



FOOD SAFETY LAB / MILK QUALITY IMPROVEMENT PROGRAM

Standard Operating Procedure

Title: **Automated Editing of DNA Sequences using SeqTrace**

SOP #:

Revision: **00**

Revision Date: **n/a**

Effective Date: **2019-10-11**

Authors: **Sam Reichler and Jordan Skeens**

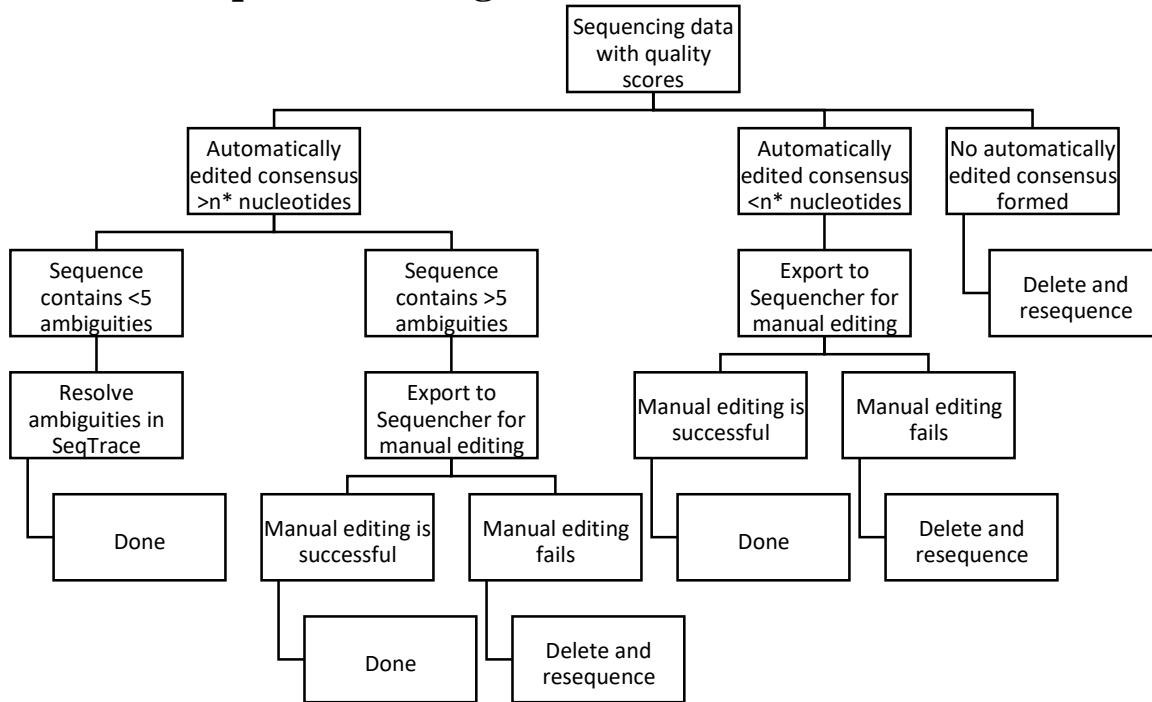
Approved by:

Automated Editing of DNA Sequences using SeqTrace

FILE NAME: Automated Sequence Editing SOP_10112019.doc



Automated Sequence Editing Decision Tree



***Recommended consensus length cutoff values:**

Gene	Length (n)
ITS	500
NLE	500
16S	600
<i>rpoB</i>	632
<i>sigB</i>	630



TABLE OF CONTENTS

1. AUTOMATED SEQUENCE EDITING DECISION TREE
2. INTRODUCTION
 - Purpose
 - Scope
 - Definitions
 - Safety
3. MATERIALS
4. PROCEDURE
5. TROUBLESHOOTING
6. REFERENCES



SECTION 1 - INTRODUCTION

1.1 Purpose

To provide a functional automated platform for the alignment, editing, and resolution of sequence ambiguity necessary to efficiently process and BLAST large numbers of short DNA sequences, such as those produced by high-throughput Sanger sequencing.

1.2 Scope

This SOP may be used by all lab members at the discretion of the PI.

1.3 Definitions

- KB Basecaller:** A computer algorithm used to interpret the electropherograms produced by the Sanger sequencing instrument into a DNA sequence. It is more advanced than the standard ABI basecaller, and is capable of accurately extracting a greater number of bases from both the 3' and 5' ends of the sequence. It assigns a quality score to every base it calls, and is also capable of calling mixed base positions through the use of standard IUB ambiguity codes. More information can be found here:
http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_041491.pdf
- Quality (Confidence) Score:** The quality score of an individual base call provides an indication of the confidence that the call is correct. The quality score Q is defined as a property which is logarithmically related to the base-calling error probabilities P : $Q = -\log_{10} P$, or $P = 10^{-\frac{Q}{10}}$. The table below demonstrates how quality scores are interpreted:

Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%



3. **SeqTrace:** A free, open-source program that can be used for the automated alignment, editing, and ambiguity resolution of Sanger-sequenced DNA. It performs these tasks based on the quality scores of the individual base calls in the sequences.

1.4 Safety

There are no health or safety risks associated with this SOP.



SECTION 2 - MATERIALS

Materials

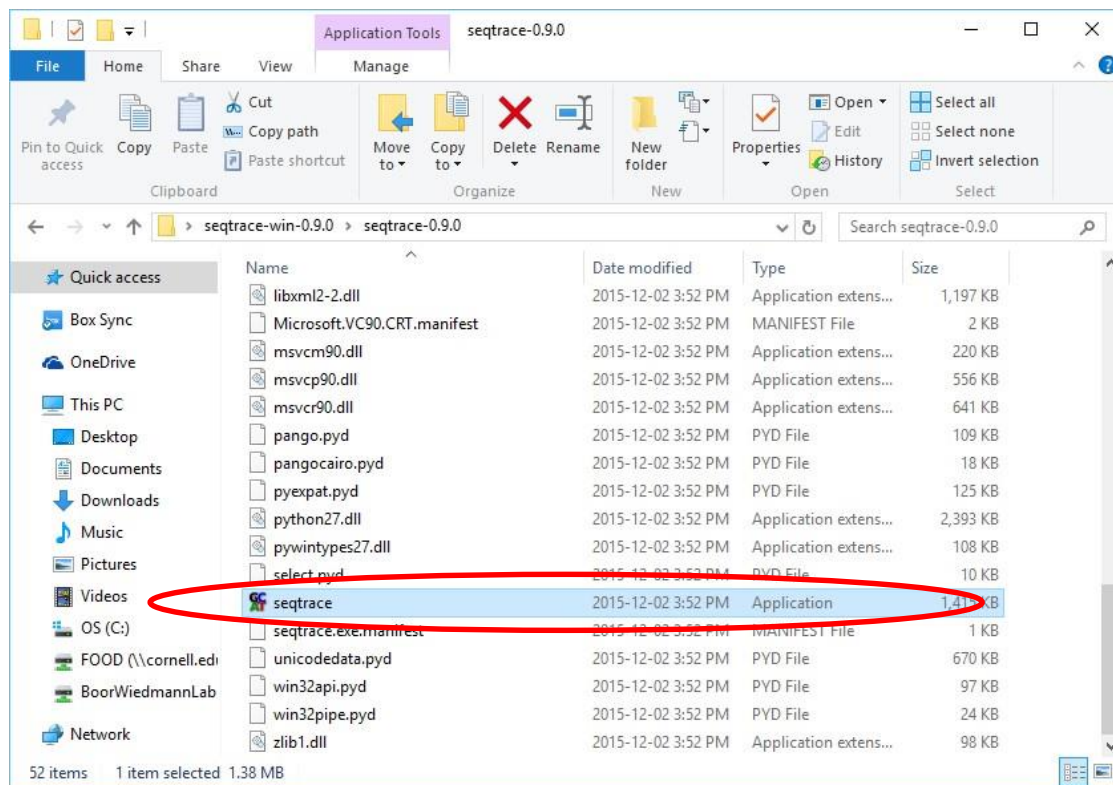
- A PC running Windows or Linux with network drive access and internet access
- PCR product or Ready-to-Load sequencing reactions ready to submit to the BRC, or DNA sequences already received from the BRC



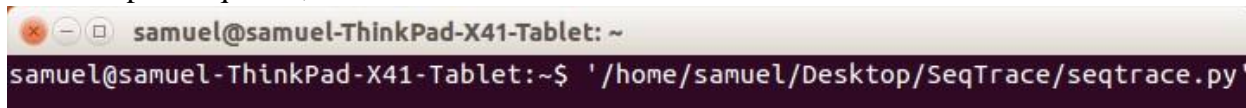
SECTION 3 - PROCEDURES

1. Software installation and running the software

- 1.1. Go to the following website: <https://code.google.com/p/seqtrace/downloads/list>
- 1.2. Download the version of SeqTrace 0.9.0 appropriate for your operating system.
- 1.3. Extract the compressed file to an easily-accessible location on your computer. No installation is required.
- 1.4. **On Windows:** Open Seqtrace by navigating to the folder you extracted the downloaded file to and double-clicking on the SeqTrace application file, as shown in the image below:



- 1.5. **On Linux:** Open a terminal window by pressing **CTRL+ALT+T**, and provide the path to the file *seqtrace.py*. This can be done easily by navigating to the file in the graphical interface and dragging the file into the terminal window. Press enter to execute the file and open SeqTrace, as shown below:



2. Obtaining Correctly Formatted Input Sequences

- 2.1. When filling out the online order form to submit samples to the BRC for Sanger sequencing, include the following comment in the “General Comments” section:



“Please use KB Basecaller software for trace processing, with mixed base identification set at a 70% detection level.” It is also helpful to the BRC staff if you note the presence of this comment to them when dropping off your samples.

- 2.2. If you have recently received sequences from the BRC that were not processed using KB Basecaller, as evidenced by their lack of quality scores, you may contact the BRC and ask them to reprocess your sequence traces post-hoc using KB Basecaller. The BRC staff may be contacted either by using the contact form on their website (<http://www.biotech.cornell.edu/contact-us>, selecting the Genomics Facility as the message recipient), or directly at genomics@cornell.edu.

- 2.2.1. The following message may be used: “Please reprocess the traces from order [order number] using the KB Basecaller with mixed base identification set at a 70% detection level. Thank you!”

3. Using SeqTrace for Automated Editing

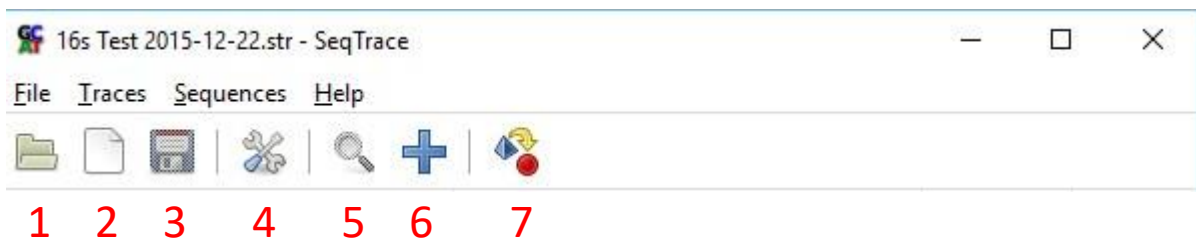
- 3.1. Full software documentation for SeqTrace is available at <https://code.google.com/p/seqtrace/wiki/DocumentationIntro>

- 3.2. Open the SeqTrace program as described in §1.4 or §1.5.


3.3. SeqTrace Basic Features

- 3.3.1. The buttons at the top of the SeqTrace window have the following functions, from left to right:


- 3.3.1.1. Open a project file
 - 3.3.1.2. Create a new project file
 - 3.3.1.3. Save the current project
 - 3.3.1.4. View and change the settings for the current project
 - 3.3.1.5. View the selected trace file(s)
 - 3.3.1.6. Add trace files to the project
 - 3.3.1.7. Export all in-use sequences

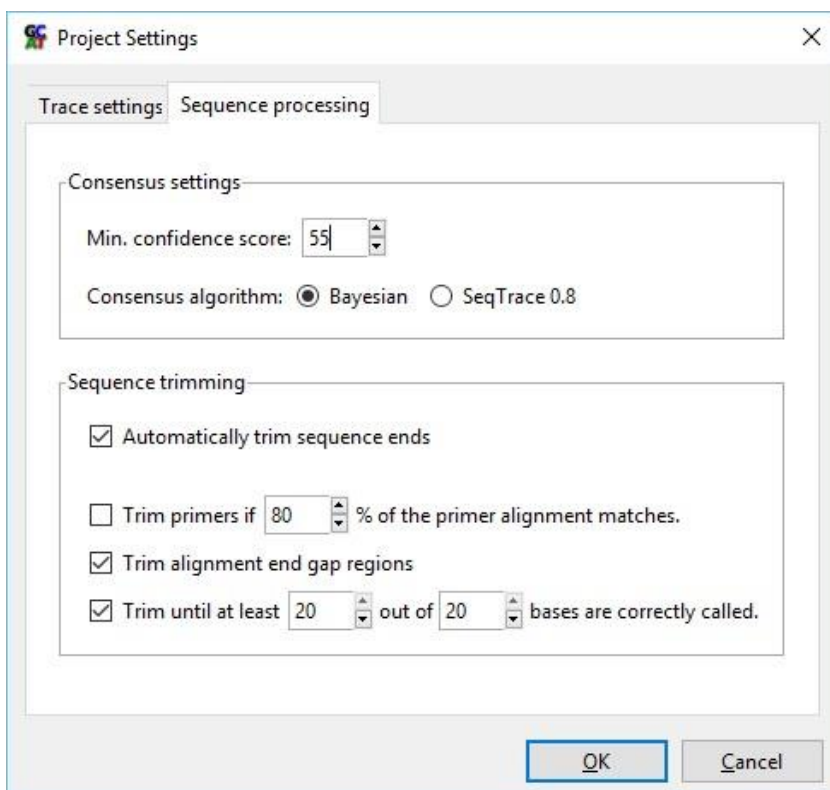
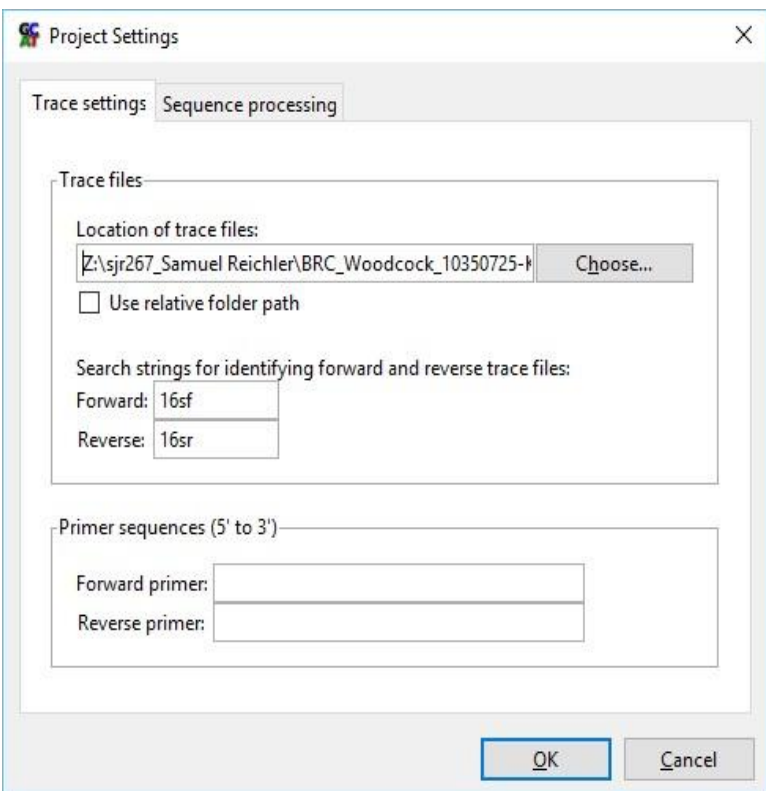


3.4. Creating a Project and Setting Parameters

- 3.4.1. Create a new project file by clicking the  icon.
- 3.4.2. In the Trace settings tab of the Project Settings window, navigate to and select the folder containing the sequences you wish to process.
- 3.4.3. Assign search strings for identifying the forward and reverse trace files.




- 3.4.3.1. For instance, if the names of your forward and reverse trace files are r10222216sf.ab1 and r10222216sr.ab1, respectively, then your forward search string would be “16sf” and your reverse search string would be “16sr.”
- 3.4.4. It is not necessary to enter the forward and reverse primer sequences if you do not want to. If you enter these sequences, they will be mapped onto your consensus sequence but it will not affect the resulting consensus sequence.
- 3.4.5. In the Sequence processing tab of the Project Settings window, set the minimum confidence score to 55.
- 3.4.6. Ensure that the Bayesian consensus algorithm is selected, that “Automatically Trim Sequence Ends” is selected, and that “Trim alignment end gap regions” is selected.
- 3.4.7. Ensure that “Trim primers if __% of the primer alignment matches” is NOT selected.
- 3.4.8. Ensure that “Trim until at least __ out of __ bases are correctly called” is selected, and change both values to the number 20.
- 3.4.8.1. Note: You must enter the second number before you will be allowed to enter the first number.
- 3.4.9. Click “OK” to finalize the settings.
- 3.4.10. Click the  icon to save your project settings to a file.






3.5. Importing and Processing Sequences

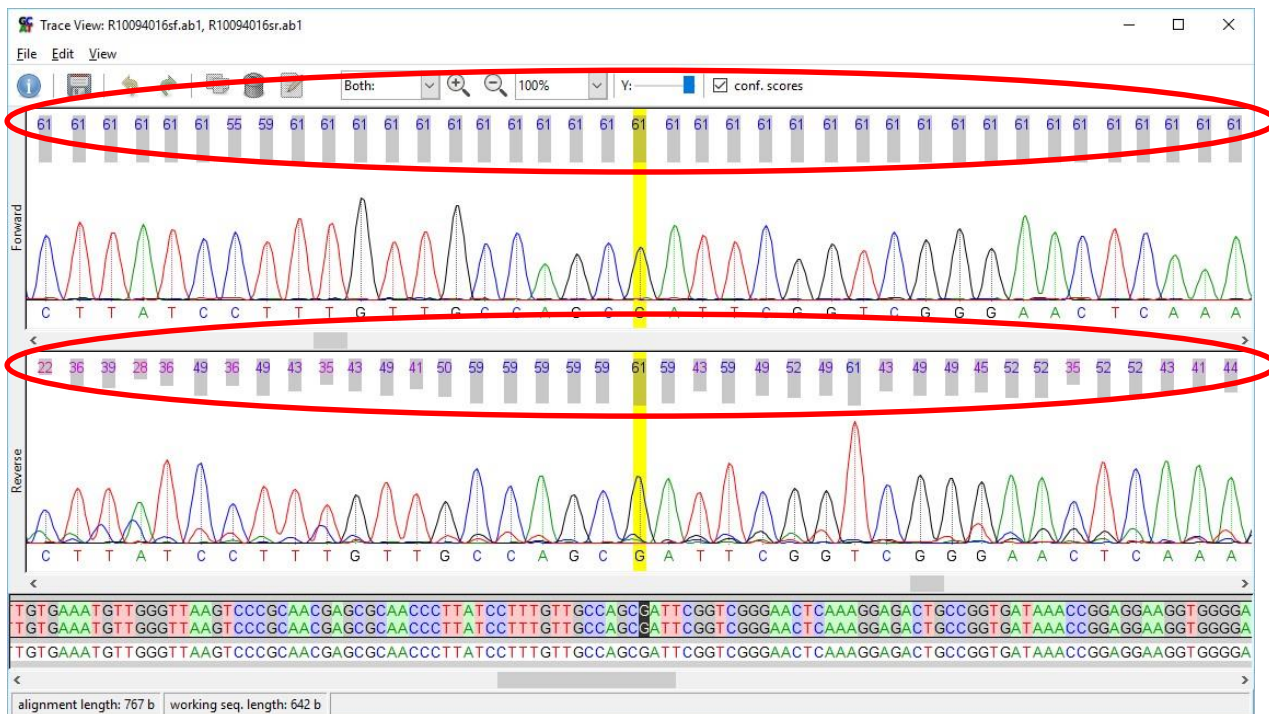
3.5.1. Click the  icon at the top of the window to add sequences to the project. Sequences should be in the *.ab1 file format or the *.ztr file format. Multiple sequences can be selected by clicking on the first sequence desired, holding down the Shift key, then clicking on the last sequence desired. Click the Add button to add the sequences to the project.

3.5.2. At this point, it is a good idea to check a sequence trace file to verify that the KB Basecaller has been used and quality scores have been assigned to each base call.

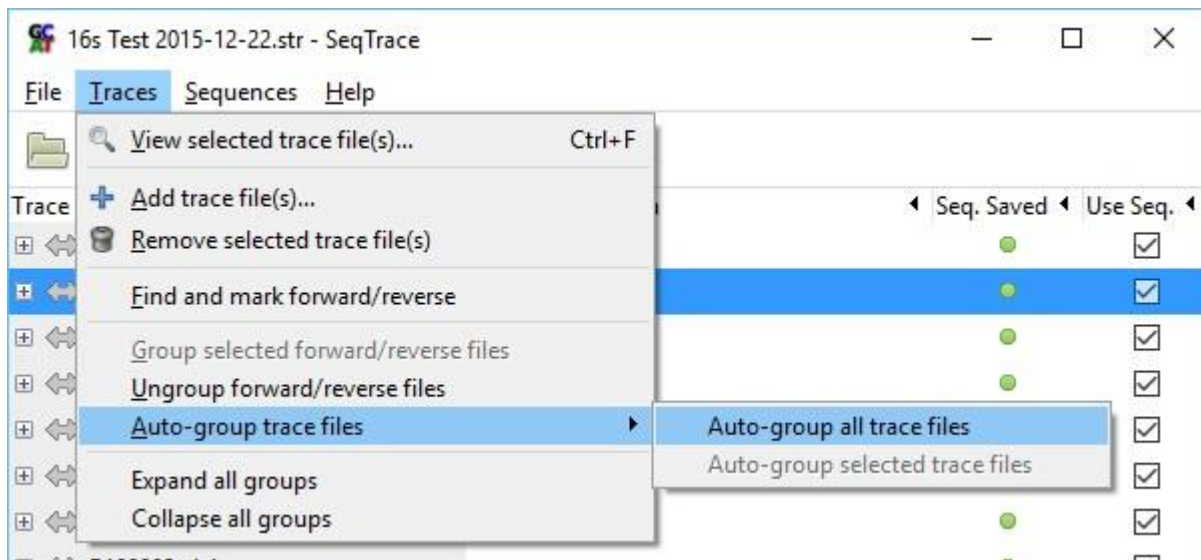
3.5.2.1. Click on a sequence trace file in the window and then click on the  icon. This will open a separate Trace View window.

3.5.2.2. The quality scores, if present, should be clearly visible above the electropherogram. If all quality scores are zero, this means that the sequences were not processed with KB Basecaller and must be reprocessed as described above.

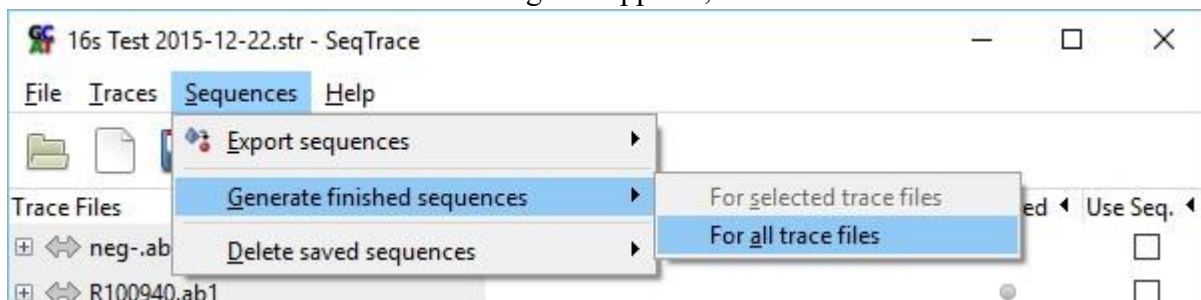
3.5.2.3. Close the Trace View window.




3.5.3. From the “Traces” menu at the top of the window, select “Auto-group trace files, and then “Auto-group all trace files.” A dialog box will appear that contains the first forward-reverse grouping done by SeqTrace. If the forward and reverse pairing is correct, you can click “Yes to All” for all the pairings to be done automatically. If the pairing is incorrect, confirm that the Project Settings match those listed above.




3.5.4. Generate the consensus sequences for all paired traces by selecting “Sequences” from the top of the screen, then “Generate finished sequences,” then “For all trace files.” If a conformational dialog box appears, select “Yes.”



3.5.5. At this point, you may view the consensus sequences by selecting a trace file in the main window and then clicking the  icon. The alignment length and the consensus sequence length are both displayed at the bottom of the Trace View window. The forward and reverse electropherograms are displayed at the top of the window, and below them is the consensus sequence. Clicking on a pair of bases in the consensus sequence causes those bases to be highlighted in the electropherograms.



3.6. Exporting and Quality-Checking Consensus Sequences


3.6.1. To export all of your consensus sequences as a multi-FASTA file, click the  icon at the top of the main SeqTrace window. Save this file and open it in a text editor such as Notepad

3.6.2. Use the Search function (Ctrl+F in Notepad on Windows) to search your multifasta for any ambiguities (“N” nucleotides) in your sequences. These indicate that SeqTrace was unable to generate a consensus for that location in the sequence.

3.7. Editing Consensus Sequences in SeqTrace

3.7.1. Return to SeqTrace.

3.7.2. You may choose to delete the consensus sequences below a certain length from the project file by clicking the small plus sign to the left of the consensus sequence in the main window, selecting the forward and reverse sequences, and then selecting “Remove selected trace file(s)” from the Traces menu.

3.7.3. Consensus sequences that contain ambiguities should be manually inspected to see if it is possible to correct them. Select the consensus sequence you wish to view and then click the  icon in the menu bar to open the Trace View window.

3.7.4. Scroll horizontally through the consensus sequence, looking for the ambiguity you found. When one or more is located, you may proceed in one of three ways:

3.7.4.1. If, upon visual inspection of the electropherograms, there is no doubt that there is consensus between the forward and reverse base calls, as highlighted below, you may edit the sequence to contain this base call.

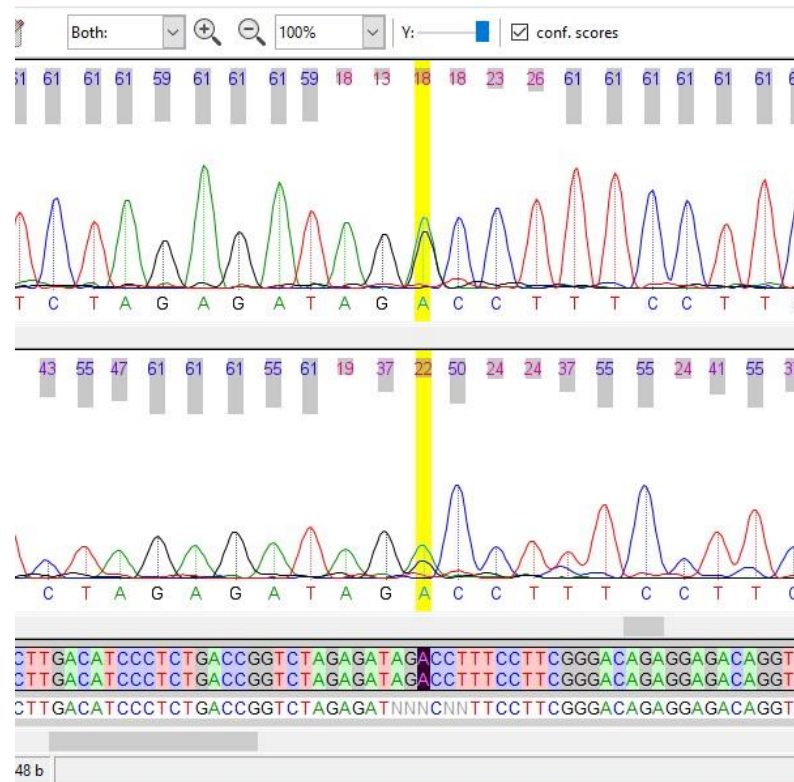





3.7.4.2. If, upon visual inspection of the electropherograms, there appears to be a disagreement between the two or if they contain double peaks, as is highlighted in the example below, the base call may be edited to contain an IUB ambiguity code:

IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T



V	A or C or G
N	A,C,T,G



- 3.7.4.3. If, upon visual inspection of the electropherograms, you can make no distinction between A, T, C, and G for a base call, you may leave it in the consensus as an N.
- 3.7.5. To edit the consensus sequence, click and drag to highlight the desired bases in the consensus sequence. Then click the  icon at the top of the window. Type the desired bases into the dialog box and then click OK.
- 3.7.6. If you wish to delete bases from the consensus sequence, highlight them as described above, then click the  icon at the top of the window.
- 3.7.7. Once you have finished editing the sequence, click the  icon at the top of the window before closing the Trace View window.
- 3.7.8. Repeat the above steps 3.7.3-3.7.7 to edit all consensus sequences containing ambiguity.
- 3.7.9. Once you believe you have addressed all instances of ambiguity, resave your consensus sequences and search for any missed ambiguities as described in 3.6., and if necessary, repeat the above editing and rechecking process.



4. Exporting and BLASTing Consensus Sequences

4.1. Export all edited consensus sequences as a multi-FASTA file as described above in step 3.6.1, saving the file wherever you want to.

4.2. BLASTing Consensus Sequences

4.2.1. Instructions for conducting a BLAST using SequenceServer can be found on the Cornell Food Safety wiki at :

<https://confluence.cornell.edu/pages/viewpage.action?spaceKey=MQIP&postingDate=2015%2F9%2F9&title=SequenceServer>.

If you cannot reach this site, navigate directly to the SequenceServer here:

<http://zeus.foodscience.cornell.edu:4567/>

4.2.2. It is not necessary to follow the instructions for concatenating all sequences files into a multi-FASTA file as described in this tutorial, as SeqTrace does this automatically.



SECTION 4 – TROUBLESHOOTING

- Certain sequences that have had quality scores added post-hoc have been observed to crash SeqTrace. If SeqTrace freezes while processing your consensus sequences, look at the checkmarks in the “Use Seq.” column of the main window to see which sequence was last successfully processed. The next unchecked sequence in the list is the one causing the issue. Restart SeqTrace, delete the forward and reverse traces for this sequence, and rerun.
- Double peaks present an issue when editing using SeqTrace. KB Basecaller calls double peaks with ambiguity codes, and the quality scores for these calls typically falls below the set cutoff in SeqTrace. SeqTrace will therefore call double peaks as N in most cases, even when the ambiguity codes called in the trace files match and are legitimate. This is a particular issue for genes with multiple copies, such as 16S. Sequencher will call ambiguity codes in the consensus sequence at a lower quality score cutoff, so using Sequencher to manually edit sequences known to possess numerous double peaks may be recommended.
- If consensus sequences are too short or contain a large number of ambiguities, it is very likely that the sequences were noisy or of low relative quality. If it is desired to use them with automated editing regardless, the pros and cons of modifying the project parameters described above to obtain better results must be carefully weighed.



SECTION 5 - REFERENCES

<https://code.google.com/p/seqtrace/>

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3413935/>

http://www.chick.manchester.ac.uk/SiteSeer/IUPAC_codes.html