Sequencing a Genome:

Inside the Washington University Genome Sequencing Center

Activity Supplement

From Signal to Sequence

(Electropherogram Interpretation)

Project Outline

The multimedia project *Sequencing a Genome: Inside the Washington University Genome Sequencing Center* is aimed at increasing the scientific literacy of biology students in the technology of genomic sequencing.

The following four video pieces are included on VHS cassette or CD:

- A guided tour of the Washington University Genome Sequencing Center, providing a look at the labs and offices that make up the preparation, production, and data management facilities. Includes animated explanations of the processes used to sequence genomic DNA.
- Exploration of current genomic research in pathogenic bacteria through an interview with a molecular microbiologist.
- Information about careers available at the Genome Sequencing Center presented through interviews with actual employees.
- An animated explanation of the chemistry of cycle sequencing using dideoxynucleotides.

Additional CD features include scripts of the video pieces, links to additional resources, and a glossary of terms.

As the scientific procedures presented in the video tour are complex, simple activities were specifically designed to better explain and reinforce the key concepts of restriction fragment mapping, PCR, sequencing, and electropherogram interpretation. The following activity is an inexpensive and simple solution to presenting how sequence data is derived from electropherograms.

Acknowledgments
This project was funded in part by a Professorship Award to Dr. Sarah C.R. Elgin from Howard Hughes Medical Institute (HHMI). Additional support was provided by the National Human Genome Research Institute (NHGRI) through its funding of the Washington University Genome Sequencing Center Outreach Program.

Editors
Sarah C.R. Elgin, Washington University Department of Biology
Susan K. Flowers, Washington University Science Outreach

Writers
Juanita Chambers, Saint Louis Public Schools
Carla L. Easter, Washington University Genome Sequencing Center
Susan K. Flowers, Washington University Science Outreach
Gabriella Farkas, Washington University Department of Biology
Anna Cristina Garza, biology major Washington University Arts and Sciences

Paper Model Illustrator
Gabriella Farkas, Washington University Department of Biology

Inquiries about this project may be directed to Susan Flowers, Washington University Science Outreach, Campus Box 1137, One Brookings Drive, St. Louis, Missouri 63130, flowers@wustl.edu, (314) 935-4217

ELECTROPHEROGRAM INTERPRETATION
TEACHER MANUAL

Lesson Overview

This activity is designed to provide your students with an understanding of the importance of electropherogram interpretation in genomic sequencing projects through reading, a paper activity, and completion of discussion questions. The students will read about electropherograms and the role of finishers in creating accurate DNA sequence data. They will work in pairs to score sample electropherograms. The discussion questions will ask them to define and explain terms and concepts covered in the reading and answer questions related to DNA sequencing accuracy.

Timeline

The background reading, worksheet, and discussion questions require 50 minutes to complete.

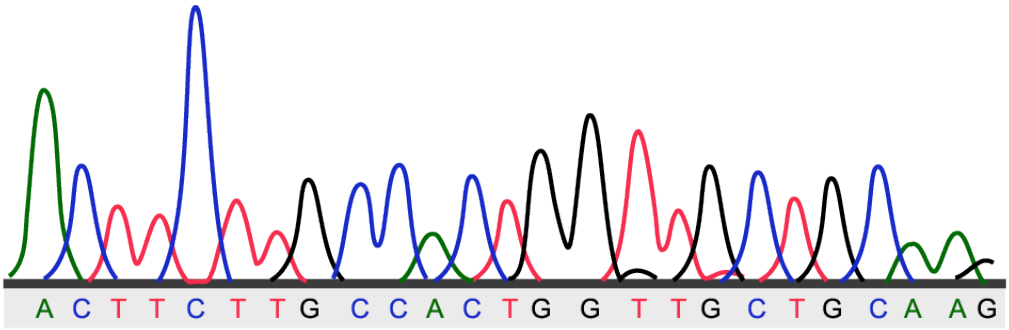
Advance Preparation

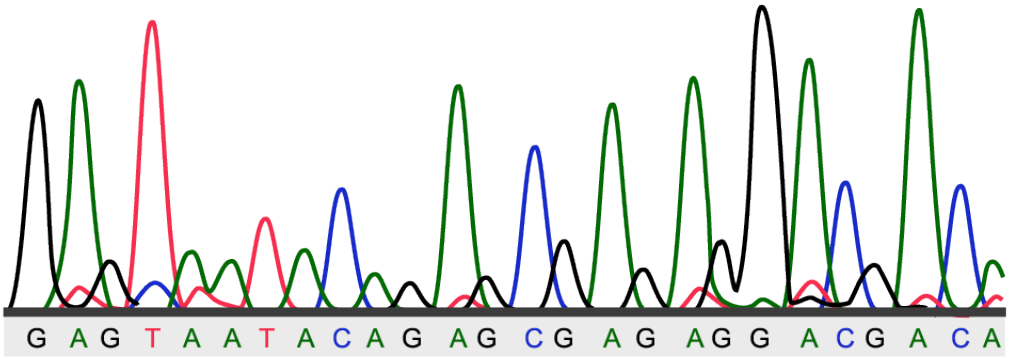
The student manual may be printed using a standard printer. You will need one manual per student if each student will be turning in a completed worksheet and discussion questions.

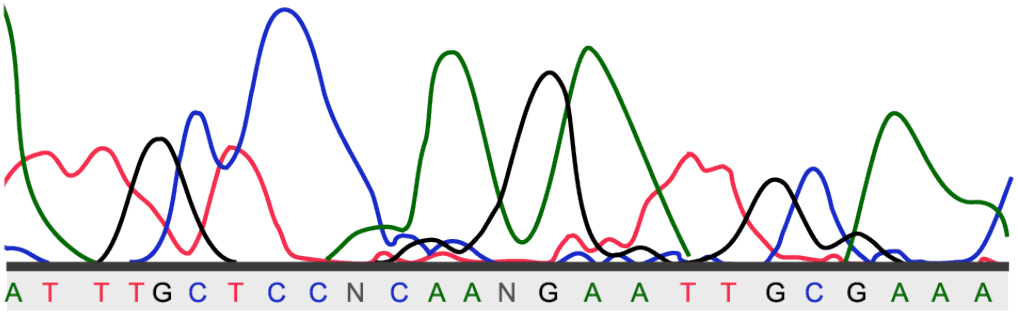
Hints and Troubleshooting

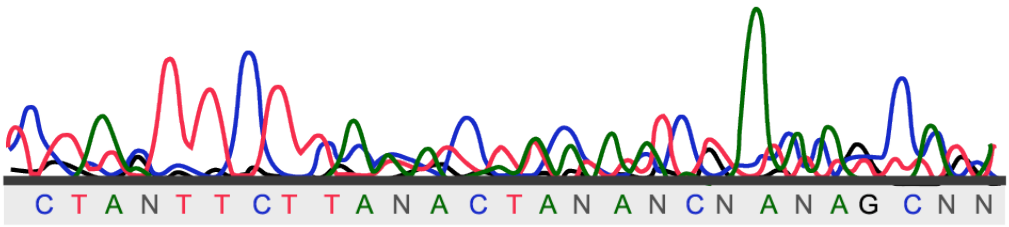
You may find it necessary to go over the background information and demonstrate one sequence reading exercise before allowing the student pairs to work on their own.

Electropherogram Interpretation Worksheet Solutions

1. 

2. 

3. 

4. 

Answers to Discussion Questions

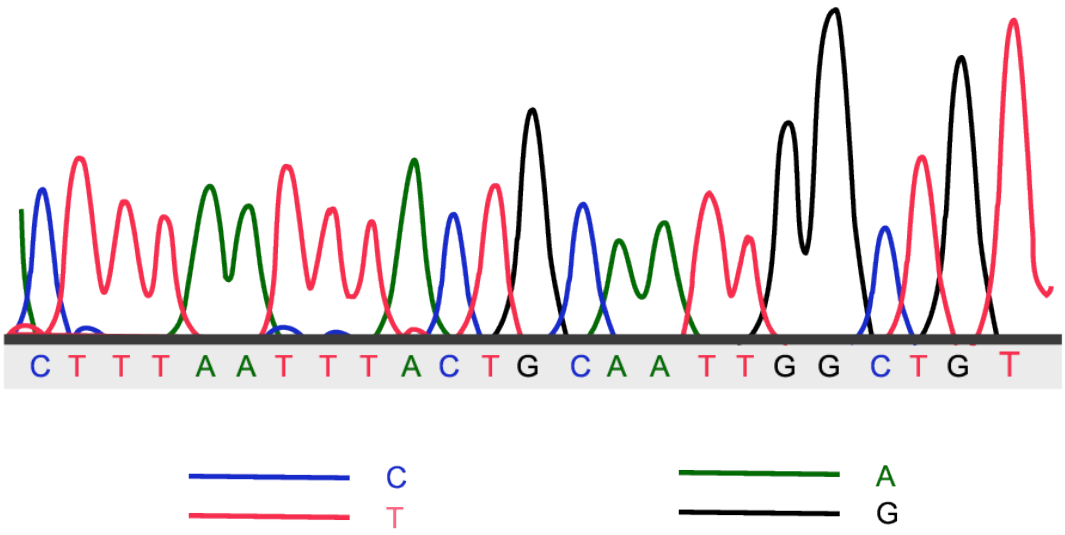
1. What is the role of a finisher in the process of DNA sequencing?
   A finisher analyzes an electropherogram in comparison to the consensus presented by the computer programs Phred and Phrap, to determine if the consensus accurately reflects the nucleotide sequence; when the data is incomplete or inadequate, the finisher calls more reactions in order to gather more data.
2. What is an electropherogram? What is a consensus sequence, and how is this different from an electropherogram?
   An electropherogram is a graphical representation of data received from a sequencing machine and yields one read. A consensus is a sequence that has been generated from the alignment of multiple reads; it is an actual sequence of letters, as opposed to a graphic image.
3. What is one cause of uncertainty in data received from an electropherogram? Explain why this particular characteristic or feature is problematic.
   Many answers are possible; all can be found in the second half of the reading. For example, it is acceptable for the student to name and explain one type of repetition. All of these are problematic for the same reason: the number of times a repetition occurs is difficult to determine, and there has to be absolute certainty that unique sequence data isn't "hidden" among the repeats. Transposons, gaps, etc. can also be mentioned, as long as the accompanying explanations seem reasonable, based on the reading.
4. Consider the activity you have just completed, and compare your results for electropherogram #1 to those for electropherogram #4. Could sequence data be accurately read from electropherogram #4? Why or why not?
   No. There are too many "N" nucleotides in the resulting sequence; the reaction should be run again to collect better data.

ELECTROPHEROGRAM INTERPRETATION
STUDENT MANUAL

Background Information

As you can probably recall from the activity "Paper Terminators," sequencing a stretch of DNA involves (1) a **dye terminator reaction**, to synthesize an array of DNA fragments of differing lengths, all complementary to the original DNA sequence of interest and ending in fluorescently labeled terminator nucleotides, and (2) gel **electrophoresis**, to separate the fluorescently labeled DNA fragments by size. Within the sequencing machine itself, dye terminator reactions are loaded into **capillaries** (one reaction per tube) containing polyacrylamide gel. As gel electrophoresis progresses, the DNA fragments in each tube migrate down the capillaries, the smaller traveling faster than the larger; this orders the fragments so that the smallest fragment exits first. As each fragment exits the capillary tube, it is hit with a laser beam that excites the fluorescent dye attached to its terminator nucleotide. A camera captures an image of this fluorescence. (Keep in mind that each of the four nucleotides has its own unique fluorescent dye and thus there are four possible fluorescent images.)

The images captured by the camera are converted to a readable form called an **electropherogram**, from which the sequence of the DNA of interest can then be determined. The electropherogram is a graphical representation of data received from a sequencing machine and is also known as a **trace**. An example of an electropherogram is shown below. (Note that in a real electropherogram, guanine would be represented by yellow, rather than black.) Each line represents one of the four nucleotides, and the peaks in the lines indicate the strength of the signal given off from the laser beam as it hits the DNA fragment. In other words, each peak corresponds to a fluorescently labeled nucleotide base, and the order of the peaks is the nucleotide sequence (because the fragments have previously been ordered by gel electrophoresis).



You might be wondering if the images captured by the camera and the electropherogram created by the computer result in a nucleotide sequence that is 100% correct and clear. Unfortunately, electropherograms are not always as high quality as the example shown to you above. The incoming signals may not be as strong and evenly spaced out. However, with the help of other computer programs and a **finisher**, the sequence often can be determined, even if the electropherogram isn't entirely clear. A finisher is someone whose job it is to analyze the raw sequence data presented by an electropherogram and create a high quality sequence by editing and calling for additional reactions to gather better data where needed. (This means that when there is simply not enough data to complete the sequence desired, even after careful analysis, the finisher informs people at the beginning of the sequencing pipeline that additional dye terminator reactions are necessary to create more, and hopefully better, data in this region of the genome.)

Before editing can begin, computer algorithms known as **Phred** and **Phrap** are used to align the data from multiple dye terminator reactions. (An algorithm is a problem-solving procedure used by the computer to perform certain tasks.) Remember, sequencing projects require multiple data sets to be generated in order for accuracy to be guaranteed. This is referred to as **coverage**. The data from a single sequencing reaction is called a **read**. Once all of the reads for a section of DNA have been aligned, another computer program called **Consed** is used to view and edit the assembled data. Phred and Phrap generate a sequence that represents the best guess at the correct sequence of the DNA of interest; this is known as the **consensus sequence**. When the data is not good enough to allow Phred and Phrap to decide a nucleotide in the sequence, an "N" appears in the consensus sequence. The **editing** done by the finisher is a visual comparison to decide if the consensus accurately reflects the sequence data for a given region of DNA. The finisher is also responsible for determining the identity of "N" nucleotides, through a comparison of different reads to one another.

Uncertainty in the nucleotide base sequence determined from an electropherogram is caused by a number of problems. Some of these are due to the nature of the DNA itself, while others are inherent to the complexity of genomic sequencing. For example, much confusion is due to the repetitive nature of DNA. Repeats can occur within the nucleotide sequence when a single nucleotide is repeated many times, or when a two- or three-nucleotide pattern is repeated. **Tandem repeats** exist where larger nucleotide sequence patterns are repeated adjacently. **Inverted repeats** occur when the same sequence is found more than once in a stretch of DNA but sometimes in reverse order. Finally, when a significantly long stretch of DNA sequence is found to exist more than once, the repetition is known as a **duplication**. Any of these kinds of repeats can be confusing during sequencing because a repeat must be sized, to determine how many copies of the pattern there are exactly, and the data must be checked to ensure that there truly is no unique sequence data within a region considered to be only repeats.

Other complications are also encountered during the complex process of sequencing. For instance, **transposons** are mobile pieces of DNA that can randomly

insert themselves into stretches of DNA. As you might recall from the "Golden Path" activity, bacterial artificial chromosomes (BACs) are commonly used as vectors in sequencing projects, and thus there is a possibility of transposons "jumping" from bacterial DNA into the DNA of interest. These must be recognized by the finisher and removed from the sequence data. There can also be regions of the DNA in which there is only sequence data for one strand of the DNA, a situation known as **single-stranded coverage**. Often times, data from sequencing projects that overlaps must be compiled, and discrepancies between these must be resolved. Regions in which there is no sequence data are called **gaps**, and when existing data cannot be compiled to close a gap, additional reactions are called by the finisher, as discussed previously.

As you can now see, computer programs and finishers play a complementary and necessary role in interpreting and analyzing data presented by electropherograms, to guarantee as much as possible that a published sequence is accurate.

Activity Overview

In this activity, you will be playing the role of Phred and Phrap, the computer programs that interpret the sequence trace files (also known as electropherograms), assign the correct nucleotide base to each corresponding electropherogram peak, and thereby generate the nucleotide sequence to be analyzed by the finisher. When you feel that a nucleotide cannot accurately be determined from the data given, assign that nucleotide as "N." By completing the Electropherogram Interpretation Worksheet, you will get an idea of how important it is for sequencing reactions to be performed carefully, and how important the finisher is in interpreting data that is even only slightly imperfect. When you have finished the worksheet, answer the accompanying Discussion Questions.

Please note: The 4 electropherograms on the worksheet are not identical and thus should not have identical nucleotide base sequences.

Name _____ Class Hour _____ Date _____

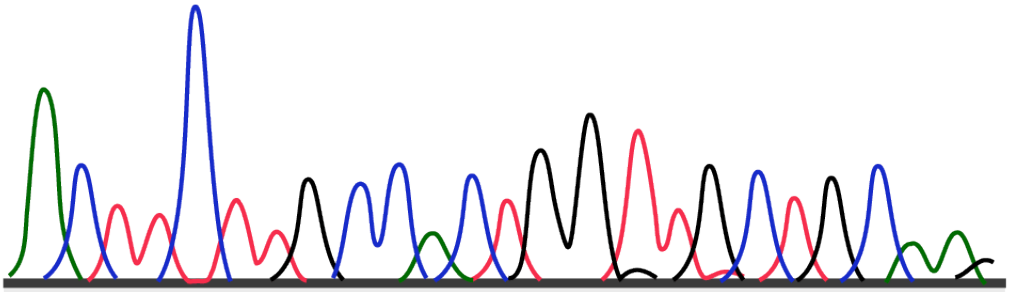
Electropherogram Interpretation Worksheet

Directions: Using the key provided, interpret the the nucleotide base sequences in the four electropherograms below. Write in each base below its corresponding peak.

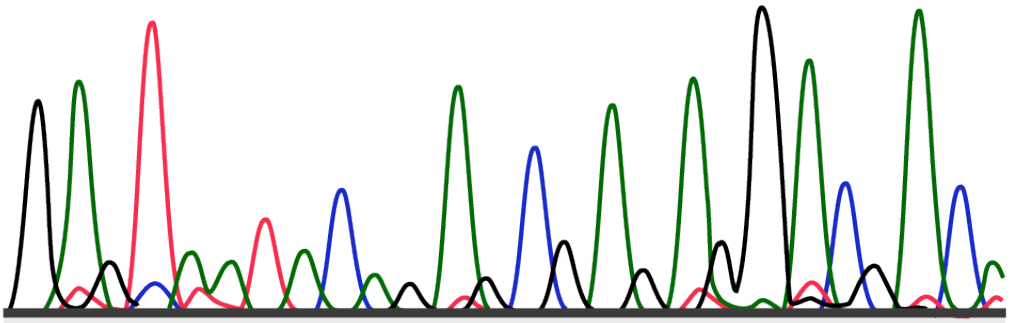
KEY

C
T
A
G

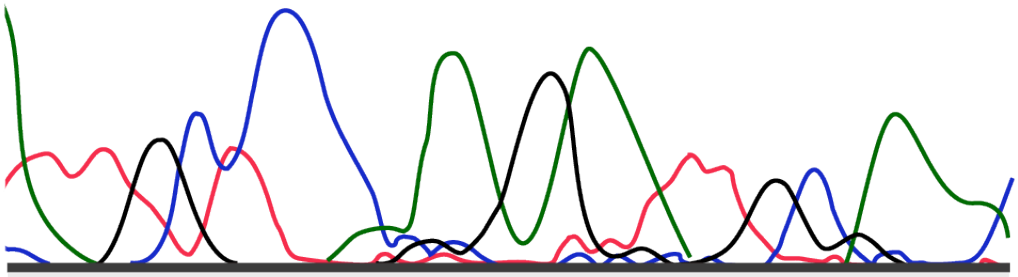
1.



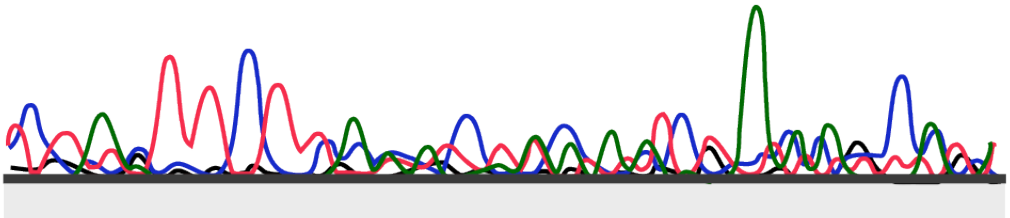
2.



3.



4.



Name _____ Class Hour _____ Date _____

From Signal to Sequence Discussion Questions

1. What is the role of a finisher in the process of DNA sequencing?

2. What is an electropherogram? What is a consensus sequence, and how is this different from an electropherogram?

3. What is one cause of uncertainty in data received from an electropherogram? Explain why this particular characteristic or feature is problematic.

4. Consider the activity you have just completed, and compare your results for electropherogram #1 to those for electropherogram #4. Could sequence data be accurately read from electropherogram #4? Why or why not?