The Wolbachia Project

Discover the Microbes Within!

Lab 5: Bioinformatics I

Sanger Sequence Analysis

Project Guide

The *Wolbachia* Project

1. Arthropod Identification
2. DNA Extraction
3. PCR
4. Gel Electrophoresis
5. Bioinformatics

The *Wolbachia* Project: Discover the Microbes Within! was developed by a collaboration of scientists, educators, and outreach specialists. It is directed by the Bordenstein Lab at Vanderbilt University.

https://www.vanderbilt.edu/wolbachiaproject

Activity at a Glance

Goal

To analyze and interpret the quality of Sanger sequences, and generate a consensus DNA sequence for bioinformatics analyses.

Learning Objectives

Upon completion of this activity, students will (i) understand the Sanger method of sequencing, also known as the chain-termination method; (ii) be able to interpret chromatograms; (iii) evaluate sequencing Quality Scores; and (iv) generate a consensus DNA sequence based on forward and reverse Sanger reactions.

Prerequisite Skills

While no computer programming skills are necessary to complete this work, prior exposure to personal computers and the Internet is assumed.

Teaching Time: One class period

Recommended Background Tutorials

- DNA Learning Center Animation: Sanger Method of DNA Sequencing (https://www.dnalc.org/view/15479-sanger-method-of-dna-sequencing-3d-animation-with-narration.html)
- YouTube video: The Sanger Method of DNA Sequencing (https://www.youtube.com/watch?v=FvHRio1yyhQ)
- Khan Academy: DNA Sequencing (https://www.khanacademy.org/science/high-school-biology/hs-molecular-genetics/hs-biotechnology/a/dna-sequencing)

Required Resources

- Computer with internet browser, such as Firefox or Chrome
- DNA analysis software, such as SnapGene Viewer* - https://www.snapgene.com/snapgene-viewer/
- DNA Sequence Files: https://www.vanderbilt.edu/wolbachiaproject/lab-5-dna-sequences/

* Multiple software options are available for DNA sequence analysis. SnapGene Viewer is highlighted here due to its user-friendly interface and cross-platform accessibility. Another highly recommended tool is MEGA X (https://www.megasoftware.net/home), although a MacOS version was not available during the development of this lab activity.

Technical Overview

File Extensions

- **.ab1** (ABI sequencer data file): Known as the *trace file*, it includes raw data that has been output from Applied Biosystems' Sequencing Analysis Software. **.ab1** files include quality information about the base calls, the chromatogram (also called the electropherogram), and the DNA sequence.
- **.scf** (Standard Chromatogram Format): Like .**ab1** files, .**scf** files are also *trace files* that include quality information about the base calls, the chromatogram (also called the electropherogram), and the DNA sequence.
- **.seq**: Known as the *sequence file*, it is a plain text file containing the DNA sequence.
- **.fasta:** A text-based format for representing either nucleotide or peptide sequences. The file often starts with a description or header line that begins with '>' and provides information about the sequence.

Quality Scores

Quality scores indicate the probability that an individual base is called incorrectly during DNA sequencing. For this lab, we recommend a Q score ≥ 40.

Q score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%

Module 1: Analyzing Sanger Sequences

Materials

Example Sanger sequences:

- ❑ Example1-WSpecF.ab1 (forward)
- ❑ Example2-WSpecR.ab1 (reverse)
- ❑ Example3-HCO2198.ab1 (reverse)
- ❑ Example4-LCO1490.ab1 (forward)
- ❑ Your sequences, if applicable

Computer with:

- ❑ SnapGene Viewer Software
- ❑ Internet Access (NCBI)

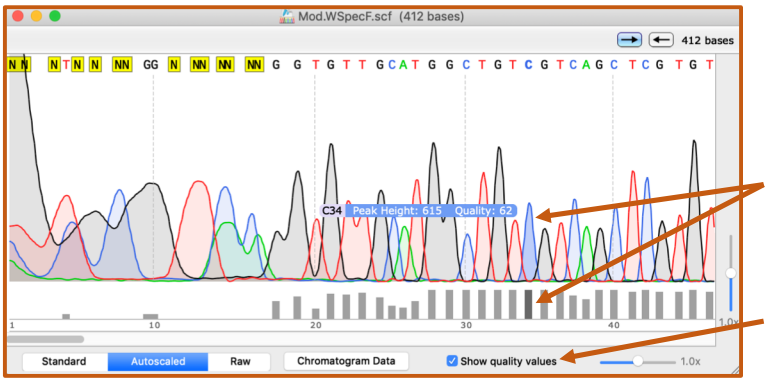
Getting Started

1. Download Example Sanger sequences to a folder on your desktop (see **Note 5.1**).
   - https://www.vanderbilt.edu/wolbachiaproject/lab-5-dna-sequences/
2. Download SnapGene Viewer to your computer:
   - https://www.snapgene.com/snapgene-viewer/
3. Open SnapGene Viewer.

Note 5.1: *For arthropod sequences, recall that LCO1490 is the forward primer and HCO2198 is the reverse primer.*

Edit the Forward Trace File

4. Select Open >> Open Files >> Example1-WSpecF.ab1.
5. Select File >> Save As >> Mod.WSpecF.scf (or name of your preference). Rename the sequence to create a copy of the original .ab1 file (See **Note 5.2**). Since SnapGene Viewer does not have an .ab1 option, use the comparable .scf extension.
6. Select "Show quality values" in the lower right hand corner. The bars correlate to quality score. Hover the cursor over each bar to visualize the quality score.



7. Use the bottom scroll bar to scan the sequence. Confirm that the majority of the sequence contains unique peaks with quality values ≥ 40.
8. The ends of the sequence will likely be low-quality (as seen above). Therefore, it is necessary to trim/delete poor base calls. Beginning at the 5'-end (left), identify the beginning of the "high quality sequence" (see **Note 5.3**).
9. Using the cursor, highlight ALL bases prior to this sequence.
   For this example, we will apply the most conservative guidelines and select for the contiguous sequence with ≥ 40 quality scores. Therefore, we will trim the first 54 bases.
10. Hit 'Delete.'
    ONLY trim from the ends, not the interior portion of the sequence!
11. Repeat steps #7-9 for the 3'-end (right).
    Applying the most conservative guideline, we will trim the last 17 bases.

Note 5.2: *Always make a copy of raw data prior to editing. Keep the originals in case you make a mistake along the way or need to refer to the raw sequences in the future.*

Note 5.3*: Determining where to trim based on the chromatogram and quality scores requires some personal judgement. For example, scroll to base 54. It has a distinct 'T' peak, but the Quality Score is only 21. According to our ≥ 40 cutoff, there are three possible options ranging from most to least conservative:*

- *Trim everything before base 55.*
- *Include this region, but change to 'T' to 'N'*
- *Perform an alignment; if the complementary strand is 'A', keep the 'T' base call.*

Most importantly, maintain consistency throughout your analysis. Define guidelines, record them in your lab notebook, and apply them to all sequences.

Continued on page 6...

Module 1: Analyzing Sanger Sequences

Edit the Forward Trace File

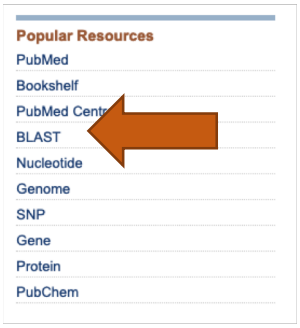
12. Scroll through the sequence. Are all quality scores ≥ 40?
    If there is a low-value base call in the middle of the sequence, DO NOT DELETE. Use your judgment here. Does the peak look unique? Is the value near 40 (i.e., 37-39)? If yes, you can leave as is. If you are not confident with this base call, highlight with your cursor and type 'N'. This will replace the base call with 'N', indicating that the exact base is unknown.
13. Select File >> Export >> FASTA Format.
14. Check your folder. You should now have 3 files for this sequence: the original trace file (.ab1), the modified trace file (.scf), and the FASTA file.

Edit the Reverse Trace File – repeat the same steps from the forward file

15. In SnapGene Viewer, select Open >> Open Files >> Example2-WSpecR.ab1.
16. Select File >> Save As >> Mod.WSpecR.scf (or name of your preference).
17. Select "Show quality values" in the lower right hand corner.
18. Use the bottom scroll bar to scan the sequence. Confirm that the majority of the sequence contains unique peaks with quality values ≥ 40.
19. Beginning at the left, identify the beginning of the "high quality sequence."
20. Using the cursor, highlight ALL bases prior to this sequence.
21. Hit 'Delete.'
22. Repeat steps #19-21 for the right end of the sequence.
23. Scroll through the sequence. Are all quality scores ≥ 40?
24. Select File >> Export >> FASTA Format.
25. Check your folder. You should now have 3 files for this sequence: the original trace file (.ab1), the modified trace file (.scf), and the FASTA file.

Generate a Consensus Sequence

26. Open NCBI in your web browser: https://www.ncbi.nlm.nih.gov/
27. Select "BLAST" from the right-hand 'Popular Resources' menu
28. Select "Nucleotide BLAST."
29. *(optional)* Enter a Job Title.
30. Click " Align two or more sequences" at the bottom of the first box.
31. Load your forward FASTA file in the top box and the reverse FASTA file in the second box. Hit BLAST.
    - *The lower box shows the alignment of the two sequences.*
    - *Bases that are gray and lower case indicate low complexity regions.*
32. Check the % Identity. It should be 100%. If not, refer back to the trace files and investigate the discrepancy.
33. If your identity is 100%, select the Arrow next to "Download" and download FASTA (aligned sequences). Save. You have now generated a **Consensus Sequence**. You will use this high quality DNA sequence in Part 2 of the Bioinformatics Lab.



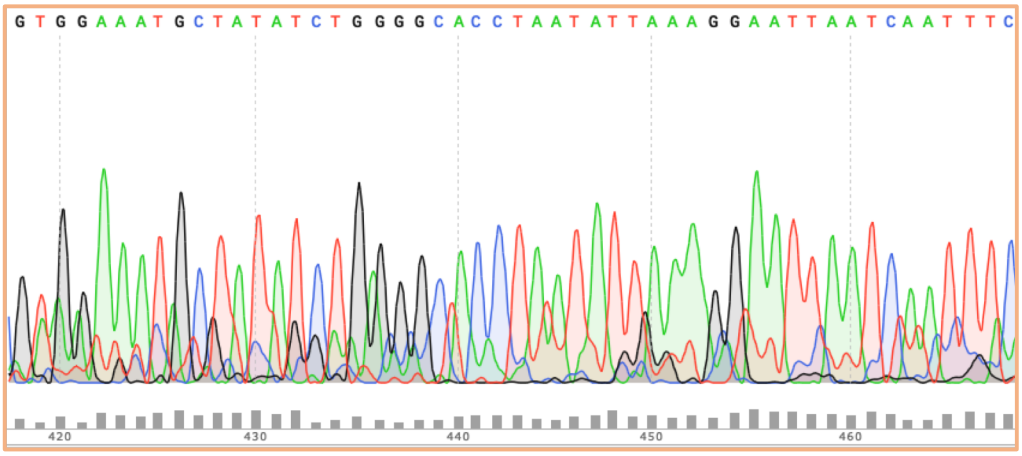
See Illustrated BLAST Analysis on page 8.

Analyze a "Less than Perfect" Sanger Sequence

34. In SnapGene Viewer, select Open >> Open Files >> Example3-HCO2198.ab1.
35. Select File >> Save As >> Mod.HCO2198.scf (or name of your preference).
36. Select "Show quality values" in the lower right hand corner.
    Notice the low Q scores.

Continued on page 7…

Module 1: Analyzing Sanger Sequences



Analyze a "Less than Perfect" Sanger Sequence

37. Not only are Q scores < 40, but there are also multiple peaks for each base call.
    - *This is a low-quality sequencing run and base calls should not be trusted.*
38. Open the complementary strand In SnapGene Viewer: select Open >> Open Files >> Example4-LCO1490.ab1.
39. Select File >> Save As >> Mod.LCO1490.scf (or name of your preference).
40. Select "Show quality values" in the lower right hand corner.
    - *This sequence is a better run and can be trusted.*
41. Repeat steps 7-14 for this sequence.
42. You may use this sequence for downstream bioinformatics analyses, but note that the sequence was obtained from a single direction.

You are now ready to repeat steps 1-33 with your own sequences.

What causes a low-quality run?

Many variables might contribute to a low-quality Sanger run. These include, but are not limited to:

- Non-specific primer binding
- Amplification of both the COI gene and nuclear mitochondrial pseudogenes (numts)
- Not enough DNA template
- DNA degradation (refer to nuclease discussion in Lab 2)
- Inhibitory contaminants (refer to salt discussion in Lab 2)

Are both forward and reverse sequences necessary?

This depends on the overall goal of your project.

- If the sequences are intended for an online data repository or publication, we highly recommend both forward and reverse reactions to validate quality.
- If the sequences are part of an informal and/or pilot study, you may use sequence data from one direction. However, ALWAYS present data in the context of a single sequencing run that was not verified with both primers.

Continued on page 8...

Module 1: Analyzing Sanger Sequences

Illustrated BLAST Analysis

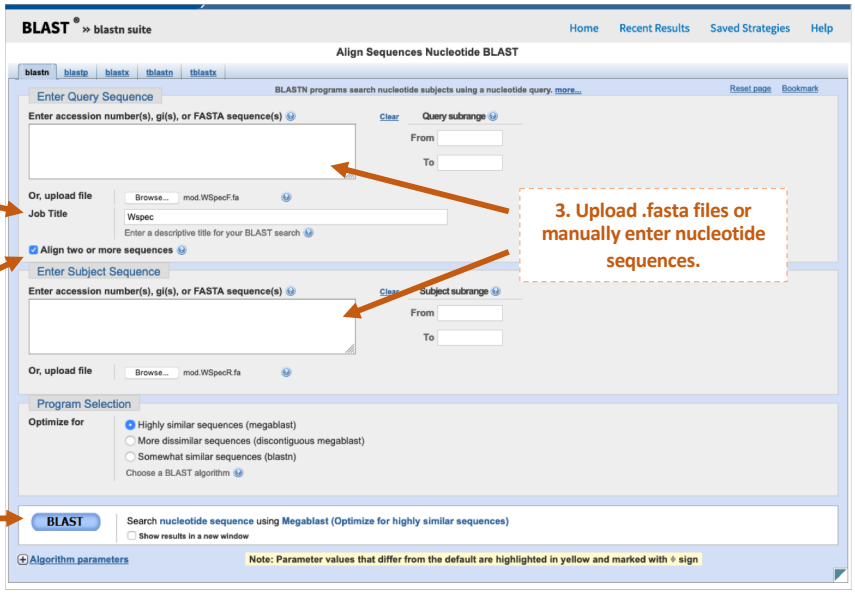
QUERY

1. Enter "Job Title" *(optional)*

2. Check "Align two or more sequences"

3. Upload .fasta files or manually enter nucleotide sequences.

4. Select "BLAST"



RESULTS

5. Confirm that identity is 100%

6. Download "FASTA (aligned sequence)"

6. The "Consensus Sequence" is the aligned 296 bases that are shared between forward and reverse strands.