9 Analyzing DNA Sequences and DNA Barcoding

Introduction

DNA sequencing is performed by scientists in many different fields of biology. Many bioinformatics programs are used during the process of analyzing DNA sequences. In this lesson, students learn how to analyze DNA sequence data from **chromatograms** using the bioinformatics tools **FinchTV** and **BLAST**. Using data generated by students in class or data supplied by the Bio-ITEST project, students will learn what DNA chromatogram files look like, learn about the significance of the four differently-colored **peaks**, learn about data quality, and learn how data from multiple samples are used in combination with quality values to identify and correct errors. Students will use their edited data in BLAST searches at the NCBI and the Barcode of Life Databases (BOLD) to identify and confirm the source of their original DNA. Students then use the bioinformatics resources at BOLD to place their data in a **phylogenetic tree** and see how phylogenetic trees can be used to support sample identification. Learning these techniques will provide students with the basic tools for inquiry-driven research.

Learning Objectives

At the end of this lesson, students will know that:

- DNA sequences can be used to identify the origin of DNA samples.
- DNA data is generated by a process called **DNA sequencing**.
- DNA sequencing produces data in the form of a chromatogram, a series of four differently colored peaks, with each color corresponding to a different DNA base.

At the end of this lesson, students will be able to:

- Describe how DNA sequencing, barcoding, and BLAST are being used to identify the origin of a wide variety of samples.
- Describe what is meant by a **quality value** when this term is used in connection with a DNA sequence.
- Describe how data are used to guide decision making when reconstructing a DNA sequence, including resolving any ambiguities or uncertain base calls.
- Use BLAST to compare two sequences and identify differences between the two sequences.
- Use BLAST to identify a DNA sequence.
- Determine where a sequence fits in a phylogenetic tree.

Class Time

3 class periods minimum (approximately 50 minutes each); however, up to 6 class periods may be required. If additional time is needed, portions of the student assignment may be assigned as homework.

Prior Knowledge Needed

- DNA sequence data is needed to answer genetic research questions and evaluate hypotheses (*Lesson One*).
- *COI* is the barcoding gene used for animals (*Lesson Two*).

Prior Skills Needed

- How to perform multiple sequence alignments (*Lessons Three* and *Four*).
- How to generate and interpret phylogenetic trees (*Lesson Three*).

Key Concepts

- The process of determining a DNA sequence involves copying DNA *in vitro* and comparing sequences from multiple samples (including sequencing both strands of DNA) to reconstruct the original sequence.
- Some DNA sequencing instruments store data in the form of DNA chromatograms, in which each base of DNA is represented as a different colored peak.
- Scientists use bioinformatics programs to process data from DNA sequencing instruments and create a representation of the original sequence.
- The DNA sequences that are obtained from different strands often contain differences; consequently, scientists use multiple samples (including sequences from both strands of DNA) to reconstruct the original sequence.
- Quality values are used to measure the probability that a base at a specific position has been correctly identified. Even a base with a relatively high quality value, for example Q = 30, has a certain probability (1/1000) of being incorrect.
- Scientists use data, such as quality values, to guide decisions when determining which bases were most likely to be present in the original DNA sequence.
- BLAST can be used to identify the origin of a DNA sample by comparing a new sequence to a database of sequences.

Materials

Materials	Quantity
Class set of Student Handout—*Analyzing DNA Sequences Instructions*	1 per student (class set)
Copies of Student Handout—*Data Table for Editing DNA Sequences*	1 per student
Teacher Resource—*Installing FinchTV* [**Note:** If students will be asked to install FinchTV, make a class set of copies of this handout.]	1 -or- 1 per student

Computer Equipment, Files, Software, and Media
Computer and projector to display PowerPoint slides. ***Alternative:*** Print PowerPoint slides onto transparencies and display with overhead projector.
Lesson Nine PowerPoint Slides—*Analyzing DNA Sequences.* Available for download at: http://www.nwabr.org/curriculum/advanced-bioinformatics-genetic-research.
A student version of lesson materials (minus *Teacher Answer Keys*) is available from NWABR's Student Resource Center at: http://www.nwabr.org/students/student-resource-center/instructional-materials/advanced-bioinformatics-genetic-research.
Optional: "Sanger Method of DNA Sequencing" video freely available from the Howard Hughes Medical Institute (HHMI). This video was also presented as an optional exercise in *Lesson One*, but may be helpful for students to review. The video is 51 seconds long and requires an internet connection and speakers. Available at: http://www.hhmi.org/biointeractive/dna/DNAi_sanger_sequencing.html.
DNA Chromatogram files from the Bio-ITEST website under the ***Resources*** tab. Available at: http://www.nwabr.org/curriculum/advanced-bioinformatics-genetic-research. [**Note:** It is important that students download both the "F" and "R" files for each sample (one sequence for each strand of DNA). Alternatively, you may use DNA sequence data generated in class.]
Computer lab with internet access and a simple text editing program such as Microsoft® Notepad (preferred for PC) or TextEdit (preferred for Mac). [**Note:** Use of Microsoft® Word is not recommended when performing bioinformatics analyses.]

Teacher Preparation

- Load the classroom computer with the *Lesson Nine* PowerPoint slides.
- Make copies of the Student Handouts. Student Handout—*Analyzing DNA Sequences* is designed to be used as part of a class set, with students writing their answers to questions in their lab notebooks or on a separate sheet of paper. Each student will need his own copy of Student Handout—*Data Table for Editing DNA Sequences*.
- To maximize class time for the lesson activities, teachers may find it useful to install the FinchTV program on classroom computers before class. See Teacher Resource—*Installing FinchTV* for complete instructions. Because downloading the FinchTV program requires free registration and an installation link sent to your email, we strongly recommend that the FinchTV program be installed by the teacher or the school IT department before class. Alternatively, make a class set of the *Teacher Resource* and have students install the program.
- Choose the option that you will be using to obtain the sequences:
  a. Use DNA chromatograms provided by Bio-ITEST and available to download at the Bio-ITEST website.
  b. Use DNA sequence data obtained through student research projects performed in your classroom.

[**Note:** Because the chromatogram data used in this lesson are authentic sequences, no answer key is provided with this lesson.]

[**Note:** If you have questions or need additional support, contact Bio-ITEST staff. Contact information is available on the NWABR contact page: http://nwabr.org/landing/contact-us.]

Procedure: Day One

PART I: Overview of DNA Sequence Analysis and Comparing DNA Sequences Using BLAST

1. Explain to students the ***aims of this lesson***. Some teachers may find it useful to write the aims on the board.
   a. **Lesson Aim:** Learn how to analyze DNA sequence data, using the bioinformatics tools **FinchTV** and **BLAST**.
   b. **Lesson Aim:** To identify an unknown DNA sequence.

   Teachers may also wish to discuss the *Learning Objectives* of the lesson, which are listed at the beginning of this lesson plan.

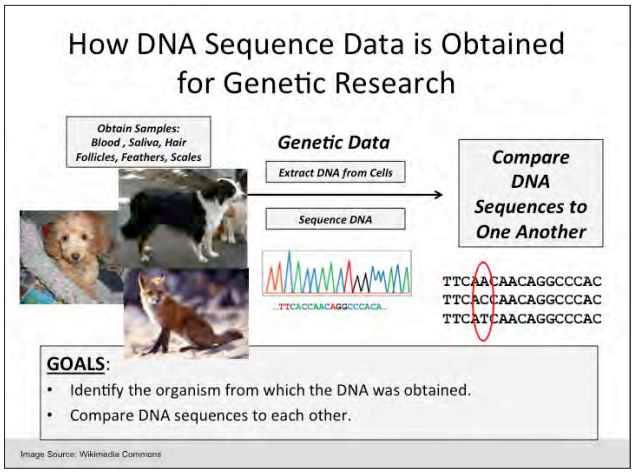
2. Show students ***Slide #1*** to remind students how DNA sequence data is obtained for genetic research:
   - Samples are obtained– almost any sample that contains cells with DNA can be used. This can include samples not found in a hospital or doctor's office, such as feathers or fish scales.
   - DNA is extracted and sequenced. Scientists obtain DNA data by a process called **DNA sequencing**, in which all of the nucleotides (A's, T's, C's, and G's) of a given region of DNA are determined or "read." This concept was first introduced in *Lesson One*.
   - These DNA sequences are the raw data used for analysis in genetic research. You may wish to show the HHMI video "Sanger Method of DNA Sequencing" for students unfamiliar with the process. This video was first presented as an optional exercise in *Lesson One*.

FinchTV: FinchTV is a chromatogram-viewing program written by scientists at Geospiza, Inc. (now PerkinElmer) that is used for presenting a graphical display of **trace files** like DNA **chromatograms**. If **quality values** are present, these can also be displayed.

DNA sequencing: The process of determining the identity and order of bases in a molecule of DNA. For more information, see *Teacher Background* and *Glossary*.

- ***DNA sequence data can be used to identify the organism from which the DNA was obtained, and to compare sequences to one another.***

Analyzing DNA: ***Slide #1***

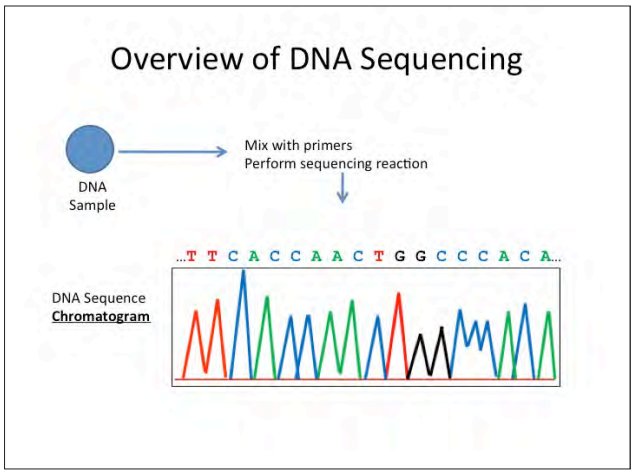


3. Tell students that today they will learn how to view and edit DNA sequences of the barcode gene *COI*, either from a sample they prepared in their classroom wet lab experiments, or from an unknown sample provided by Bio-ITEST.

Chromatogram: A chromatogram is a type of data file produced by a DNA sequencing instrument. For more information, see *Teacher Background* and *Glossary*.

4. Show ***Slide #2***, "Overview of DNA Sequencing." Explain that DNA sequencing instruments produce files called **chromatograms**. Each chromatogram file contains the data from a single sequencing reaction.

Analyzing DNA: ***Slide #2***



Peak: A point where the signal intensity from a fluorescent dye is stronger than the intensity in the surrounding areas. Each colored peak represents a different DNA nucleotide (green for adenine, red for thymine, blue for cytosine, and black for guanine).

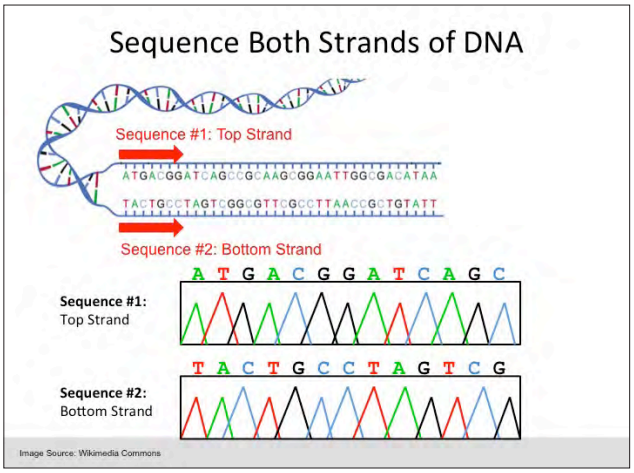
5. In the DNA chromatogram, each DNA base is represented as a **peak** of a different color. The DNA sequencing instrument "reads" the concentration measurement for each base and uses that data to determine the most likely identity for each base at each position. Sequencing instruments also produce text files showing the identity and order of all the bases (the DNA sequence).

6. Write the colors of each peak on the board. It is helpful to write each base in the corresponding color (for example, write "adenine = green" in green pen or chalk). This key is also in *Part II* of Student Handout—*Analyzing DNA Sequences* for student reference. These colors represent each DNA base in a variety of different sequence analysis programs, including FinchTV, which is used in this lesson.

   Adenine (A) = Green
   Thymine (T) = Red
   Cytosine (C) = Blue
   Guanine (G) = Black

7. Explain to students that computers do much of the DNA sequence analysis and record the computer program's interpretation of the DNA sequence. However, it is important for scientists to review that data to be sure the computer program is correct. Sometimes the computer makes errors, or cannot distinguish between two DNA peaks, requiring scientists to review additional data to identify the correct base.

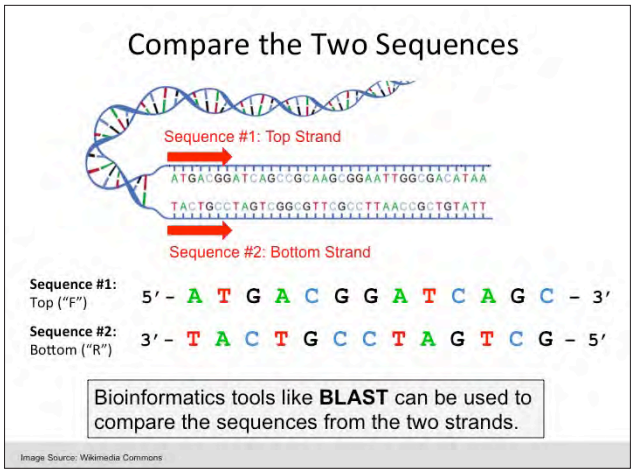
8. Show ***Slide #3***, and remind students that all DNA molecules are double-stranded. Therefore, both strands of DNA in a given sample will be sequenced. In this slide, ***Sequence #1*** is the sequence from the top strand of DNA, while ***Sequence #2*** is the sequence from the bottom strand of DNA. We have two DNA chromatograms and two DNA sequences. Note that the sequence of the bottom strand is shown in reverse order (3′ to 5′ instead of 5′ to 3′).



Analyzing DNA: ***Slide #3***

9. Show ***Slide #4***, which aligns the two DNA sequences from ***Slide #3***. This is where the work of the genetic researcher begins. Ask students if they believe these two sequences came from opposite strands of the same piece of DNA. Are they complementary to each other?

Analyzing DNA: ***Slide #4***

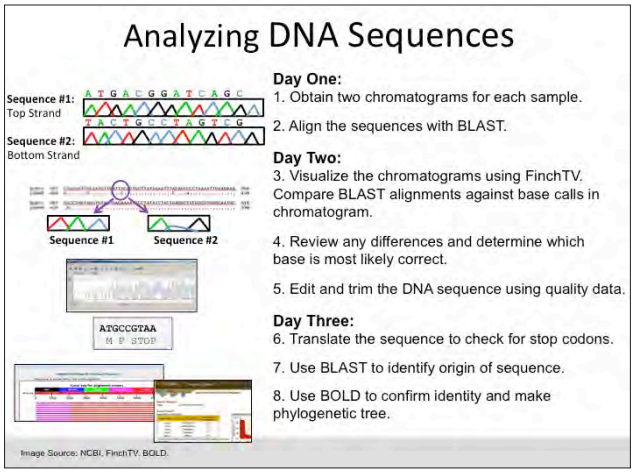


10. Point out to students that bioinformatics tools like BLAST, which they have used in previous lessons, can be used to compare the two DNA sequences.

11. Tell students that genetic researchers often refer to one of the DNA sequences as the "F" sequence, and one as the "R" sequence, as seen in ***Slide #4***. "F" stands for "Forward," and "R" stands for "Reverse," which are the names of the primers used when performing the DNA sequencing reactions.

12. Show ***Slide #5***, "Analyzing DNA Sequences," which summarizes the steps students will perform when analyzing their DNA sequence data. Today, they will perform Steps #1 and #2: obtain their DNA chromatograms from the Bio-ITEST website or from data generated in their classroom, and compare the two DNA sequences (one from each strand) using BLAST.

Analyzing DNA: ***Slide #5***



13. Tell students that they will work with the DNA chromatograms on *Day Two*, and finish their analyses on *Day Three*, when they translate their DNA sequences ***in silico*** and use BLAST and BOLD to identify the organisms from which their DNA sequences originated.

In silico: On the computer.

14. Assign students to small groups of three or four students each.

15. Pass out the class set of *Analyzing DNA Sequences*, one per student, and have students work through *Part I* of the handout on their own. This handout can be used as a class set, with students writing their answers to questions in their lab notebooks, or on a separate sheet of paper.

Closure: Day One

16. Summarize today's lesson:
    - Students completed the first portion of their DNA sequence analysis: aligning and comparing the sequences from the two strands of DNA using BLAST, and assigning portions of their BLAST alignment to each group member. Scientists sequence both strands of DNA for a given region to improve the accuracy of their data.
    - Students' next task will be to look at the raw data in the DNA chromatograms, using **quality values** and the bioinformatics tool FinchTV to resolve any **discrepancies**, or differences between the two DNA sequences. They will learn more about quality values, discrepancies, and FinchTV in *Day Two*.
17. If students have not completed Part I of Student Handout—*Analyzing DNA Sequences*, they may either complete it as homework or *Part I* may be completed in class the following day.

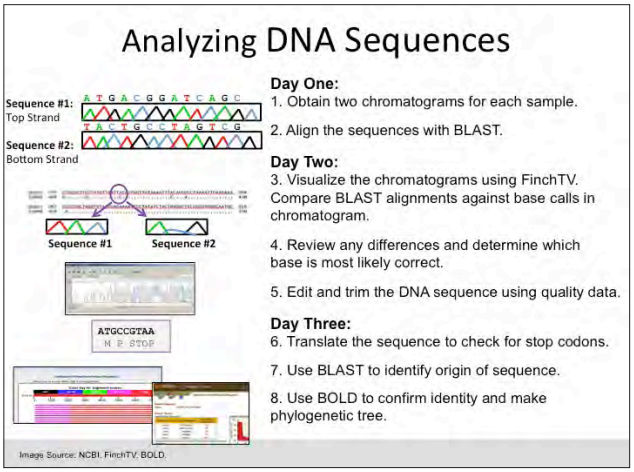
Quality values: A quality value is a number used to assess the accuracy of each base in a DNA sequence. Quality values represent the ability of the base calling software to identify the base at a given position. They are calculated by taking the log10 of the error probability and multiplying it by -10.

Discrepancy: A discrepancy in DNA sequencing is a point where the sequences from different samples or DNA strands disagree.

Procedure: Day Two

PART II: Viewing and Editing DNA Chromatograms Using Quality Values

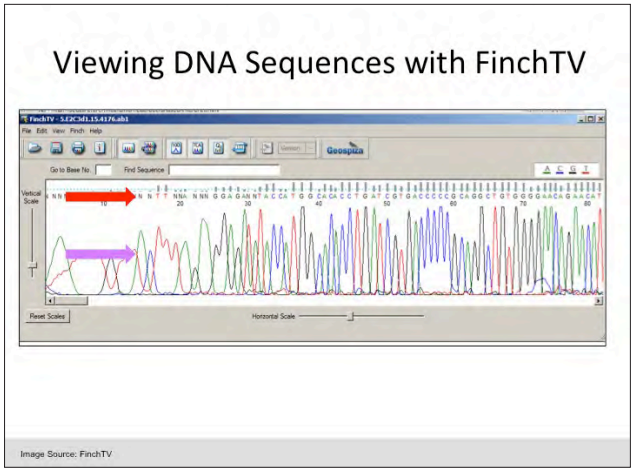
18. Show students ***Slide #5***, "Analyzing DNA Sequences," and tell them that their goals for today are:
    - Completing Steps #3, #4, and #5.
    - Reviewing the differences between the two DNA strands that they found yesterday in their BLAST alignment.
    - Using the quality values data in the DNA chromatograms to inform their decision making as they resolve these differences, or discrepancies, as they edit their DNA sequences.



Analyzing DNA: ***Slide #5***

19. Show ***Slide #6***, "Viewing DNA Sequences with FinchTV." Tell students that bioinformatics tools are used by many different kinds of scientists to view and analyze DNA sequences. FinchTV is one of these programs, and the company that created it makes it freely available for all scientists (and student researchers!) to use.

Analyzing DNA: ***Slide #6***

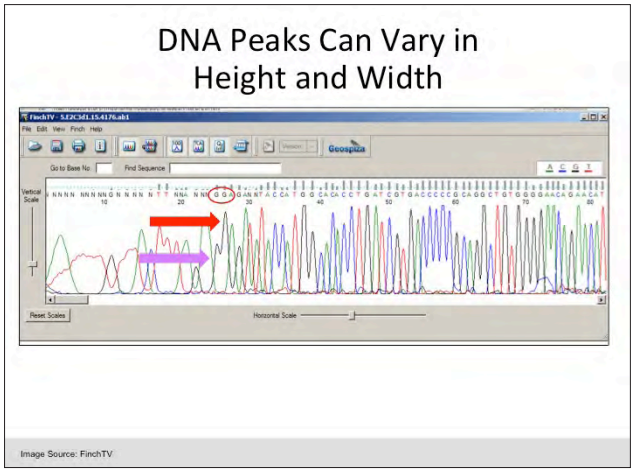


20. Explain that, as shown in ***Slide #6***, DNA analysis programs for viewing chromatograms show the four-colored DNA peaks (***purple arrow***) and the DNA sequence assigned by the computer program (***red arrow***). Most DNA analysis programs use the same color-coding for each DNA base:

    Adenine (A) = Green
    Thymine (T) = Red
    Cytosine (C) = Blue
    Guanine (G) = Black

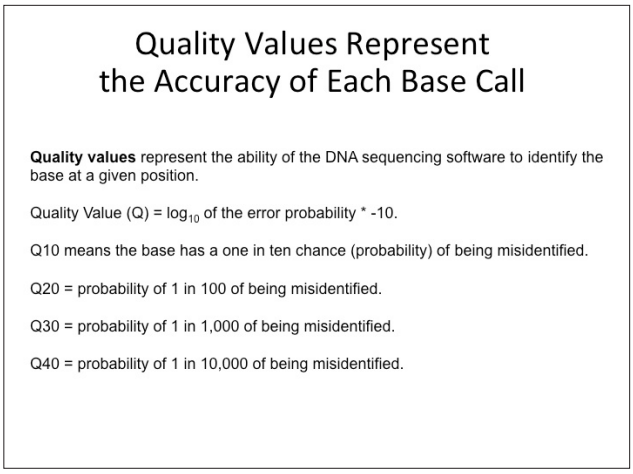
21. Show ***Slide #7***, and point out that the DNA sequence peaks may vary in both height and width, as seen with the guanine peaks at bases #25 (***red arrow***, ***tall peak***) and base #26 (***purple arrow***, ***short peak***). This is all right, as the peaks are still ***clear***, ***evenly spaced***, and most importantly, there is ***only one main peak***. However, there are important additional data to use when evaluating DNA sequence data, as seen in the next slide.

Analyzing DNA: ***Slide #7***



22. Show ***Slide #8***, and explain to students that DNA chromatograms contain not only the colored peaks that represent each base, but also important information about the ability of the DNA sequencing software to identify each base. This information is called the quality value, and is used by scientists when they analyze DNA sequence data. Quality values are calculated as the log10 of the error probability multiplied by -10. A quality value of 10 (Q10) means that there is a 1 in 10 chance that the sequencing program was wrong in identifying a given base. Q20 is a probability of 1 in 100 for misidentification; Q30 is a probability of 1 in 1,000 for misidentification; and Q40 is a probability of 1 in 10,000 for misidentification.

Analyzing DNA: ***Slide #8***

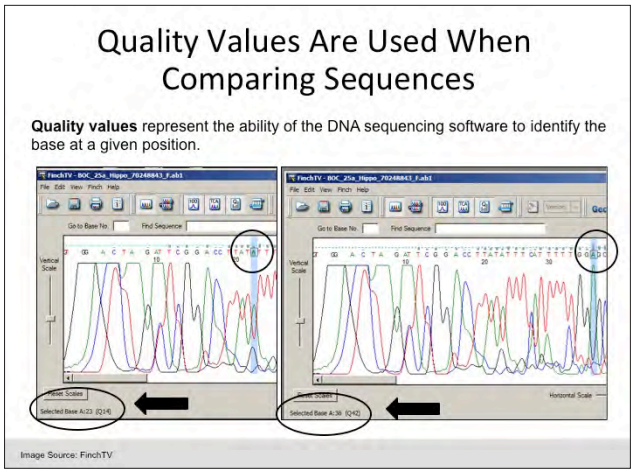


23. Tell students that scientists that wish to publish their DNA sequence data at the NCBI must have sequences with quality values of Q30 or higher. Generally speaking, scientists compare the quality values of each discrepant base to inform their decisions when comparing the DNA sequences from both strands of DNA.

24. Show ***Slide #9***, which shows how FinchTV displays quality values. When a base is selected (***black circles***), the information for the base is displayed in the lower left corner. In the chromatogram on the left, we can see "Selected Base A:23 (Q14)" which means that the base selected (or highlighted) is an A at position number 23, and has a quality value of 14. In contrast, in the chromatogram on the right, base A36 has a quality value of 42. The quality value is also represented as a small histogram above each DNA base.

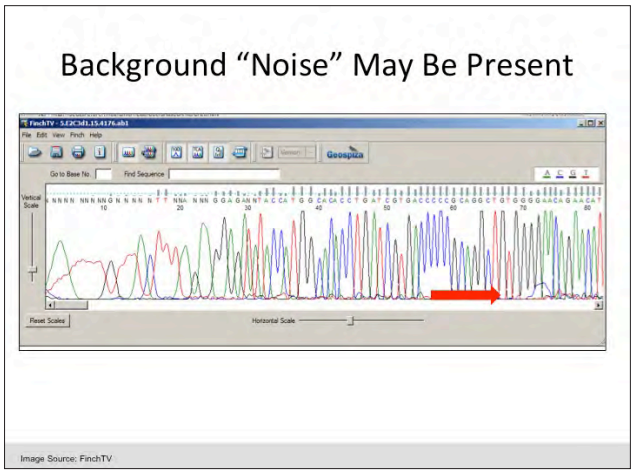
    If a scientist were comparing the DNA sequences from two strands of DNA, at any position where there is discrepancy they would choose the base with the higher quality value. For example, if the F sequence from one strand of the DNA shows an A at position 34 with a quality value of 25, while the R sequence from the other DNA strand shows a G at position 34, but only has a quality value of 12, the scientist would assume that the A at position 34 is correct.

Analyzing DNA: ***Slide #9***



25. Show ***Slide #10*** and bring students' attention to the tiny peaks at the bottom of the sequence (***red arrow***). This is called "noise" (like static in the background on the radio) or "background peaks." A good quality DNA sequence should not have much noise. It is important for students not to confuse the noise with the sequence of the gene they are analyzing.

Analyzing DNA: ***Slide #10***



26. Show students ***Slide #11***, which highlights the poor quality peaks at the beginning of the DNA sequence file (***purple circle***). These peaks are irregularly shaped and irregularly spaced. If the software in a DNA sequencing instrument determines that a base should be at a certain position, but is unable to identify that base, it calls that base an "N" (for "nucleotide").

Analyzing DNA: ***Slide #11***

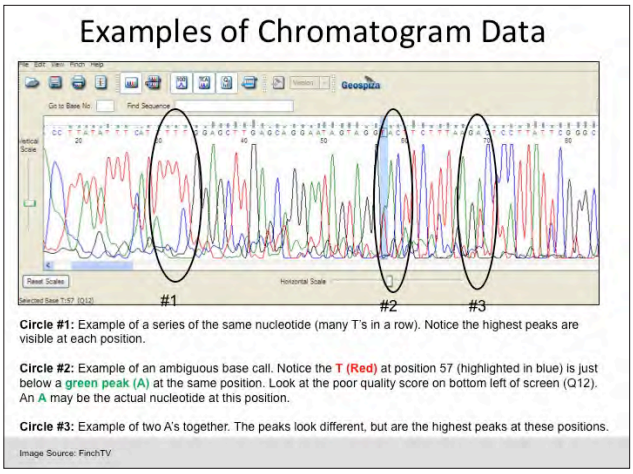
27. Tell students that many errors in DNA sequencing tend to occur at the beginning or the end of DNA sequences, and scientists often use bioinformatics programs to trim or remove these poor quality portions before performing other analyses. This is one of the steps the student researchers will do today.

28. Show students ***Slide #12***, which shows more examples of some common challenges when analyzing chromatograms:

- ***Circle #1*** shows a series of multiple T's; point out to the students that they can still identify the tops of each peak at each nucleotide position, but a repeating series of a single base can lead to unevenly spaced peaks. The highest peak (the peak with the strongest signal), is usually the correct base at these positions.
- ***Circle #2*** shows an example of an ambiguous **base call** at position 57. Point out to the students that there is both a red peak (T) and a green peak (A) at this position, with a poor quality score (Q12). This is where it is handy to use the second chromatogram to help the students identify the true base at this position. This is why scientists routinely sequence both strands of DNA.
- ***Circle #3*** shows an example of two adjacent A peaks (green) that have different intensities (different peak heights), yet are still the best calls for the bases at these positions.

Base calls: The process of identifying the base that produced the strongest signal at a given point in the DNA sequence. DNA sequencing instruments contain bioinformatics software that analyzes the data produced when the instrument is run. In Sanger sequencing, the software records the intensity of the fluorescent signal from each base, determines which base produced the strongest signal (i.e., "calls" the base), and records the identity of that base (green = adenine; red = thymine; blue = cytosine; black = guanine).

Analyzing DNA: ***Slide #12***



29. Tell students to work through *Parts II-V* of Student Handout—*Analyzing DNA Sequences* with the other members of their group. It may take students two class periods to complete this portion of the lesson.

Closure: Day Two

30. Summarize today's lesson:
    - Students have learned that DNA sequencing data is presented in chromatograms, in which each base is represented by a different color.
    - They have also learned how to assess DNA sequence quality and use the quality value data in DNA chromatograms to inform their decision making about discrepancies between two DNA sequences.
    - In the last section, students will learn how to use a new BLAST program called **blastx** to translate their edited DNA sequence *in silico* to determine whether any stop codons were introduced during the editing process, such as by accidentally adding or removing a base. They will then use blastn to identify the kind of organism from which their DNA was obtained, as they did early in the *Genetic Research* curriculum, and confirm those findings at BOLD.

blastx: This program compares the six-frame translation products of a nucleotide query sequence (both strands) against a protein sequence database at the NCBI.

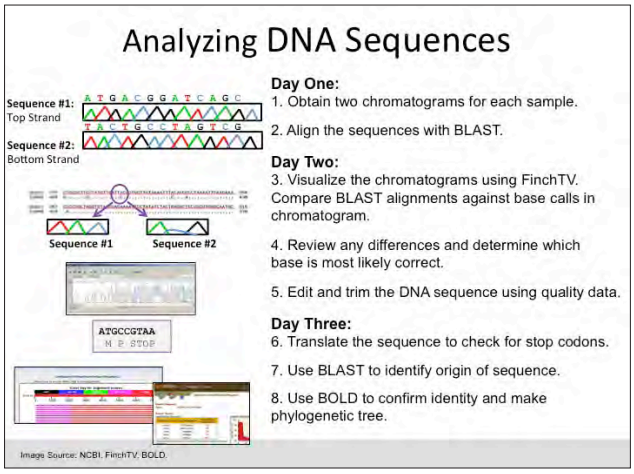
31. If students have not completed *Parts II-V* of Student Handout—*Analyzing DNA Sequences*, they may either complete them as homework or *Parts II-V* may be completed in class the following day. As FinchTV would be required to complete these portions of the *Student Handout*, it may be best to provide additional class time to complete these sections.

Procedure: Day Three

PART III: Reviewing the Edited Sequence by Translating Sequences with blastx and Identifying the Organism From Which the DNA Was Obtained

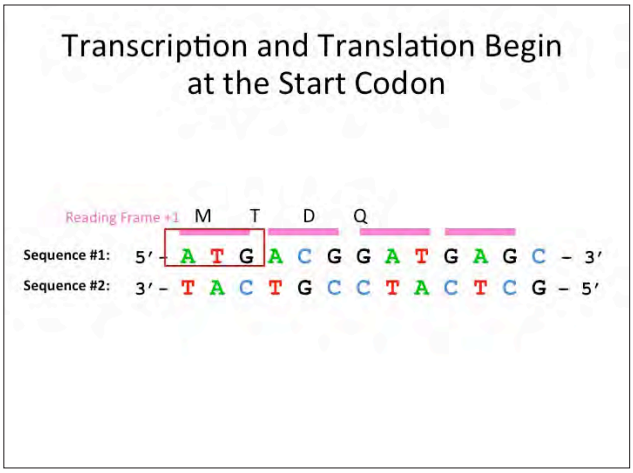
32. Show ***Slide #13***, and remind students that in the previous activities for this lesson, they have obtained their DNA chromatograms, aligned and compared their F (Forward) and R (Reverse) sequences using BLAST, and edited one of their DNA sequences using the quality value data to resolve any base discrepancies.

Analyzing DNA: ***Slide #13***



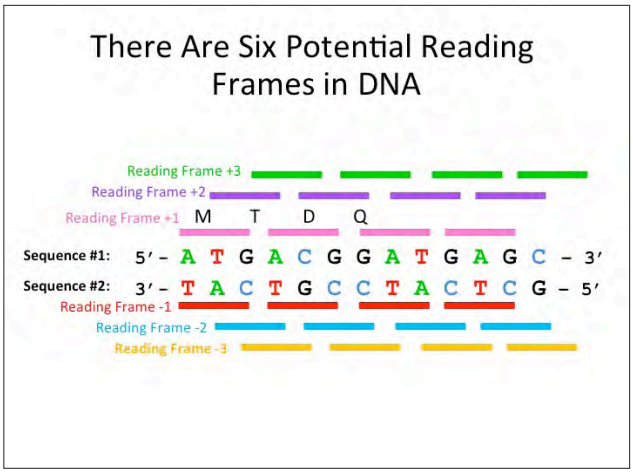
33. Tell students that they will conclude the lesson today by translating their edited DNA sequence *in silico*, as they did in *Lesson Four*, to confirm that they did not accidentally introduce any stop codons to their sequence while they were editing.

34. Show ***Slide #14***, and remind students that transcription and translation often begin at the ATG "start codon" and proceed in a 5′ to 3′ direction.



Analyzing DNA: ***Slide #14***

35. Show ***Slide #15***, and remind students that for any given DNA molecule, there are six potential reading frames: three on the top strand, and three on the bottom strand.

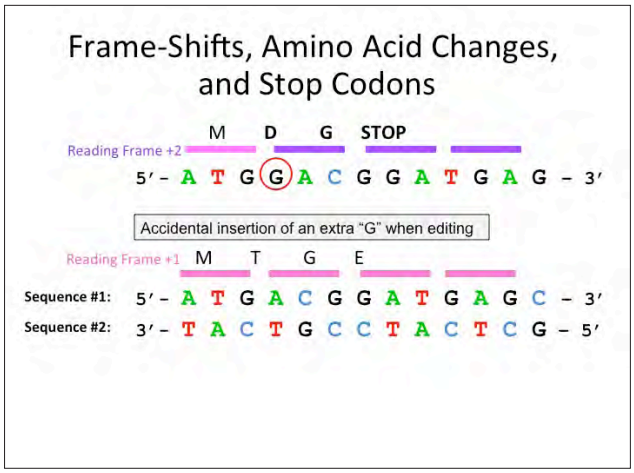


Analyzing DNA: ***Slide #15***

36. Show ***Slide #16*** and explain to students that **frame shifts** occur when one or two bases are inserted or deleted, "shifting" the reading frame from one frame to another. As seen in ***Slide #16***, if a scientist editing a DNA sequence accidentally adds an extra G (***circled in red***) at position #4, after the start codon, the remaining DNA would be translated in reading frame #2 instead of reading frame #1. This would change the amino acid sequence, as well as introduce an early stop codon.

Frame shift: The addition or deletion of one or more bases in a DNA sequence that alters the reading frame of the gene from the point of insertion or deletion forward.

Analyzing DNA: ***Slide #16***



37. Tell students that to check for these types of errors, they will translate their DNA sequences *in silico* using another BLAST program called blastx. Blastx translates a DNA sequence into all six possible reading frames, similar to what students did with ORFinder in *Lesson Four*, and then compares those sequences to the sequences in the NCBI protein database. If students find an error in their DNA sequence, such as an extra nucleotide, they can review their chromatograms and try to correct the error by updating their "Edited" chromatogram file.

Nucleotide BLAST: A type of BLAST program that compares a nucleotide query sequence to a database of nucleotide sequences at the NCBI.

38. Explain to students that the final steps in this activity will be to use **nucleotide BLAST** and BOLD (the Barcode of Life Database) to identify the organism from which the DNA was obtained, and use features of the Barcode of Life Database to generate a phylogenetic tree using other *COI* DNA barcode sequences currently in the BOLD.

39. Tell students to work through *Parts VI* and *VII* of Student Handout—*Analyzing DNA Sequences* with the other members of their group.

Closure: Day Three

40. As a class or within small student groups, discuss the phylogenetic trees students obtained in *Part VII* of Student Handout—*Analyzing DNA Sequences*. Questions for discussion include:
    - Which organisms are most closely related to the organism from which your DNA was obtained?
    - Were you familiar with any of the organisms in your phylogenetic tree before this lesson?
    - If you were familiar with any of these organisms, was the tree you obtained from BOLD similar to the tree you expected, based on what you already knew?
    - What do you find most surprising about this phylogenetic tree?
    - What surprising or interesting fact did you learn during your research about these organisms?

41. Summarize today's lesson in the context of the lesson as a whole:
    - Students have learned how to analyze and identify DNA sequences, aligning and comparing sequences using nucleotide BLAST, and viewing DNA chromatograms with quality values to inform their decision making to determine the correct sequence for a particular piece of DNA using FinchTV.
    - Using the bioinformatics tool blastx, students translated their DNA sequence *in silico* to confirm the results of their sequence analysis, and then used nucleotide BLAST and BOLD to identify the organism from which their DNA was isolated.
    - Using the phylogenetic tree-building tool together with the *COI* DNA barcode sequences available at BOLD, students were able to evaluate the evolutionary relatedness of their organism to other organisms that have been barcoded.
    - Students have now come full circle in the process of genetic research, from raw DNA sequence data to making conclusions about evolutionary relationships.

42. Finally, at the end of Student Handout—*Analzying DNA Sequences*, students are asked to reflect on what they have learned about the process of genetic research throughout the DNA barcoding lessons. Lead a discussion with students about this question, guiding them to reflect on the process of science. The discussion points listed below also refer to the *Key Concepts* first introduced in *Lesson One*:
    - Scientists in many different fields use bioinformatics to answer research questions specific to their field of study.
    - Like many other kinds of research, genetic research is a ***process***.
    - Scientific experiments build on what is ***already known*** about a given subject or field, using this information and observations as background when asking scientific questions.
    - The methods of science (often called the ***scientific method***) involve asking a question, formulating a hypothesis about that question, gathering data, analyzing that data to determine whether the data supports the hypothesis, making conclusions, and revising the original hypothesis if needed.
    - Scientific experiments build on one another, with both "successes" and "failures" helping to move the scientific process forward.
    - Genetic research involves asking a research question based on observations of the natural world, generating a hypothesis, obtaining DNA sequence data, and comparing and analyzing DNA sequences to address the hypothesis.

Homework

For homework, ask students to write about the things they learned in *Lesson Nine* in their lab notebooks, on another sheet of paper, or in a word processing program like Microsoft® Notepad or Word which they then provide to the teacher as a printout or via email. This can serve as an entry ticket for the following class.

Have them complete these prompts:

a. Today I learned that…

b. An important idea to think about is…

c. Something that I don't completely understand yet is…

d. Something that I'm really confident that I understand is….

Extension

Students can research some of the new DNA sequencing technologies, including high-throughput DNA sequencing techniques and "next generation" DNA sequencing. High-throughput technologies make it possible for scientists to run many DNA sequences at the same time, much more quickly than was ever possible before.

Teacher Background: DNA Sequencing

DNA sequencing is the process of determining the identity and order of bases in a molecule of DNA.

A common method for sequencing DNA involves: purifying DNA from a sample; making a copy of that DNA *in vitro*; separating the new DNA molecules by their size; and identifying the base at the end of each DNA molecule by measuring the intensity of the fluorescent signal. This entire process is commonly known as **Sanger sequencing** after Fred Sanger, the biochemist who developed the method for using **dideoxyribonucleotide triphosphates (ddNTPs)** to create DNA molecules of random sizes.

Automated DNA sequencing instruments use capillary electrophoresis to separate the differently sized molecules of DNA. Capillary electrophoresis separates DNA molecules in a small capillary tube instead of in an agarose gel. Automated DNA sequencing instruments also contain a laser that excites the fluorescent dye attached to each DNA base, instruments that capture and measure the intensity of fluorescence, and software for processing the fluorescent signal and creating a **chromatogram**. A key point to note is that the DNA bases that are measured are produced by synthesizing new DNA in vitro, and might contain differences from the original due to errors during DNA synthesis. Scientists use **chromatogram-viewing** programs like **FinchTV** to view and analyze their chromatograms and associated DNA sequence data. They use sequence assembly programs to reconstruct a model of the original sequence.

Glossary

3' hydroxyl group: During DNA synthesis, **DNA polymerase** catalyzes the formation of a phosphodiester bond between the 3' hydroxyl group on the deoxyribose at the end of a DNA strand and the phosphate group attached to the 5' carbon of the deoxyribose on the new DNA nucleotide.

Base calls: The process of identifying the base that produced the strongest signal at a given point in the DNA sequence. DNA sequencing instruments contain bioinformatics software that analyzes the data produced when the instrument is run. In Sanger sequencing, the software records the intensity of the fluorescent signal from each base, determines which base produced the strongest signal (i.e., "calls" the base), and records the identity of that base (green = adenine; red = thymine; blue = cytosine; black = guanine).

blastx: This program compares the six-frame translation products of a nucleotide query sequence (both strands) against a protein sequence database at the NCBI.

Chromatogram: A chromatogram is a type of data file produced by a DNA sequencing instrument. Chromatograms contain many types of information. The most important types of information for DNA sequence analysis are the signal intensities and the **base calls**. Chromatograms often contain **quality values** as well that can be used to evaluate the accuracy of the data. If a chromatogram file is produced by a sequencing instrument made by the company ABI (now Life Technologies), it can be recognized by the extension ".ab1" (for example, CoyoteCOI.ab1). Since chromatograms contain the signal intensities for each base, and that information can be used to generate a graph, they can also be described as a type of **trace file**.

Consensus sequence: A sequence that shows the amino acid or **nucleotide** found most often at each position in a set of aligned sequences. When performing DNA sequencing, the consensus sequence represents the best agreement between sequence data from multiple samples, such as the results from sequencing both strands of DNA. A consensus sequence can be obtained from multiple samples by using **sequence assembly** programs.

Deoxyribonucleotide triphosphates (dNTPs): These are the bases that can be used for making DNA. They are abbreviated as dATP (deoxyadenosine triphosphate), dCTP (deoxycytosine triphosphate), dGTP (deoxyguanine triphosphate), and dTTP (deoxythymidine triphosphate). A mixture containing all four deoxyribonucleotide triphosphates can also be described as a "set of dNTPs." See also **Dideoxyribonucleotide triphosphates (ddNTPs)** and **DNA sequencing**.

DNA polymerase: The enzyme that assembles new DNA molecules, in the cell or *in vitro* in the laboratory, using a DNA template and **deoxyribonucleotide triphosphates (dNTPs)**.

DNA sequencing: The process of determining the identity and order of bases in a molecule of DNA.

Dideoxyribonucleotide triphosphates (ddNTPs): Dideoxyribonucleotide triphosphates (abbreviated ddNTPs) are similar to the normal **deoxyribonucleotide triphosphates** that are used for making DNA with one change: they are missing the **3' hydroxyl group** on the deoxyribose sugar. The 3' hydroxyl group is necessary for DNA synthesis because **DNA polymerase** builds a chain of DNA by catalyzing the formation of a **phosphodiester bond** between the first phosphate group on the new dNTP and the 3' hydroxyl group at the end of the DNA strand. If the last dNTP in a DNA strand is missing the 3' hydroxyl group, DNA polymerase will be unable to add a new base and DNA synthesis will stop. When ddNTPs are used for automated DNA sequencing, they are also labeled with a fluorescent dye. When the names of these bases are abbreviated, an extra "d" is added to indicate that they are missing the 3' hydroxyl group. The abbreviations for these bases are: ddATP, ddCTP, ddGTP, and ddTTP. The normal bases are dATP, dCTP, dTTP, and dGTP.

Discrepancy: A discrepancy in DNA sequencing is a point where the sequences from different samples or DNA strands disagree. Some examples would be where one DNA sequence contains an extra base relative to another, or contains a different nucleotide at the same position. Discrepancies exist because errors can occur during DNA synthesis, or because there is a real difference between samples.

FinchTV®: FinchTV is a chromatogram-viewing program written by scientists at Geospiza, Inc. (now PerkinElmer) that is used for presenting a graphical display of **trace files** like DNA **chromatograms**. If **quality values** are present, these can also be displayed.

Frame shift: The addition or deletion of one or more bases in a DNA sequence that alters the reading frame of the gene from the point of insertion or deletion forward.

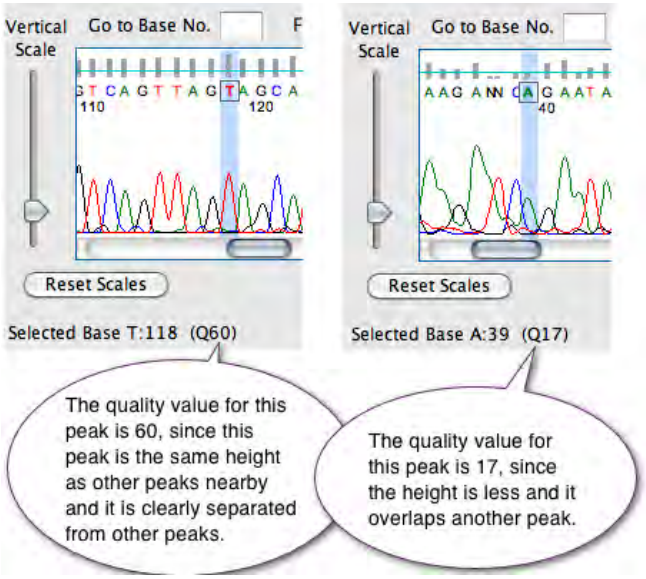
In silico: On the computer.

Nucleotide BLAST: A type of BLAST program that compares a nucleotide query sequence to a database of nucleotide sequences at the NCBI.

Peak: A point where the signal intensity from a fluorescent dye is stronger than the intensity in the surrounding areas. Each colored peak represents a different DNA nucleotide (green for adenine, red for thymine, blue for cytosine, and black for guanine).

Phosphodiester bond: A phosphodiester bond is a strong covalent bond between a phosphate group and two five-carbon ring carbohydrates. **DNA polymerase** catalyzes the formation of phosphodiester bonds between the **3' hydroxyl group** on the deoxyribose at the end of a DNA strand and the phosphate group attached to the 5' carbon of the deoxyribose on the new DNA **nucleotide**.

Quality values: A quality value is a number used to assess the accuracy of each base in a DNA sequence. Quality values represent the ability of the base calling software to identify the base at a given position. They are calculated by taking the $\log_{10}$ of the error probability and multiplying it by -10. A base with a quality value of 10 has a one in ten chance of being misidentified. Bases with quality values of 20, 30, and 40, have error probabilities of one in 100, one in 1,000, and one in 10,000, respectively. Many databases ask that submitted DNA sequences have an average quality value close to 30 or higher. Quality values can be used to help guide decisions about the **discrepancies** between different sequences.



Query: When searching databases like those at the NCBI, your "query" is the sequence you are searching with and trying to match. In this case, your query is your unknown sequence.

Sanger sequencing: A method of DNA sequencing that uses **dideoxyribonucleotide triphosphates (ddNTPs)** to end DNA synthesis at random positions.

.scf files: ".scf" is a file extension used to identify files that only store a portion of the DNA sequence data found in a **chromatogram**. SCF stands for "Staden Compressed Format."

Sequence assembly: DNA sequence assembly is the process of constructing a model of the original DNA sequence by aligning and comparing sequence data from multiple samples. See also **consensus sequence**.

Subject: When searching the databases at the NCBI, the subject sequences are sequences from the database that match the query. In this case, if a subject sequence is identical to your query, your query sequence probably came from the same creature that contributed the subject sequence.

Trace file: The term "trace file" is used informally to describe a data file that contains sufficient information for drawing a "trace" or graph of the signal intensities for each DNA base. Both **chromatograms** and **.scf files** are considered trace files.

Resources

The Howard Hughes Medical Institute (HHMI) **Biointeractive** site offers a variety of freely-available videos and animations. The following animations may be of particular interest for this lesson:

- Sanger Method of DNA Sequencing (0:52 minutes):
  http://www.hhmi.org/biointeractive/dna/DNAi_sanger_sequencing.html
- Polymerase Chain Reaction (1:27 minutes):
  http://www.hhmi.org/biointeractive/dna/DNAi_PCR.html
- Human Genome Sequencing (1:48 minutes):
  http://www.hhmi.org/biointeractive/dna/DNAi_human_genome_seq.html

Credit

Teacher Background: DNA Sequencing adapted from Wikipedia, "DNA sequencing." Accessed July 20, 2010.

FinchTV, Version 1.4, is a free DNA sequence analysis program provided by Geospiza, Inc. (now PerkinElmer, Inc.).

9 Analyzing DNA Sequences Instructions

Student Researcher Background:

DNA Analysis and FinchTV

DNA sequence data can be used to answer many types of questions. Because DNA sequences differ somewhat between species and between individuals within a species, DNA sequences are widely used for identification. In this activity, you will use bioinformatics programs to work with DNA sequences and identify the origin of a DNA sample.

Discrepancy: A discrepancy in DNA sequencing is a point where the sequences from different samples or DNA strands disagree.

Quality values: A quality value is a number that is used to assess the accuracy of each base in a DNA sequence. Quality values can be used to help guide decisions about the **discrepancies** between different sequences. For more on quality values, see *Part II*.

Aim: Today, your job as a researcher is to:

1. Obtain two chromatograms for each sample. The two chromatograms should represent sequences from both strands of DNA.
2. Use BLAST to compare the sequences from the two chromatograms.
3. Review each **discrepancy** between the two sequences and use **quality values** to determine which base is most likely to be correct.
4. Edit and trim the DNA sequence by using quality data from the two chromatograms.
5. Translate the sequence to check for stop codons.
6. Use BLAST to identify the origin of the DNA sequence.
7. Use BOLD to confirm the identification of the species (or genus) and place the sample in a phylogenetic tree.

Instructions: Write your answers to the questions in your lab notebook or on a separate sheet of paper, as instructed by your teacher.

PART I: Compare Sequences from the Two DNA strands with BLAST

It is a common practice in many labs to sequence both strands of DNA. These sequences can then be compared to identify and correct any potential errors. Together, the two files contain sequence data from both strands of DNA.

1. ***Working with your group members, record the name of each of your data files in your lab notebook or on a separate sheet of paper.*** These files may be provided by your teacher, or you may visit the Bio-ITEST website to retrieve a set of unknown samples to analyze at: http://www.nwabr.org/curriculum/advanced-bioinformatics-genetic-research.

   Click on the ***Resources*** tab and scroll down to Lesson Nine to select your DNA chromatogram.

   [**Note:** As each DNA sequence file is from a different stand of DNA, one file will include "F" in the file name, while the other file will include "R" in the file name.]

[**Note:** When DNA sequences are obtained from different strands, the file names frequently include a letter such as "F" or "R" to show that the sequences are from different strands. These come from the PCR primers used to sequence the DNA, Forward ("F") and Reverse ("R").]

2. ***If you are using data from your classroom experiments***, right-click the name of each DNA sequence file and choose ***Save as*** to download the file and save it to your desktop.

   If you are using data from the Bio-ITEST website, click the file link on the website and, when prompted, select ***Save***, and then save the file to your desktop.

3. The "R" sequence file will be used as your ***subject sequence*** in your BLAST comparison below. Record the name of this file in your lab notebook or on a separate sheet of paper, and be sure to label it "subject sequence."

4. Open your ***subject sequence*** file in FinchTV using ***one*** of the methods listed below. This file will have the file extension ".ab1" and the file icon may look like a small DNA chromatogram.

   To open the .ab1 chromatogram file, use one of these methods:

   - Double-click the sequence file.
   - Right-click the sequence, choose ***Open with*** and navigate to find FinchTV.
   - Open FinchTV, then open the ***File*** menu in FinchTV, choose ***Open***, and navigate to find and choose the file.
   - Open FinchTV, and then drag and drop the sequence file into the sequence viewer window of FinchTV where it says, "Drag chromatogram file here to get started."

5. Open the FinchTV ***File*** menu, select ***Export*** and choose ***DNA sequence: FASTA***.

6. Save the exported FASTA file on your desktop. This file will have the file extension ".seq" and contains only the text of your DNA sequence (the A's, T's, G's, and C's). This file extension may not be recognized by your computer, and will probably have an icon that looks like the image at the right.

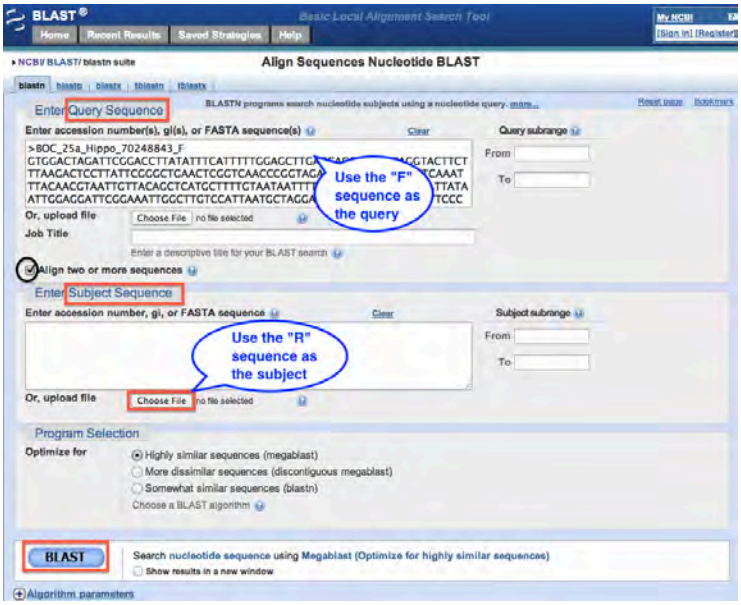
7. Open your second .ab1 DNA sequence file ("F") in FinchTV. The data in this file will be used as your ***query sequence*** in BLAST.

8. Record the file name for your ***query sequence*** in your lab notebook or on a separate sheet of paper. Be sure to label it "query sequence."

9. To use BLAST to compare the data from the two strands ("F" for your query sequence and "R" for your subject sequence), with your query sequence file open in FinchTV from the previous two steps, go to the ***Edit*** menu and choose ***Select All***.

10. Once your entire sequence is highlighted in the sequence window, go back to the ***Edit*** menu and select ***BLAST Sequence***, and then select ***Nucleotide***, ***blastn***. The NCBI BLAST page will open with your ***query sequence*** already pasted in the window (see ***Figure 1***).

Figure 1: Entering Sequences to Compare Using BLAST. Source: NCBI BLAST.



11. Check the box to "Align two or more sequences" (black circle, ***Figure 1***). A second sequence window will appear with the heading "Enter Subject Sequence" (red box, ***Figure 1***).

12. Click the ***Browse*** or ***Choose File*** button (red box, ***Figure 1***) and navigate to find your exported ***subject sequence*** file (which you downloaded to your desktop in Step #3). Be sure that this is the file with the ".seq" file extension (not an ".ab1" file).

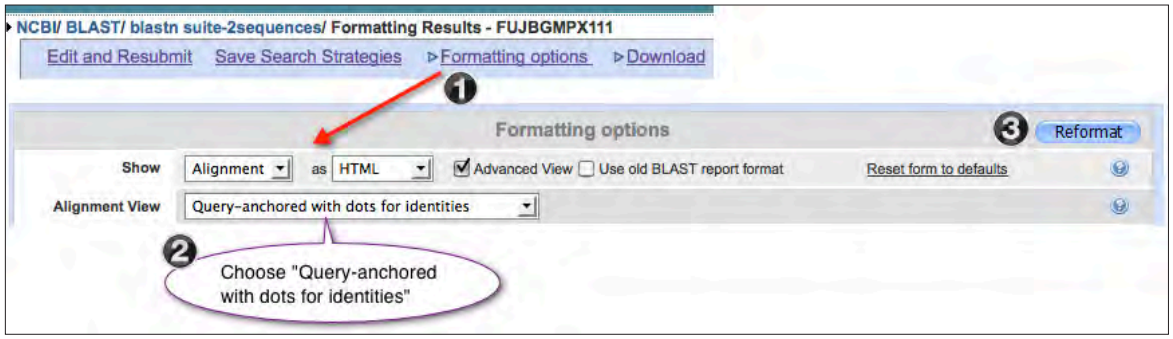
13. Open your sequence by selecting ***Open*** from the navigation menu, or by double-clicking on your .seq subject sequence file. The file name will appear in the box beside the ***Browse*** or ***Choose File*** button in the BLAST window.

14. Click BLAST (red box, ***Figure 1***). Your ***subject sequence*** will be uploaded when the BLAST alignment begins.

15. To make it easier to view these discrepancies, reformat the BLAST output (see ***Figure 2***) by following these steps:
    a. Click ***Formatting Options*** at the top of the page to open the ***Formatting*** menu (Step 1).
    b. From the ***Alignment View*** menu, select ***Query-anchored with dots for identities*** (Step 2).
    c. Click ***Reformat*** (Step 3).

[**Note:** BLAST will align and compare the sequences from the two DNA sequence files and identify points where they differ. Although both sequences came from the same DNA sample, each sequence was derived from a different strand. Sometimes errors in the DNA sequencing reactions can lead to differences, or **discrepancies**, between the two sequences.]

Figure 2: Reformatting the BLAST Results. Source: NCBI BLAST.



16. Scroll down below the format menu, and you will see that your alignment has been reformatted to show the ***query sequence*** in the top row of the alignment and the ***subject sequence*** in the bottom row. Dots are used in the alignment to show positions where the two sequences are identical. Where the sequences differ, letters representing bases, or other symbols, indicate positions where a base either differs or has been inserted or deleted relative to the other strand.

17. To print copies of your alignment (see ***Figure 3***), follow these steps:
    a. Click the Download link at the top of the BLAST results page (Step 1).
    b. Click Text. The aligned sequences will be downloaded as a Text (".txt") file (Step 2).
    c. Open the file in a text-editing program such as Microsoft® Notepad or Word and print the file.

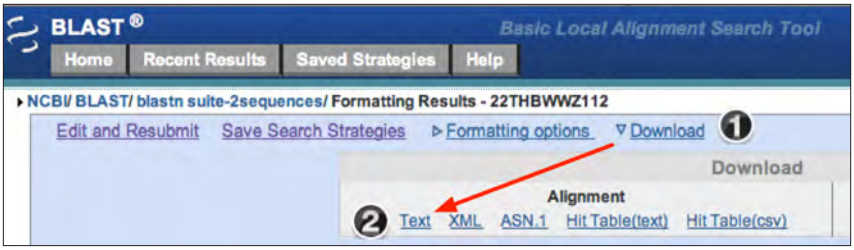


Figure 3: Downloading a Text File of Your Sequence Alignment. Source: NCBI BLAST and Microsoft® Notepad.

18. Divide up your alignment, as shown in ***Figure 4***, by highlighting or circling and labeling each part of your alignment. Each member of your group will be responsible for editing a different part of your sequence.

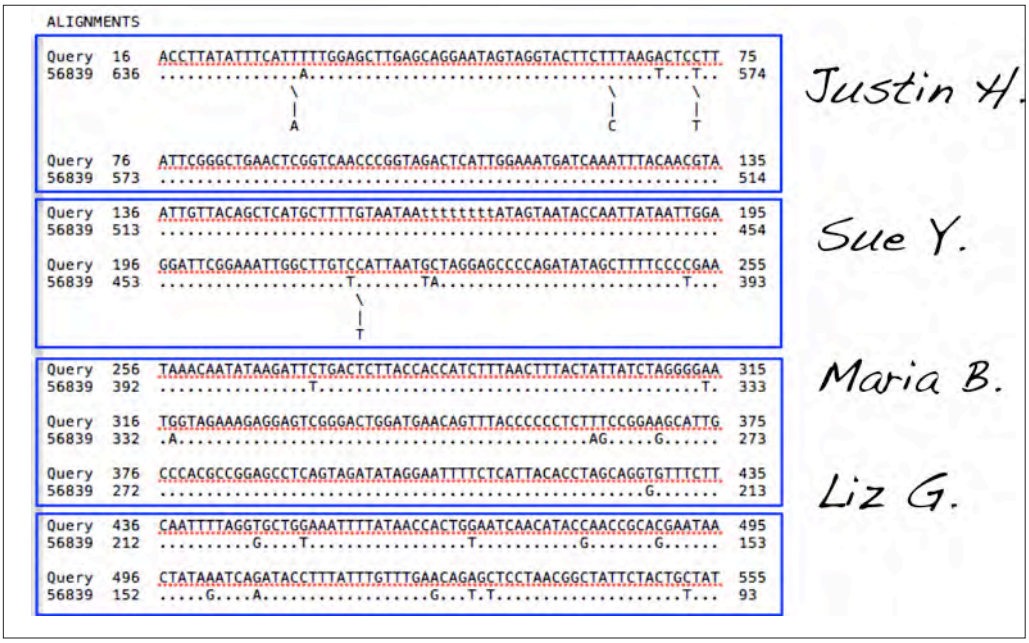


Figure 4: Assigning Portions of Your Sequence Alignment to Group Members.

PART II: Learning to Work with Sequences

Student Researcher Background:

Using FinchTV for DNA Analysis

FinchTV is a designed to allow researchers to view DNA sequence files like the **chromatograms** you are using here. In a chromatogram file, the signal intensities are presented in a graph with the four bases, each identified by different colors. Like many sequence analysis programs, FinchTV uses green for adenine, red for thymine, black for guanine, and blue for cytosine, as seen in the *DNA Sequencing Key* below.

DNA Sequencing Key

Adenine (A) = Green	Thymine (T) = Red	Cytosine (C) = Blue	Guanine (G) = Black

A. Getting Familiar with FinchTV

19. If it is not already open, open your ***subject sequence*** chromatogram file ("R" sequence with the ".ab1" file extension from *Part I* above) in FinchTV. If you need assistance finding or opening the file, look back at the instructions in *Part I*.

20. Be sure that you have recorded the name of your ***subject sequence*** file in your lab notebook or on a separate sheet of paper. Be sure to label the sequence "subject sequence."

21. Use the ***Vertical Scale*** adjustment on the left side of the program window to adjust the peak height, as shown in ***Figure 5***. It is important for you, the researcher, to be able to clearly see the DNA sequence peaks. The height of a peak corresponds to the relative concentration of that base, at that position in the sequence. The height should be high enough for you to see clearly, but not so high that the background or "noise" peaks at the bottom of the chromatogram (***black arrow***) overwhelm your sequence data (***white arrow***).

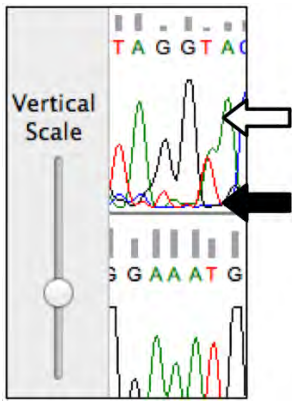


Figure 5: Vertical Scale.
Source: FinchTV.

22. Click the ***Wrapped View*** icon to view the entire sequence in one screen.

23. Click the ***Base Position Numbers*** icon to view the base position numbers throughout the sequence.

24. Click the ***Base Calls*** icon to view the base calls (i.e., what the computer program interprets the sequence to be).

25. Click the ***Quality*** icon to display the quality bar graph above each DNA sequence peak. When evaluating data, it is important to look not only at what the data is, but at the quality of the data. The **quality value** for DNA sequences is expressed as the "Q" value ("Q" for "Quality").

Quality values: A quality value is a number used to assess the accuracy of each base in a DNA sequence. Quality values represent the ability of the base calling software to identify the base at a given position and are calculated by taking the log10 of the error probability and multiplying it by -10.

- A base with a quality value of 10 has a one in ten chance of being misidentified.
- Bases with quality values of 20, 30, and 40 have error probabilities of one in 100, one in 1,000, and one in 10,000, respectively.

Many databases ask that submitted DNA sequences have an average quality value close to 30 or higher. Quality values can be used to help guide decisions about editing the **discrepancies** between different sequences, as you will do below.

B. Viewing information for a specific base

26. With the ***quality values*** displayed for your sequence, select a base by clicking it with your mouse. The selected base will be highlighted, as seen in ***Figure 6***.

27. The one letter abbreviation for that base will appear in the lower left corner, along with the sequence position and the ***quality value*** (if available). In ***Figure 6***, the selected base is a T (thymine) located at position 70 in the sequence and has a quality value of 17, which is generally accepted to be low quality.

28. Experiment by clicking on a number of different bases in your sequence. Answer these questions in your lab notebook or on another sheet of paper:

    What is the highest quality value you see?

    What is the lowest quality value you see?

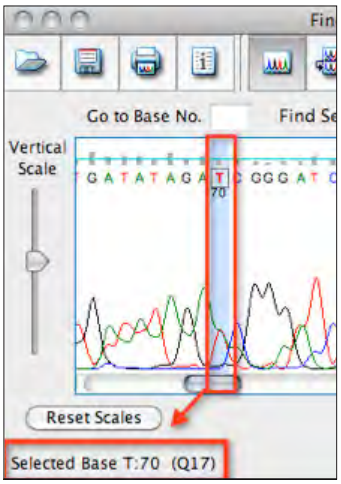


Figure 6: Quality Values. Source: FinchTV.

C. Finding a base or sequence in FinchTV

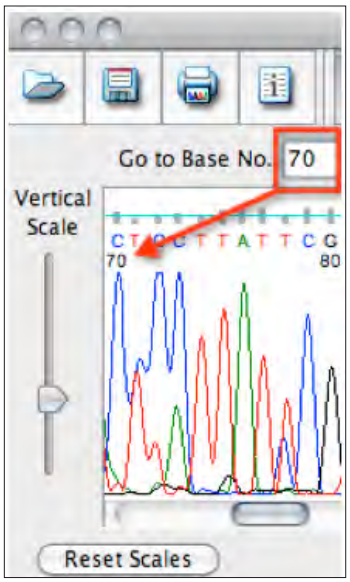
29. To find a specific base, enter the position number for that base in the ***Go to Base No***. window and click the Return or Enter key on your keyboard. The requested base will appear at the beginning of the sequence window (see ***Figure 7***).

30. Experiment by selecting a base number in your sequence.

31. Another way to find a specific base in FinchTV is to ***enter a sequence that is located near or contains your base***. In ***Figure 8***, the sequence GGTCAA was typed in the ***Find Sequence*** window and the Return key pressed. FinchTV located the sequence and highlighted it in blue.

32. Experiment with your sequence by trying to locate the sequence GGTCAA. ***Is that sequence present in your DNA sequence data?*** Record your answer in your lab notebook or on a separate sheet of paper.

Figure 7: Finding Specific Bases. Source: FinchTV.



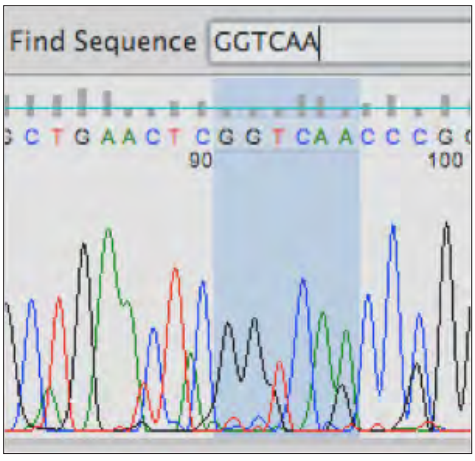


Figure 8: Finding a Specific Sequence. Source: FinchTV.

PART III: Viewing the Reverse Complement of a Sequence

Student Researcher Background:

Viewing the Complementary Sequence in FinchTV

Many times you know the sequence of bases in one strand of DNA, but wish to view that region in the complementary strand of DNA to confirm the identity or view the quality value of a specific base. To do this, you can use the methods described above for finding that base, but first you must obtain the reverse complement of the sequence in the file.

Figure 9 shows the relationship between a sequence, the complementary strand, and the sequence of the complementary strand in the "reversed" order.

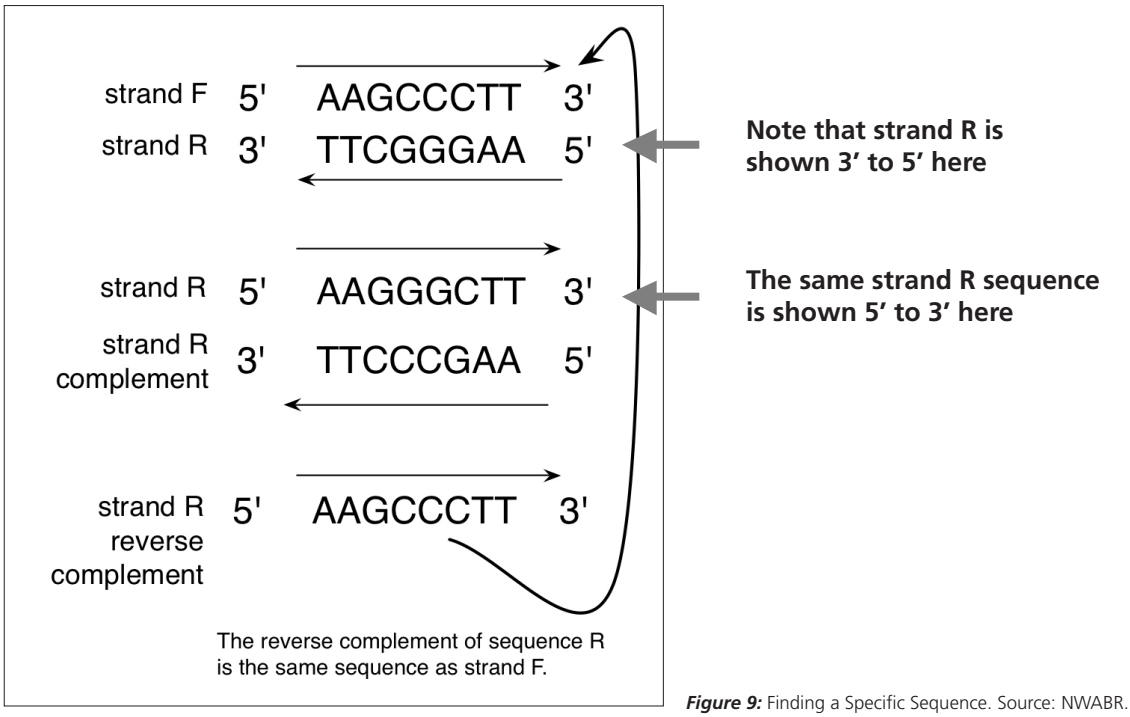


Figure 9: Finding a Specific Sequence. Source: NWABR.

Notice that the sequence of strand R, in a 5' to 3' direction, is not identical to the sequence of strand F. To see the same sequence from strand R that we see in strand F, we need to determine the complementary sequence and then view the sequence in a 5' to 3' direction. FinchTV will make that change for us.

33. Make the **reverse complement** of your **subject sequence** ("R" sequence) by clicking the ***Reverse Complement*** icon.

34. Save your reverse complement DNA sequence. Your new file name should include the ***subject sequence name*** and the phrase "RevComp" at the end, such as:

    "BOC_25a_Hippo_70248843 R RevComp."

PART IV: Reviewing Discrepancies and Recording Your Data

Student Researcher Background:

Resolving Discrepancies between Two DNA Sequences

A discrepancy in DNA sequencing is a point where the sequences from different samples or DNA strands disagree.

In this next section, you will use FinchTV to review each position where the sequences disagree and record your results. For each discrepant base, you will:

1. Find the discrepant base in the sequence file for the query sequence and record the position and quality value.
2. Find the discrepant base in the reverse complement of the subject sequence. Obtain and record the position and quality value for that base.
3. Make a decision about editing and record your decision. You may decide to keep the original base or change the sequence based on the quality values and your visual review.

There are ***three types of discrepancies*** that you are likely to see (as shown in ***Figure 10***).

a. In the most common case, the base in the subject sequence is different from the base in the query (Red box #3, ***Figure 10***).

b. Sometimes, the base may be missing in one sequence and present in another. If the base is present in the query sequence, the subject sequence will contain a dash (-) at the position of the missing base (Red box #2, ***Figure 10***).

c. If the base is present in the subject sequence, but not the query, then there will be an inserted base below the sequence line with a line above the inserted base (Red box #1, ***Figure 10***, where the T is present in the subject sequence between the two C's).

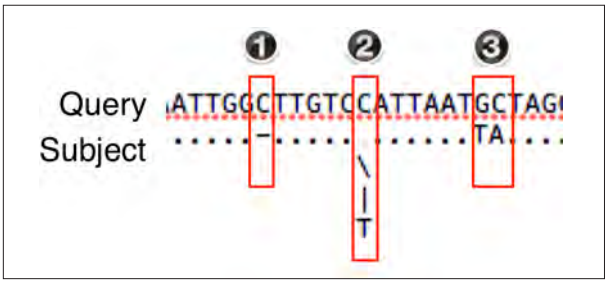


Figure 10: Three Types of Discrepancies.

The results from comparing sequences from two example chromatograms are shown below in ***Figure 11***.

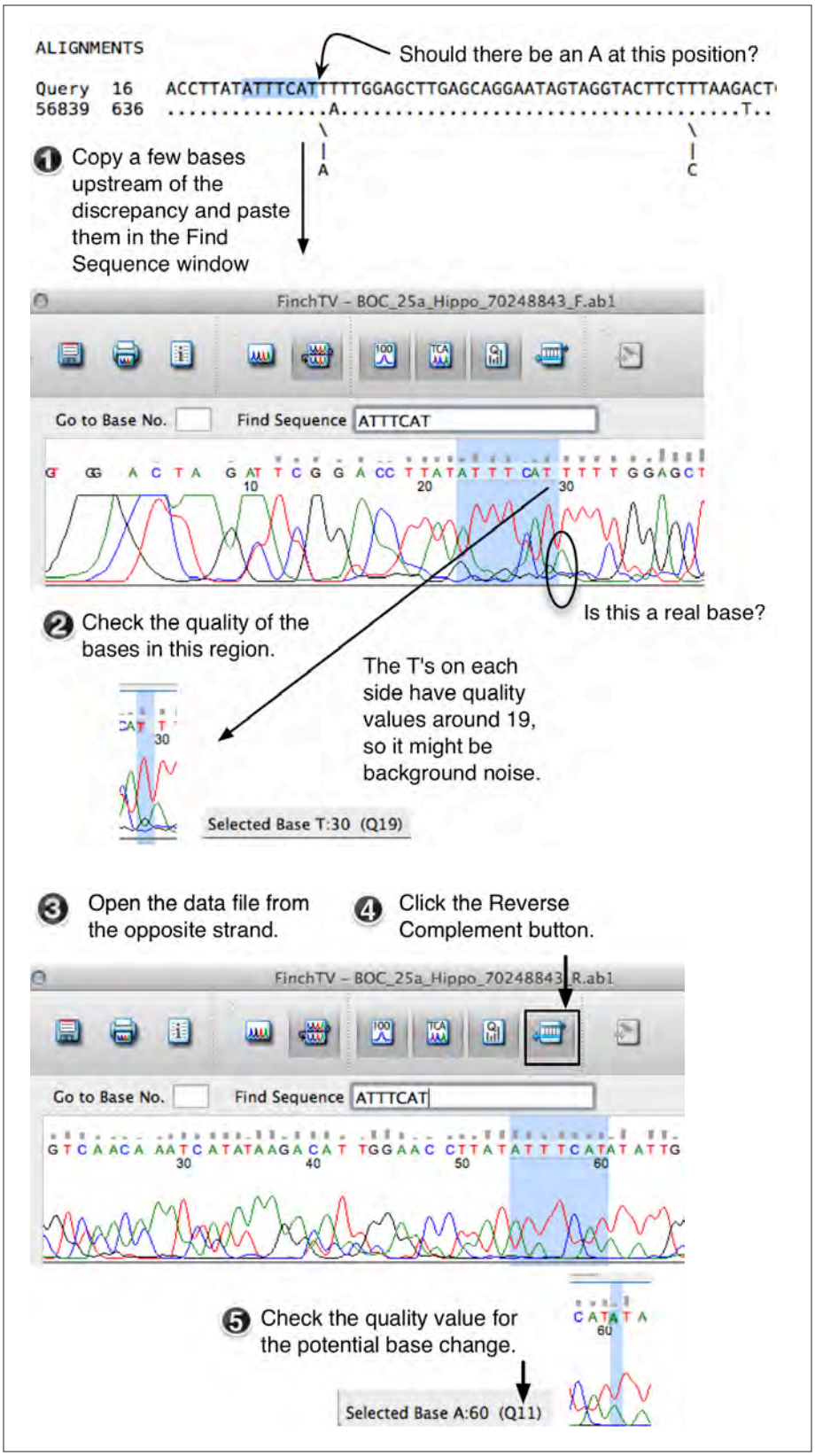


Figure 11: Analyzing Discrepancies in Your DNA Sequences.

Example Data Table for Editing DNA Sequences

Query Sequence = BOC_25a_Hippo_70248843_F
Subject Sequence = BOC 25a Hippo 70248843 RRevComp

1. Sequence position in query sequence	2. Base or note in query strand	3. Quality value	4. Sequence position in subject sequence	5. Base in subject strand	6. Quality value	7. Notes from query strand	8. Notes from subject strand	9. Editing decision	10. Reviewed by:
34	A	22	70	T	17	Good Quality	Poor Sequences	No change	SGP
75	T	13	111	C	31	Good Sequence	Fair Sequence	Replace T with C	TMS

35. For each position that you review in your portion of your group's alignment, follow the steps provided below. You may want to keep track of the reviewed bases in your alignment by using a highlighter or pen to mark each base in the alignment after review. You will record your data on Student Handout—*Data Table for Editing DNA Sequences*.

a. Open the ***query sequence*** chromatogram file ("F" sequence with the ".ab1" extension) in FinchTV. It may be helpful to view the sequence in the ***Wrapped View*** format.

b. Search for the sequence ***immediately next to the discrepancy*** you wish to analyze (see *Steps #1* and *2* in ***Figure 11*** and green box in ***Figure 12***) using the ***Find Sequence*** search box. Enter the sequence (black arrow, ***Figure 12***), and then press ***Enter*** or ***Return***.

c. Record the position number of the discrepancy in the query sequence (*Box 1* of Student Handout—*Data Table for Editing DNA Sequences*).

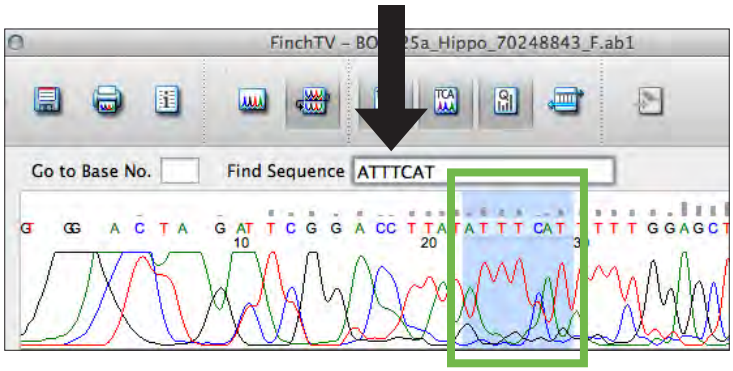


Figure 12: Finding Your Sequence. Use the ***Find Sequence*** search box to locate the region of the DNA chromatogram next the discrepancy you identified in the printout of your alignment. Source: FinchTV.

d. Record the base or note (such as a gap where a base is missing) found in the query sequence (*Box 2*).

e. Record the quality value for that base in *Box 3*.

f. Repeat *Steps b-e* for each of the discrepancies in the query sequence. If you need to make any special notes to yourself about the query sequence, you can do this in *Box 7*.

g. In FinchTV, open the chromatogram data file from the opposite strand, [this will most likely be the "R" file].

h. Click the ***Reverse Complement*** button to view the complementary sequence for the R file in the same 5' to 3' order as viewed from the F file (Step #3 and Step #4 in ***Figure 11***).

i. Search for the sequence immediately next to the discrepancy (see *Steps #1* and *2* in ***Figure 11***) using the ***Find Sequence*** search box. Be sure that you enter the correct sequence ***next to the discrepancy***. If you include the discrepancy found in the query sequence, the search will not match the reverse complement of the subject sequence.

j. Record the position number of the discrepancy in the reverse complement of the subject sequence (*Box 4*).

k. Record the base or note (such as a gap where a base is missing) in the reverse complement of the subject sequence (*Box 5*).

l. Record the quality value for that base in *Box 6* (Step #5, ***Figure 11***).

m. Repeat *Steps h-k* for each of the discrepancies in the reverse complement of the subject sequence. If you need to make any special notes to yourself about the subject sequence, you can do this in *Box 8*.

n. For each discrepancy, use the quality value data in *Boxes 3* and *6* to inform your base editing decision: ***Which base will you choose for your final sequence? Which base has the higher quality value?*** Enter this decision in *Box 9*.

o. Be sure to put your name or initials in *Box 10*. If anyone else helps you with this part, such as other group members, you can add their name, too.

PART V: Edit and Trim One of the DNA Chromatogram Files

Now it is time to update one of your DNA sequence files with the discrepancies you resolved in Student Handout—*Data Table for Editing DNA Sequences*.

36. Find the file that contains your query sequence chromatogram ("F" sequence with the "ab1" extension).

37. Make a copy of the file that contains your query sequence and rename the copy so that the new file name begins with word "Edit."

38. For each position that will be edited, use the information in Student Handout—*Data Table for Editing DNA Sequences* to locate the position and edit the sequence.
    a. To change a base in FinchTV, click that position and type the letter for the new base.
    b. To delete a base, select that base and click the delete key.
    c. To insert a base, click the position in the sequence, right click, choose ***Insert before base***, and enter the letter of the new base.

39. Save your edited file.

40. Chromatograms often contain low quality sequences at the 5′ and 3′ ends that are removed by trimming (deleting the bases). Trim your sequences by selecting the bases to be trimmed and clicking the Delete key.
    a. Trim bases from the 5′ end until the last 20 bases contain fewer than 3 bases with quality values below 10.
    b. Trim bases from the 3′ end until the last 20 bases contain fewer than 3 bases with quality values below 10.

41. Save your edited DNA chromatogram file (which will include the "ab1" extension).

PART VI: Perform a blastx Search to Translate Your Sequence and Check for Stop Codons

To be sure that extra bases were not accidentally added or removed during editing, scientists often translate their DNA sequences *in silico* (on the computer) to check for stop codons.

We can use the tool **blastx** to do this for us. Blastx will automatically translate our DNA sequence into a protein sequence.

42. Open your edited DNA chromatogram file (if it is not already open).

43. To view the DNA sequence, click the ***Chromatogram Info*** button (***Figure 13***), and a new window will appear that contains your edited DNA sequence.

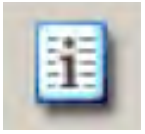


Figure 13: Chromatogram Info button

44. Open the FinchTV Edit menu and choose ***BLAST sequence***, ***Translated***, ***BLASTx***. This will open blastx at the NCBI and paste your sequence in the query box, as shown in ***Figure 14***.

45. Blastx will translate your sequence and compare it to a database of nucleotide sequences. If there are mistakes in the sequence, blastx will help identify the positions of those errors.

46. Under ***Genetic code***, select the option you think is most likely to match your organism (see ***Figure 15***). If your organism has bones, genetic code ***Vertebrate Mitochondrial (2)*** is the best choice.

47. Click ***BLAST***.

48. Review your results to identify places where there might be mistakes in the sequence. In the example below, we see a break around nucleotide position 210, indicating that sequencing errors may be located in that region (black arrow in ***Figure 16***).

49. When we look at the alignment in ***Figure 17***, we see there are many differences between our query sequence and the subject sequence (identified by blastx). This confirms that there may be a frame shift in the query sequence, either from a sequencing error or because we accidentally introduced an extra base when editing.

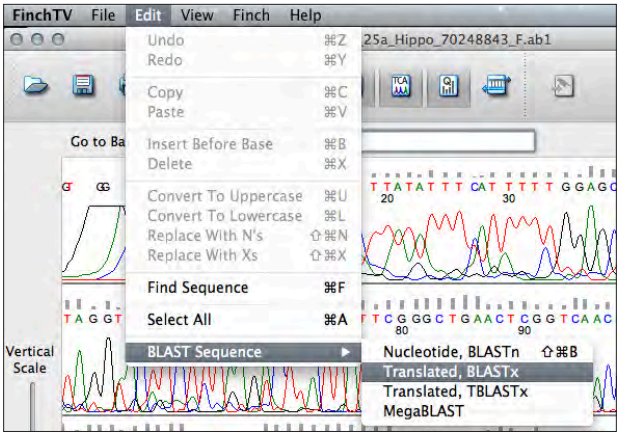


Figure 14: Choosing blastx from FinchTV. Source: FinchTV.

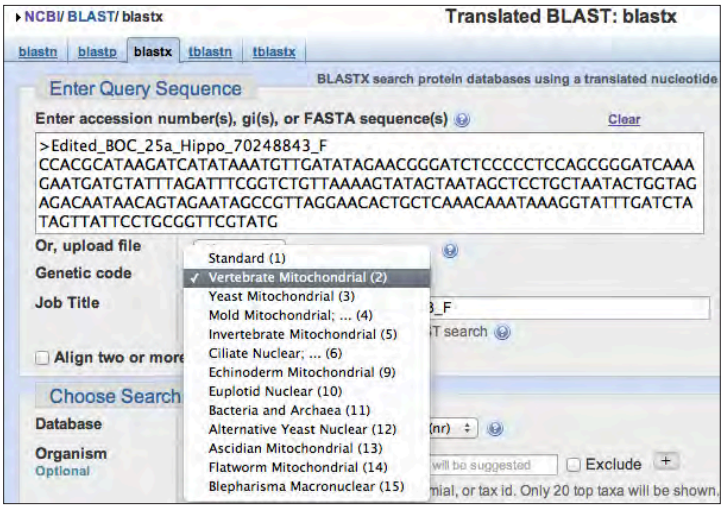


Figure 15: Using blastx to Translate DNA. Source: NCBI blastx.

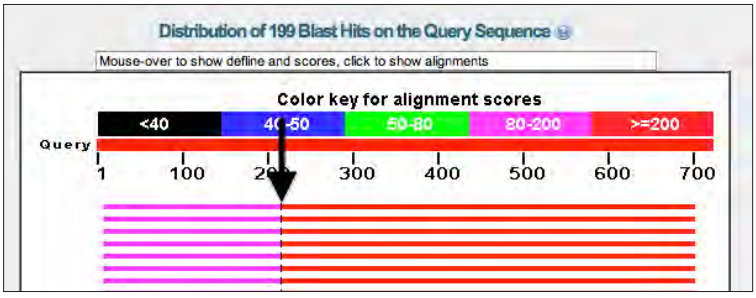


Figure 16: Blastx Translation and Search of the NCBI Protein Database Reveals Errors in DNA Editing. The pink lines represent where our sequence matches one sequence in the protein database, while the red lines indicate where the rest of our sequence matches a second sequence in the protein database. We must have introduced a frame shift at the "break" (black arrow). Source: NCBI blastx.

50. If you find any errors in your blastx alignment, go back to Student Handout—*Data Table for Editing DNA Sequences*. Review your notes from editing and look again at your chromatograms to see if an error was made. Correct any errors in your edited sequence (".ab1" file).

```
>ref|YP_004563971.1| cytochrome c oxidase subunit I, partial (mitochondrion) [Homarus
americanus]
 gb|ADP08193.1| cytochrome c oxidase subunit I [Homarus americanus]
Length=512

 GENE ID: 10743685 COX1 | cytochrome c oxidase subunit I [Homarus americanus]

                                                  Sort alignments for this subject sequence by:
                                                    E value  Score  Percent identity
                                                    Query start position  Subject start position
 Score =  223 bits (568),  Expect(2) = 1e-75, Method: Compositional matrix adjust.
 Identities = 132/163 (81%), Positives = 144/163 (88%), Gaps = 0/163 (0%)
 Frame = +3

Query  213  CPLMLGAPDMAFPRMNNM*FWllppsltllls*GMVE*GVGTGWTVYPPLSGSIAHAGAS  392
Sbjct  82   V.................S.....F......TS....S.............AA.......  141

Query  393  VDMGIFSLHLAGVSSILGAGNFMTTGINMPTARMTMNQMPLFVWTElltaillllplPVS  572
Sbjct  142  ..L.................V.....A...RSKG...DR......SVFI..V....S...L  201

Query  573  WGAITMLMAERNLNTSFFDPAGGGDPDLYQHLFWFFCHPEVYL  701
Sbjct  202  A......LTD.................V.........G.....I  244


 Score = 86.3 bits (212),  Expect(2) = 1e-75, Method: Compositional matrix adjust.
 Identities = 63/69 (91%), Positives = 66/69 (96%), Gaps = 0/69 (0%)
 Frame = +1

Query  10   IRTLYFIFGAWAGMVGTSL*LLIRAELGQPG*LIGNDQIYNVIVTAHAfvmiffmvmpim  189
Sbjct  14   .G.................S.V.........S...D......V.................  73

Query  190  iggfgNWLV  216
Sbjct  74   .........  82
```

Figure 17: Blastx Alignment Shows Many Mismatches. This indicates that there may be a frameshift in our query sequence. Source: NCBI blastx.

PART VII: Identify Your Sequence and Place It in a Phylogenetic Tree by Comparing It with Sequences in the BOLD Database

51. Open your edited DNA chromatogram file (if it is not already open).

52. View the DNA sequence by clicking on the ***Chromatogram Info*** icon.

53. Select the sequence and copy it.

54. Go to the BOLD database at http://www.barcodinglife.com. BOLD uses BLAST to compare sequences you enter to a database of sequences that meet the internationally agreed upon criteria for DNA barcoding.

55. Choose ***Identification*** from the menu at the top of the homepage.

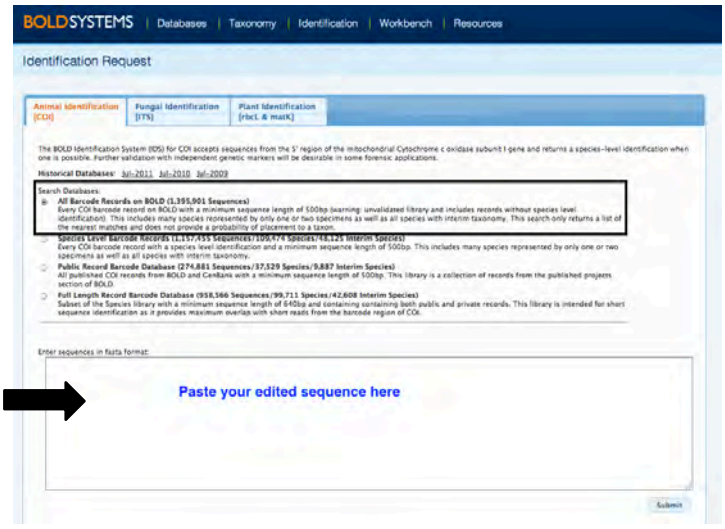


Figure 18: Entering your DNA Sequence to Identify It Using BOLD. Source: BOLD.

56. Select ***All Barcode Records on BOLD*** from the ***Search Databases*** menu (black box, ***Figure 18***).

57. Paste your sequence in the text area labeled "Enter sequence in fasta format" (black arrow, ***Figure 18***) and click ***Submit***.

No sequences in the database matched ours closely enough to be considered a match by the BOLD identification algorithms. However, we can look at the data and find related sequences. The image in black box in ***Figure 19*** shows that our sequence was 53.49% similar to a sequence from *Aphis craccivora*.

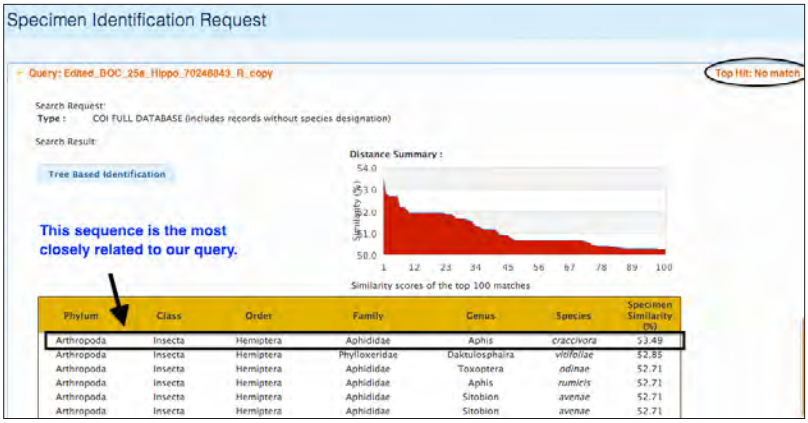


Figure 19: Results from a Search of the BOLD Database. Source: BOLD.

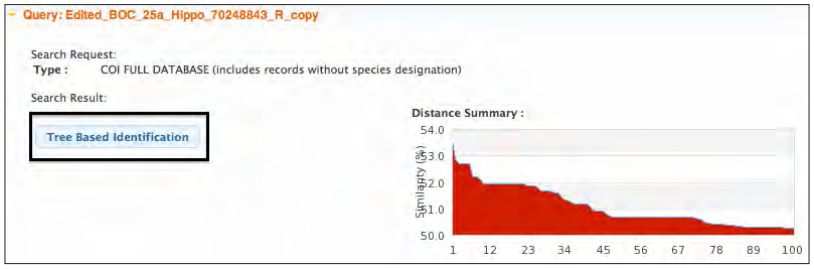


Figure 20: Select ***Tree Based Identification*** Button to See Where Your Sequence Fits in with the BOLD Phylogenetic Tree. Source: BOLD.

58. Look at your BOLD search results. ***What species matches your sequence most closely?*** What ***genus does that species belong to?*** Include the complete scientific name (Genus and species) in your lab notebook or on a separate sheet of paper.

59. The search results also provide taxonomic information about the species from which the DNA sequence was isolated, as seen in the black box in ***Figure 19***. Fill in the following information for the species that matches yours most closely, in your lab notebook or on a separate sheet of paper.

Phylum: _____

Class: _____

Order: _____

Family: _____

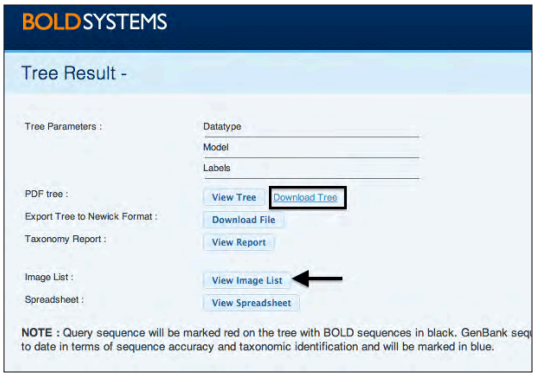


Figure 21: Select View Tree to Download PDF of Your Tree. Source: BOLD.

60. Click the ***Tree Based Identification*** button to see where your sequence fits in a BOLD-generated phylogenetic tree (black box, ***Figure 20***).

61. Select the ***Download Tree*** link to download a PDF file containing your tree (black box, ***Figure 21***). You may also wish to click the ***View Image List*** button to view images of related species.

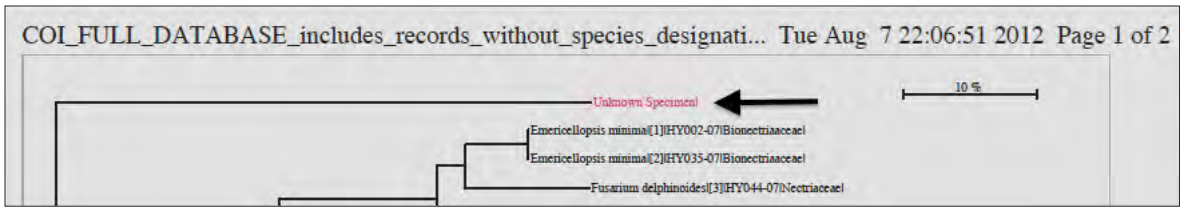


Figure 22: A Phylogenetic Tree from BOLD, with the Unknown DNA Sequence Highlighted in Red. Source: BOLD.

62. Find your sample in the tree file. Your sample will be identified in red, as seen in the example shown in ***Figure 22*** ("unknown specimen," black arrow). You will probably have to scroll to the second or third page in your PDF.

63. Copy and paste a screen capture image of your tree into a word processing document.

64. Based on what you see in your phylogenetic tree, record in your lab notebook or on a separate sheet of paper ***at least two organisms that are closely related to the species from which your sequence was obtained. Be sure to include the complete scientific name for each***.

65. Using online tools such as Google, Wikipedia, and/or the Encyclopedia of Life (http://www.eol.org), search for these closely-related organisms and ***list their common names*** in your lab notebook or on a separate sheet of paper, next to your answers to the previous question.

66. Using these same tools (Google, Wikipedia, and/or the Encyclopedia of Life), ***determine the common name of the species from which your DNA was obtained*** and record it in your lab notebook or on a separate sheet of paper.

67. Read the Wikipedia and Encyclopedia of Life entries about all three of these species. If these species are not found in Wikipedia or the Encyclopedia of Life, you may need to find other information from a Google search.

68. ***What do all of these organisms have in common? Habitat? Diet?*** In your lab notebook or on a separate sheet of paper, ***list any similarities that these organisms share. Also note any important differences, and other facts that you find interesting and/or surprising.***

69. Finally, thinking about what you have learned in all nine lessons about DNA barcoding, ***how have these lessons contributed to your understanding of the process of genetic research?*** Write your answer(s) in your notebook or on a separate sheet of paper.

Name _____ Date _____ Period _____

9 Data Table for **Editing DNA Sequences**

1. Sequence position in query sequence	2. Base or note in query strand	3. Quality value	4. Sequence position in subject sequence	5. Base in subject strand	6. Quality value	7. Notes from query strand	8. Notes from subject strand	9. Editing decision	10. Reviewed by:

9 Installing FinchTV

1. Download the FinchTV program from the Geospiza, Inc. website, available here: http://www.geospiza.com/Products/finchtv.shtml.
2. Click the ***Download it today!*** link, as shown in ***Figure 1***.
3. Complete the basic registration information, which includes your email address. A link will be sent to your email.
4. Within your email program, open the email from Geospiza, and click the link to download FinchTV, as shown in ***Figure 2***.
5. Click the link associated with your operating system, as shown in ***Figure 3***.
6. For some systems (like Microsoft® Windows) you may be prompted to open the linked file with WinZip®.
   a. You may then be prompted to purchase WinZip®. Click ***Use the Evaluation Version***, which is free.
   b. Click ***Yes*** when prompted to download the files to your folder.
   c. Click ***Next*** and then ***Unzip Now***.
   d. When unzipping is complete, click ***Finish***.
7. The file should download automatically to your program files and open. If it does not, go to your C: drive and click ***Program Files***. The FinchTV folder can be found in the "Geospiza" folder. Click the ***FinchTV.exe*** icon to launch the program.

Figure 1: Download it Today! Source: Geospiza, Inc.

Figure 2: Opening the Website Link from Your Email Program.

Figure 3: Selecting Your Operating System. Source: FinchTV.