches/Aln. Length)
<i>horsefly</i>				
<i>mosquito-1</i>				
<i>mosquito-2</i>				
<i>mosquito-3</i>				
<i>mosquito-4</i>				

- Do you notice hits that have the same statistics (e.g. alignment length, bit score, e-value, mismatches) but originate from different species? What does this tell you about the DNA Barcoding technique? (hint: read the 'More about BLAST' box).

More about BLAST

BLAST (Basic Local Alignment Search Tool) is an algorithm to search for similarities between DNA or Protein Sequences. BLASTN specifically is used to compare a query nucleotide sequence and search for close matches in a database of known sequences. When BLAST finds a similar sequence, it returns that sequence along with statistical information that help you to judge the significance of the match (i.e. is there a match because these sequences share a biological relationship or is the match more likely due to random chance).

Here are some important definitions about the information generated by the DNA Subway BLASTN search:

- **Accession number**, a unique identifier given to each sequence submitted to a database. Prefixes indicate the database name – including gb (GenBank), emb (European Molecular Biology Laboratory), and dbj (DNA Databank of Japan).

- **Organism and sequence description or gene name** of the hit. Click the genus and species name for a link to an image of the organism, with additional links to detailed descriptions at Wikipedia and Encyclopedia of Life (EOL).
- Several statistics allow comparison of hits across different searches. The number of mismatches over the length of the alignment gives a rough idea of how closely two sequences match. The **bit score** formula takes into account gaps in the sequence; the higher the score the better the alignment. The **Expectation or E-value** is the number of alignments with the query sequence that would be expected to occur by chance in the database. The lower the E-value, the higher the probability that the hit is related to the query. For example, an E-value of 0 means that a match this good (or better) would be expected to turn up 0 times by chance in a search of a database of the size you searched. BLAST results are ranked so that matches with the highest alignment length and bit score as well as the lowest e-value and mismatch number are at the top of the list (the best matches). In some cases, multiple species will have identical BLAST results – it is in these cases we see that BLAST is limited in differentiating very closely related sequences (and by extension, species).
- Examine the last column in the report called “**Mismatches.**” For barcodes, this is the informative column, with the best hits being those with the lowest number of mismatches. Note that hits with low numbers of mismatches can sometimes be lower on the list, as the bit scores are used to arrange the hits in the table. High bit scores can occur when the alignment length is longer, even when there are more mismatches than for other hits.

More about novel sequences and the limitations of BLAST

If there are zero mismatches between your sequence and a BLAST result, the sequence probably matches because it is in the same taxonomic group as your sample. Check to see if any matching sequences are from species that seem reasonable for your sample. If your sequence matches that of multiple related species, you may only be able to identify your sample as being a member of a particular genus, rather than a species. In this case, more than one barcoding locus may be needed to identify your sample at the species level.

If your best matches include some mismatches, you may have identified a novel barcode -- a species that is not yet in the database! The more mismatches you find, especially in regions of the sequence with high quality scores, the more likely that your sequence is unique. However, sequencing errors could explain the difference, so it will be important to reexamine the trace files at any sites with mismatches to ensure that the consensus at those locations is of high quality.

It is also important to note that BLAST results alone cannot conclusively determine the identity of your species. Using a single DNA Barcoding locus (See references here: <http://www.dnabarcoding101.org/lab/index.html>) **we can only expect to identify the genus of an unknown organism.** In other words, in closely related species, a single DNA barcode may not reveal any mutations that differentiate these species. To overcome this limitation, it may be necessary to use more than one DNA barcoding locus.

B. Upload Data

In this exercise, we don't have a sequence representing the *Anopheles* genus. We can import this sample from GenBank, an online database of sequences.

1. Click "Upload Data"; under "Import Sequence" select "Import sequence from Gen Bank." Paste in the accession **KM592986.1** and then click "Import." This will import a COI sequence from *Anopheles neomaculipalus*.

C. Add Reference Data

Now that we have identified our sequences, we will add some additional species to our analysis to use for comparison. These common insects contain representatives from many of the large insect orders. This additional step allows us to reconstruct evolutionary relationships among species and help us confirm that we have identified species correctly. It is important to note that genomes do not evolve uniformly, that is some genes have higher mutation rates than others. The relationships that we reconstruct in this exercise are based only on analysis of the COI gene. We may estimate different relationships using other genes.

1. Click "Reference Data" and select "Common insects." Next click "Add ref data".

Note: This will add a selection of COI sequences from the following species:

Species (Common name)	Genus
American Cockroach	Blattaria
American Grasshopper	Orthoptera
Caddisfly	Trichoptera
Cicada	Hemiptera
Dragonfly	Odonata
Fruit Fly	Diptera
Honey Bee	Hymenoptera
Lacewing	Neuroptera
Mayfly	Ephemeroptera
Monarch Butterfly	Lepidoptera
Praying Mantis	Mantodea
Scorpionfly	Mecoptera
Stag Beetle	Coleoptera
Stonefly	Plecoptera
Termite	Isoptera

IV. Analyze Sequences: Compare sequences using a multiple sequence alignment and phylogenetic tree

In this section, we will assemble and align all of our collected sequences. A multiple sequence alignment allows you to compare several DNA sequences at once, visualizing similarities and differences between the sequences. From this information, we will construct a phylogenetic tree, a hypothesis about how these sequences are related.

Note: In section IV, we will be repeating steps in stops A (select data), B (align with Muscle), and C (Create NJ Phylogenetic tree with Phylip). These steps can be repeated with different data selections to draw and examine a variety of possible phylogenetic trees.

A. Select Data

1. Click “Select Data” and select user data (Horsefly, and mosquito samples 1-4); then click “Save Selections” to move those sequences to the next steps.

B. Align with MUSCLE: Generating a multiple sequence alignment

1. Click “MUSCLE” to generate the multiple sequence alignment. When the “V” appears, click MUSCLE again to view the alignment results. MUSCLE is a software tool that takes several DNA sequences and repositions them (adding gaps where necessary) to generate a multiple sequence alignment (the alignment of 3 or more DNA sequences). Assuming the DNA sequences share a common origin, alignments of DNA sequence can reveal mutations between different sequences, including insertions, deletions, or single nucleotide polymorphisms (SNPs).



Figure 7. DNA Subway Alignment Viewer. *Sequence Conservation* displays a histogram across the displayed sequences. At positions where most nucleotides are the same, the histogram approaches 100%; dips in the histogram are more variable regions. *Sequence Variation* displays the nucleotides that occur at that position relative to the consensus sequence; the colors of the bars (Green = A, Red = T, Black =G, Blue =C) display what alternative nucleotide(s) appear at that position (e.g. a 100% blue bar, would mean one or more nucleotides at that position are not identical to the consensus nucleotide, and the alternative nucleotide is a C; a 50% blue, 50% black bar would indicate the alternative nucleotide is either a G or a C). The *Consensus* bar is light gray in positions where all sequences shown contain the same nucleotide; colors appear to indicate any nucleotide position that is not at 100% consensus with other aligned sequences. Missing sequence is indicated in each row by a dark grey block on either end of the sequence. Light gray spaces indicate agreement with the consensus sequence.

How to interpret the multiple alignment

A consensus sequence shows the most common nucleotide at a given position along the length of the match. Using the (+) button or the (ATCG) button to zoom down to the nucleotide level, the entire consensus sequence is shown, and nucleotides in the rows of sequences are only shown if they differ from the consensus. At this point, it should be possible to see clear patterns of sequences that share mutations (polymorphisms). You can also identify conserved regions, where there are no differences among sequences.

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

2. Click the “TRIM ALIGNMENT” button to trim the alignment to a region where all the selected DNA sequences overlap. Without this trimming step, those missing regions of sequences would be interpreted by the phylogenetic tree building algorithm as true deletions in the sequence, rather than missing data. In this step, we are trimming again to account for the BLASTN matches we introduced, which may have different sequence lengths from the data we generated.

Section Questions

- **Looking at the sequence conservation histogram, are some regions in the barcode more conserved or less conserved than other regions?**

C. Create a Neighbor-Joining phylogenetic tree

In this section, we will use a software tool to build a phylogenetic tree. This particular software (Phylip NJ) uses a strategy called ‘Neighbor joining’ builds trees by placing the most similar sequences close together on the tree.’ A Phylogenetic tree is a hypothesis about how sequences are related. Although a phylogenetic tree may look like a family tree, the analogy is incomplete, since in a family tree we (usually) know for certain what the relationships between members are. In a phylogenetic tree, we are using only some of the available evidence – in this case DNA sequence from a single barcoding region – to make an inference about how species may be related.

1. Click “PHYLIP NJ” to construct your phylogenetic tree. When the “V” appears, click PHYLIP NJ again to view the alignment results.
2. For the “Select Outgroup” function, select “horsefly”
 - In this tree, we have selected an outgroup – an assumption that one species, in this case horsefly, is the least closely related species in our selection of sequences.

More on Phylogenetic Trees

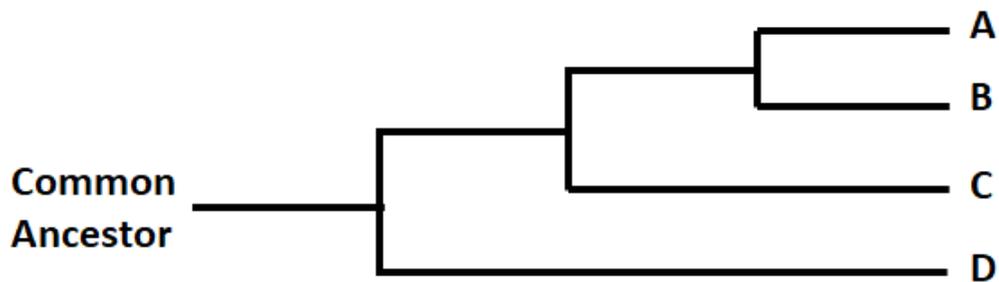
Phylogenetic trees can display inferred relationships between organisms by comparing changes in DNA or protein sequences and grouping collections of DNA or protein sequences by similarity. Often, changes in DNA or protein sequences are used to infer the overall relationships between organisms.

Trees consist of branch tips (also called leaves) that are labeled with the name of the sequence and/or organism, and nodes, which represent the point at which descendants from an inferred common ancestor diverged into different lineages.

An analysis of 10 sequences (perhaps from 10 different organisms) will result in a tree with 10 leaves. The number of nodes depends on how many hypothetical common ancestors can be inferred to exist between the sequences. Just how many common ancestors (nodes) may have existed, and what justifies hypothesizing a common ancestor (i.e. creating a node). The final tree will vary according to software and settings used.

Relatedness between any two sequences decreases as the number of nodes between them increases; sequences sharing only one node are more closely related than sequences that can only be joined by two nodes. If your analysis consists of sequences from different organisms, nodes may represent distinct taxonomic units (such as families or species).

In the example tree below, species A and B are the most closely related species shown. Species C is more distantly related to species A and B, and species D is the most distantly related species shown. In order to reach these conclusions, we must know that the tree under consideration is a 'rooted tree', meaning that we can assume that the last common ancestor from which all of these species are descended is positioned as shown. In order to create a rooted tree, we must have independent information that tells us that one of our species is an 'out-group' that is less related to the others in the tree. In this tree, species D is the out-group; in the first mosquito tree that we construct, the out-group is the horse fly.

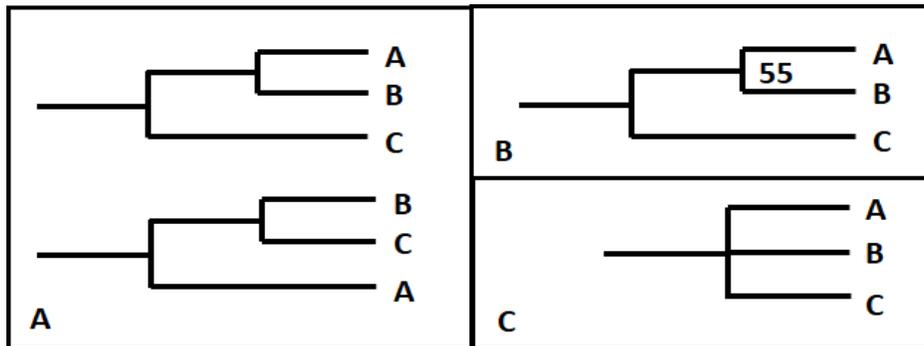


The NJ method is a "distance-based" method of tree building. First a matrix of the sequences is created and the distance between each pair of sequences in a multiple alignment is calculated; the distance is related to the number of mismatches between the sequences. To construct a tree, pairs of sequences are connected by nodes starting with the most closely related sequences. Inside the MUSCLE window you can click the "Sequence Similarity %" button in the upper right-hand corner of the alignment window to see a similarity matrix.

To find the most likely tree and determine the reliability of the branches in this tree, NJ in DNA Subway uses bootstrapping, or resampling of the sequence data. Bootstrapping is a computational technique for

assessing the accuracy of a statistical estimate. In bootstrapping, the columns in the sequence alignment are randomly resampled over and over to make many new alignments – 100 for NJ in DNA Subway – and these alignments are used to construct NJ trees. The final tree represents the “most likely” tree and shows the confidence of relationships with bootstrap levels. Each bootstrap value is the number of times that particular relationship appears in the 100 resampled trees. The values do not represent the distance between sequences. Instead, a higher value indicates that a branch of the tree is well supported, while low values indicate that the relationships are less certain. In general bootstrap values above 70 might be considered as plausible given the data, and above 95 can be considered “correct.”

For example, suppose that species A, B, and C are all somewhat related to each other. It may be unclear which of the two trees shown in panel A below is more correct. Let’s say that when a multiple sequence alignment is sampled, 55 out of 100 of the resulting trees show species A and B more closely related to each other than they are to species C. This situation would be represented as shown in panel B below. This is how DNA Subway will represent trees, to show how strong the support is for various branches. However, when scientists feel that the evidence for one pair of species being more closely related than another is this weak, they may prefer to leave out nodes with weak support, and instead show a node with more than two branches (panel C).



For more on interpreting phylogenetic trees see:

- Teaching the Process of Molecular Phylogeny and Systematics: A Multi-Part Inquiry-Based Exercise <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2995769/pdf/cbe513.pdf>
- Teaching Tree-Thinking to Undergraduate Biology Students <http://evolution.berkeley.edu/UToL/meisel10.pdf>
- Understanding Evolution: The Tree Room. https://evolution.berkeley.edu/evolibrary/article/0_0_0/evotrees_intro

Interpreting first tree results

Depending on how you trimmed your sequences in the “Consensus Editor” step, you should obtain a tree that looks like this:

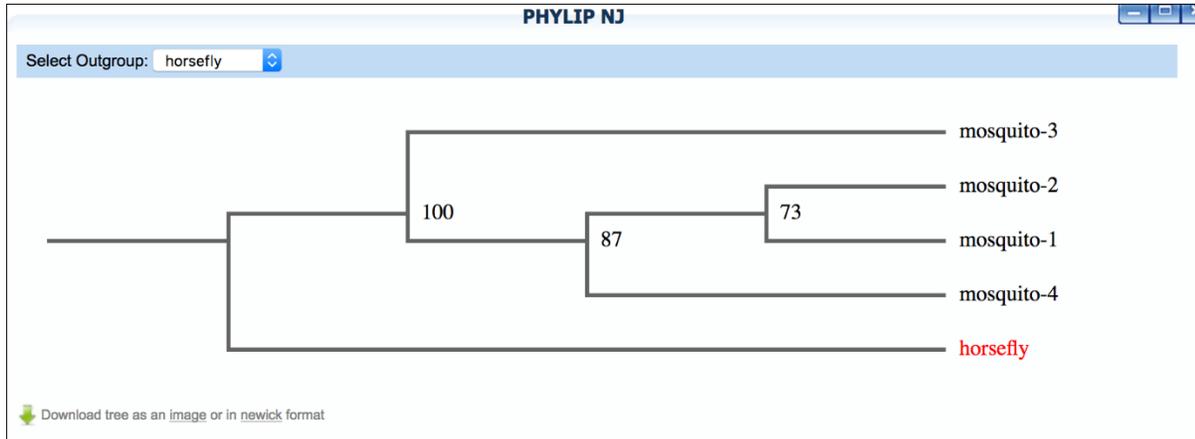


Figure 8. Phylogenetic tree of mosquito and horsefly sequences

In this tree, we see our four mosquitos, and the horsefly as an outgroup (a horse fly is in the same order – Diptera – as mosquitos, but is in a different family). The numbers on the nodes of the tree are bootstrap values and can be interpreted to represent how frequently this branch appears when the data used to create the tree (the multiple sequence alignment) are resampled. For example, if you used different samples (columns of matched sequences) from the multiple alignment sequence data to make this tree 100 times, the branch that shows mosquito-2 and mosquito-1 to be the most closely related sequences in the tree relationship between mosquito 2 and 3 would appear 73/100 times. A bootstrap value of more than 70 is considered reasonably supported. In this tree, the mosquito 1 and 2 are more closely related to each other than any other taxa (mosquito or horsefly). Mosquito 4 appears here to be more closely related to the mosquito 1/2 group, than to mosquito-3. Mosquito 3 is still more closely related to mosquito-4 than to the horsefly. The grouping that separates the mosquitos from the horsefly appears 100/100 times in our bootstrap resampling. This is not surprising given how distantly related horseflies are from mosquitos.

3. Go back to the beginning of **Section IV** (Analyze Sequences: Compare sequences using a multiple alignment and phylogenetic tree) and repeat the instructions under section A step 1, except for “Select Data”, select all the user samples (Horsefly, and mosquito samples 1-4), as well as the 5 BLAST hits (one for each mosquito and the horsefly), and the Genbank *Anopheles* sample. Then click “Save Selections” to move those sequences to the next steps. Calculate the next phylogenetic tree following **Section IV: B. Align with MUSCLE: Generating a multiple sequence alignment, C. Create a Neighbor-Joining phylogenetic tree (steps 1-2). Again, select horsefly as the outgroup.**
4. Repeat the tree building procedure this time adding the reference data samples. Use Horsefly as the outgroup.

Section Questions

- How would you describe the relationships in your tree (i.e. what two species are the most closely related, followed by next most closely related species, etc.)?
- According to one reference [3], the *Culex* genus is very ancient – it appears as much as 150 million years before *Anopheles* and *Aedes*. It would not be surprising for such an ancient genus to be quite diverse. Does this information help to explain how the *Culex*-related sequences appear in your tree?
- How should you view relationships with low bootstrap support (i.e. bootstrap value < 70)?
- When we add one of the insects in the reference data group, we see several of them that have the word fly in their name (e.g. Mayfly, Scorpionfly, etc.). True flies are members of the order Diptera. Are all the species with the term “fly” in their common name members of Diptera based on your tree?
- Given the BLAST results and the phylogenetic tree, what identifications do you propose for the 4 mosquito samples?
- Given your answer to the previous question, and the list of diseases vectored by each mosquito genus from the intro to this exercise, which of the following diseases should citizens near the location where these mosquitos were collected NOT be concerned about: Zika, West Nile, or Malaria? Explain your answer:

Conclusion

In this exercise, the mosquito sequences were analyzed, first by cleaning these sequence data, using BLASTN to search a known database of sequences for matches, and by placing these sequences in the phylogenetic contexts of known organisms. DNA Barcoding is a powerful technique for identifying organisms. However, DNA Barcoding alone may only be once piece of evidence needed to establish the identity of an unknown sequence. In general, barcoding at a single locus (such as COI) will not yield a conclusive identification for the species but should establish the genus the sample belongs to. With additional barcode sequences, it may be possible to have enough evidence for a conclusive identification. For definitive identification, DNA Barcoding, along with traditional taxonomic characters such as the shape/size/structure of the organism and its various components are used for definitive identification. Finally, using DNA sequences to create a phylogenetic tree can provide valuable insights into species relationships.

Further work

Visit the DNA Barcoding 101 website (<http://www.dnabarcoding101.org/>) for more information on DNA Barcoding, including how to use DNA Subway to analyze your own DNA sequences and even contribute them to the global online databases for DNA sequencing and barcoding.

References

1. "Vector-Borne Diseases." *World Health Organization*, 1 Oct. 2017, www.who.int/mediacentre/factsheets/fs387/en/.
2. Hebert PDN, Ratnasingham S, DeWaard JR. Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. *Proc R Soc B Biol Sci* 2003;270(SUPPL. 1):S96-9.
3. Reidenbach KR, Cook S, Bertone MA, Wiegmann BM, Besansky NJ. Phylogenetic analysis and temporal diversification of mosquitoes (Diptera: Culicidae) based on nuclear genes and morphology. *BMC Evol Biol*. 2009; 9: 298. Published online 2009 Dec 22.