



Lab 5: Bioinformatics I

Sanger Sequence Analysis

Project
Guide

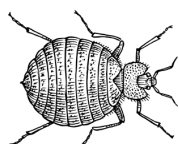


Table of Contents

Page	Contents
3	Activity at a Glance
4	Technical Overview
5-6	Part 1: Analyze a Sanger Sequence
7	Part 2: Generate a Consensus Sequence
8	Illustrated BLAST Alignment
9	Part 3: Analyze a “Less than Perfect” Sanger Sequence
10	Database Entry



Content is made available under the Creative Commons Attribution-NonCommercial-No Derivatives International License. Contact (wolbachiaproject@vanderbilt.edu) if you would like to make adaptations for distribution beyond the classroom.



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

Goals

- To analyze and interpret the quality of Sanger sequences
- To 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>).

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%

Consensus Sequence

Recall that DNA is double-stranded. Therefore, sequencing a region of DNA involves two Sanger reactions: forward and reverse. Sanger sequencing the forward strand uses only the forward primer (the same forward primer used for PCR) while sequencing the reverse strand uses only the reverse primer (the same reverse primer used for PCR). This lab activity will walk through the analysis of each sequence separately and then illustrate how to generate a consensus sequence. To generate the consensus, you will perform an alignment of both the forward and reverse sequences to confirm that bases are complementary. If the alignment is not 100% homologous, you should investigate the discrepancy in the original chromatogram files.

Is sequencing both strands required? No. Oftentimes, only one direction is sequenced because it is much more cost-effective. If the Sanger sequencing run was successful and quality scores are >40 , this DNA sequence can be trusted. Most *Wolbachia* Project participants sequence only the forward strand.

Part 1: Analyze a Sanger Sequence

MATERIALS

Example Sanger sequence:

- Example-WSpecF.ab1

Computer with:

- SnapGene Viewer Software
- Internet Access (NCBI)

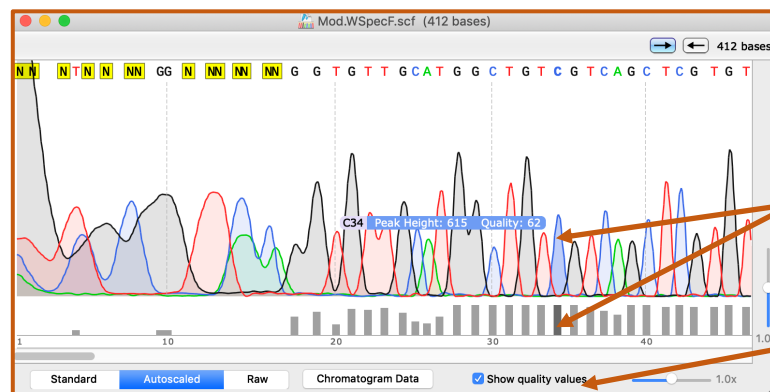
Getting Started

1. Download Example Sanger sequences to a folder on your desktop.
 - <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.

Edit a Forward Trace File

4. Select Open >> Open Files >> **Example-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. Since SnapGene Viewer does not have an .ab1 option, use the comparable .scf extension.

Note: 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.
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.



The Quality Score for this base call is 62.

Turn on Quality Values

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.”

Note: *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 (see Part 2); 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.

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; NEVER trim the interior portion of the sequence!
11. Repeat steps #8-10 for the 3'-end (right).
Applying the most conservative guideline, we will trim the last 17 bases.
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.

Part 2: Generate a Consensus Sequence

MATERIALS

Example Sanger sequences:

- Example-WSpecF.ab1
- Example-WSpecR.ab1

Computer with:

- SnapGene Viewer Software
- Internet Access (NCBI)

Edit the Forward Trace File

Download Example Sanger sequences to a folder on your desktop.

- <https://www.vanderbilt.edu/wolbachiaproject/lab-5-dna-sequences/>

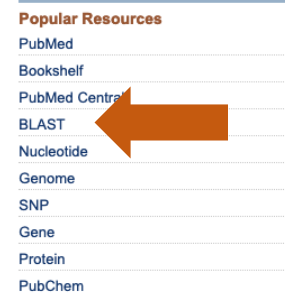
Complete Part 1 to generate a WSpecF forward FASTA file.

Edit the Reverse Trace File – repeat the same steps from Part 1

1. In SnapGene Viewer, select Open >> Open Files >> **Example-WSpecR.ab1**.
2. Select File >> Save As >> Mod.WSpecR.scf (or name of your preference).
3. Select “Show quality values” in the lower right-hand corner.
4. Use the bottom scroll bar to scan the sequence. Confirm that the majority of the sequence contains unique peaks with quality values ≥ 40 .
5. Beginning at the left, identify the beginning of the “high quality sequence.”
6. Using the cursor, highlight ALL bases prior to this sequence.
7. Hit ‘Delete.’
8. Repeat steps #19-21 for the right end of the sequence.
9. Scroll through the sequence. Are all quality scores ≥ 40 ?
10. Select File >> Export >> FASTA Format.
11. 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

12. Open NCBI in your web browser: <https://www.ncbi.nlm.nih.gov/>
13. Select “BLAST” from the right-hand ‘Popular Resources’ menu
14. Select “Nucleotide BLAST.”
15. (optional) Enter a Job Title.
16. Click “Align two or more sequences” at the bottom of the first box.
17. 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.*
18. Check the % Identity. It should be 100%. If not, refer to the trace files and investigate the discrepancy.
19. If your identity is 100%, select the Arrow next to “Download” and download FASTA (aligned sequences). Save. You have now generated a **Consensus Sequence**.



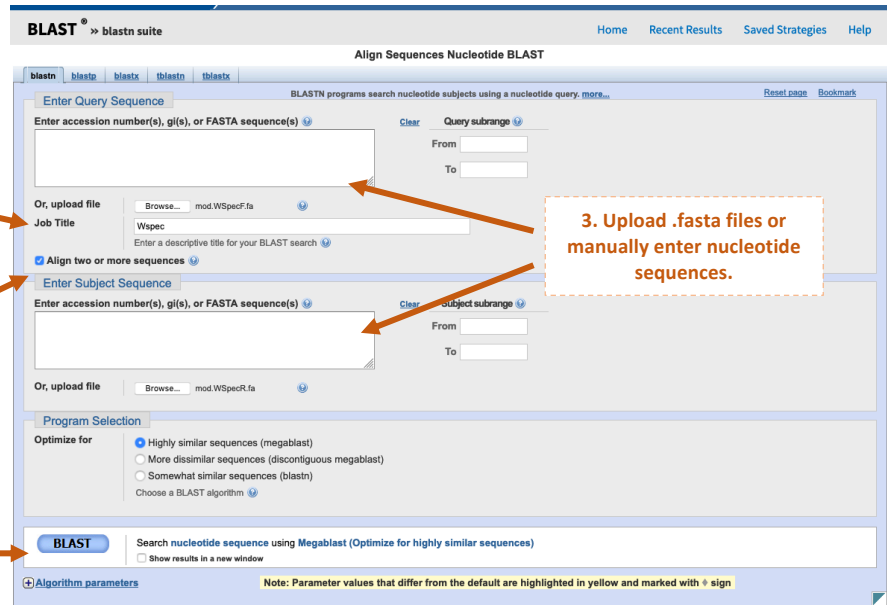
Illustrated BLAST Alignment

QUERY

1. Enter "Job Title" (optional)

2. Check "Align two or more sequences"

4. Select "BLAST"

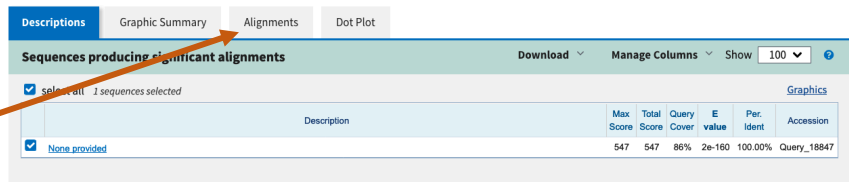


The screenshot shows the BLAST search page. Annotations include:

- An arrow pointing to the 'Job Title' field with the text '1. Enter "Job Title" (optional)'.
- An arrow pointing to the 'Align two or more sequences' checkbox with the text '2. Check "Align two or more sequences"'.
- An arrow pointing to the 'BLAST' button with the text '4. Select "BLAST"'.
- A dashed box around the 'Or, upload file' section with the text '3. Upload .fasta files or manually enter nucleotide sequences.'

RESULTS

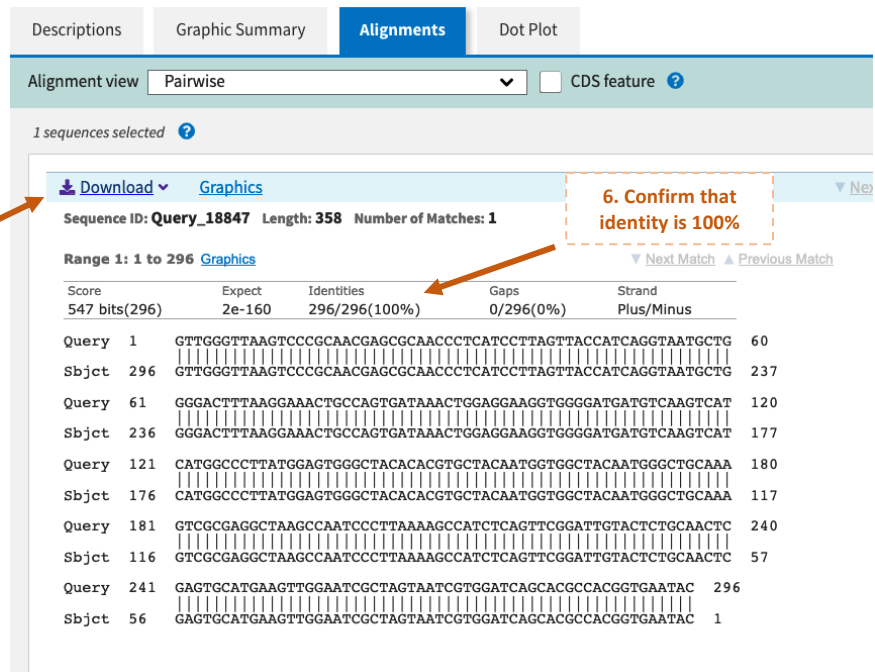
5. Select "Alignments" tab



The screenshot shows the 'Alignments' tab selected. A table titled 'Sequences producing significant alignments' is visible. The table has columns: Description, Max Score, Total Score, Query Cover, E value, Per. Ident, and Accession. One row is shown with 'None provided' in the description and '547' in the Max Score column.

7. Download "FASTA (aligned sequence)" and "Save"

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



The screenshot shows the 'Alignments' view for 'Pairwise' alignment. Annotations include:

- An arrow pointing to the 'Download' button with the text '7. Download "FASTA (aligned sequence)" and "Save"'.
- A dashed box around the 'Identity: 296/296(100%)' with the text '6. Confirm that identity is 100%'.
- A bracket on the right side of the alignment table with the text '8. The "Consensus Sequence" is the aligned 296 bases that are shared between forward and reverse strands.'

Score	Expect	Identities	Gaps	Strand
547 bits(296)	2e-160	296/296(100%)	0/296(0%)	Plus/Minus
Query 1	GTTGGGTTAAGTCCCGCAACGAGCGCAACCCCTCATCCTTAGTTACCATCAGGTAATGCTG	60		
Sbjct 296	GTTGGGTTAAGTCCCGCAACGAGCGCAACCCCTCATCCTTAGTTACCATCAGGTAATGCTG	237		
Query 61	GGGACTTTAAGGAAACTGCCAGTGATAAACTGGAGGAAGGTGGGGATGATGTCAGTCAAT	120		
Sbjct 236	GGGACTTTAAGGAAACTGCCAGTGATAAACTGGAGGAAGGTGGGGATGATGTCAGTCAAT	177		
Query 121	CATGGCCCTTATGGAGTGGGCTACACACGTGCTACAATGGTGGCTACAATGGGCTGCAAA	180		
Sbjct 176	CATGGCCCTTATGGAGTGGGCTACACACGTGCTACAATGGTGGCTACAATGGGCTGCAAA	117		
Query 181	GTCGGAGGCTAAGCCAATCCCTTAAAAGCCATCTCAGTTCGGATTGTACTCTGCAACTC	240		
Sbjct 116	GTCGGAGGCTAAGCCAATCCCTTAAAAGCCATCTCAGTTCGGATTGTACTCTGCAACTC	57		
Query 241	GAGTGCATGAAGTTGGAATCGCTAGTAAATCGTGGATCAGCACGCCACGGTGAATAC	296		
Sbjct 56	GAGTGCATGAAGTTGGAATCGCTAGTAAATCGTGGATCAGCACGCCACGGTGAATAC	1		

Part 3: Analyze a “Less than Perfect” Sanger Sequence

MATERIALS

Example Sanger sequences:

- Example-CO1R.ab1
- Example-CO1F.ab1

Computer with:

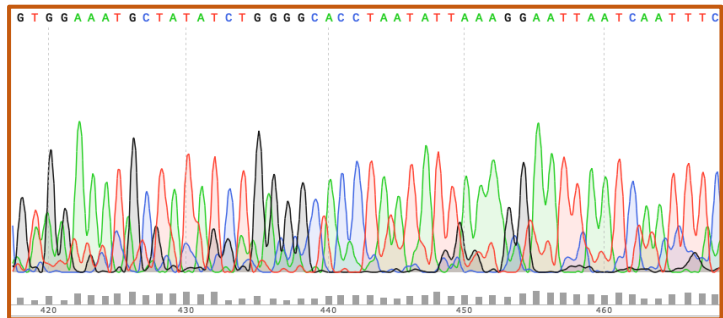
- SnapGene Viewer Software
- Internet Access (NCBI)

Getting Started

Download Example Sanger sequences to a folder on your desktop.

- <https://www.vanderbilt.edu/wolbachiaproject/lab-5-dna-sequences/>

1. In SnapGene Viewer, select Open >> Open Files >> **Example-CO1R.ab1**.
2. Select File >> Save As >> Mod.CO1R.scf (or name of your preference).
3. Select “Show quality values” in the lower right-hand corner. *Notice the low Q scores.*
4. 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.*
5. Because the Sanger sequence is poor quality, we will need to perform Sanger sequencing using the other PCR primer (in this case, use the forward primer).
6. Open the complementary (forward) strand In SnapGene Viewer: select Open >> Open Files >> **Example-CO1F.ab1**.
7. Select File >> Save As >> Mod.CO1F.scf (or name of your preference).
8. Select “Show quality values” in the lower right-hand corner.
This sequence is a better run and can be trusted.
9. Follow the steps in Part 1 for this sequence.



Many variables can cause a low-quality run including, but not limited to:

- Non-specific primer binding
- Contamination with other samples during DNA extraction and/or PCR
- Arthropod-specific: Amplification of both the COI gene and nuclear mitochondrial pseudogenes (numts)
- *Wolbachia*-specific: The arthropod is infected with more than one *Wolbachia* strain (co-infection)
- Not enough DNA template
- DNA degradation
- Inhibitory contaminants (salts, enzymes)

Database Entry

After completing this lab, analyze your own Sanger sequences and complete corresponding entries in The *Wolbachia* Project Database. A comprehensive guide is located under the Resources tab.

<https://wolbachiaprojectdb.org/>

Database fields to review and update

- Wolbachia* positive?
- Confidence level
- Explain your confidence level

Database fields to complete

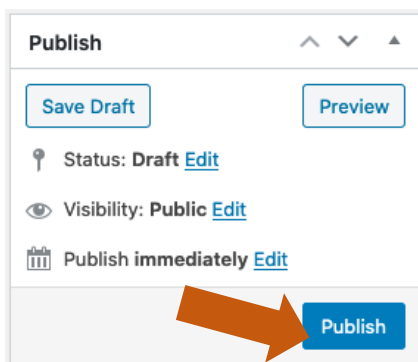
If *Wolbachia*-positive and DNA was sequenced:

- Wolbachia* 16S .fasta upload
- Wolbachia* 16S .ab1 upload
- Wolbachia* 16S sequence (text)

If Arthropod DNA was sequenced:

- Arthropod 16S .fasta upload
- Arthropod 16S .ab1 upload
- Arthropod 16S sequence (text)

Share with the world



Publish

Save Draft Preview

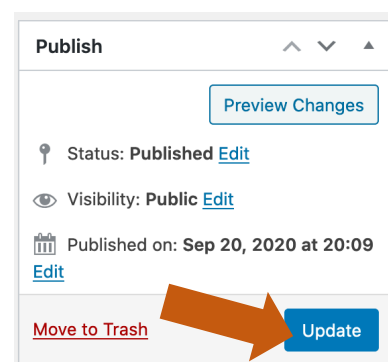
Status: Draft [Edit](#)

Visibility: Public [Edit](#)

Publish immediately [Edit](#)

Publish

**Publish
or
Update**



Publish

Preview Changes

Status: Published [Edit](#)

Visibility: Public [Edit](#)

Published on: Sep 20, 2020 at 20:09 [Edit](#)

[Move to Trash](#) **Update**