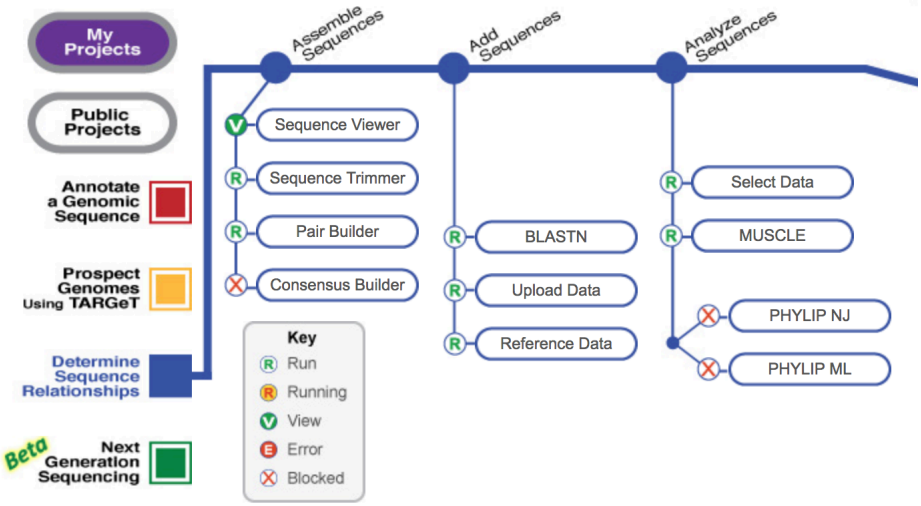
Sequence Analysis Guide

Note that there are lots of ways to analyze sequence data, and many research groups have created software to work with this type of data. Below is one curated route that will get you started to address your research question. If you have a research question that these techniques can't address, ask your TA for help to explore other bioinformatics algorithms.

1. **Reading the data.** Sequence data is available on the course website as two filetypes:
    a. sequence files in .ab1 format with interactive electropherograms (the .ab1 filetype is uploadable into other application for visualization and analysis)
    b. electropherograms in a PDF (created from the .ab1 files using the Mac platform software 4Peaks at https://nucleobytes.com/4peaks/index.html. It is not necessary for you to download this software for this course.)

2. **Assess data quality.** Using the electropherogram PDFs, describe the quality of the sequences: how much noise is present? Are there any portions of your sequence missing or that appear to be mis-called? How high is your confidence in the sequence calls? Are there regions of the sequence of lower quality?

3. **Align forward and reverse sequences for identical samples.** Navigate to DNA Subway (https://dnasubway.cyverse.org/) and create a free account. On their homepage, select Determine Sequence Relationships. Select project type "DNA" and choose files to upload ABI1 trace files. Use Ctrl+ or Alt+ mouse click to select all files you wish to compare and upload (we recommended a maximum of 12 files, or 6 unique samples, for your first use). Name your project and continue. The screen capture below shows your options in DNA Subway, which is a GUI that can push your data out to different sequence analysis sites. We'll use DNA Subway to **Assemble Sequences**, including view, trim, pair, and build consensus FASTA files.



a. For each forward/reverse pair follow the instructions below to view, align, pair, and build a consensus. Once you have a satisfactory consensus sequence, select it with your mouse and save it as a txt file in notepad or text edit.
   i. Click Sequence Viewer and compare the sequences to your electropherogram PDFs. You'll see N's for ambiguous calls, etc. Note if any sequences are missing. No action is needed at this step, so after viewing you may close the window.
   ii. Click Sequence Trimmer. This application will trim off the low quality sequence calls at either end of your sequences, but remove nothing from the middle of the sequence. Once trimmed, you'll see a new version of your sequences with the statement "Your sequences have been trimmer." Visually inspect, then close the window.
   iii. Click Pair Builder. For each reverse sequence, select the hyperlinked "F" to convert it to an "R." Check the boxes next to two sequences you want to pair and response Yes to the "Pair them?" dialog box. Repeat for every pair, then Save your pairs.
   iv. Click Consensus Builder. Wait until the samples has processed (R converts to C), when click again to view. For each sample, view the consensus, decide whether to trim or not, then select and copy the name and consensus sequences into a text editor, formatted as a FASTA file like this:

```
>SEQUENCE_1Name
AGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTC
>SEQUENCE_2Name
AGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGT
```

   v. At this point, add each consensus to the class google sheet using the link shared in class.

4. **Check for similar sequences in the NCBI database.** Using the BLAST online interface (https://blast.ncbi.nlm.nih.gov/Blast.cgi), select Nucleotide BLAST, and paste your FASTA sequence(s) into the box and click BLAST. The algorithm will search a large database of sequence information and return a list of the most similar. This is an excellent way to check for contamination and identify off target PCR amplification.

5. **Sequence Similarity.** What sequences in the database are most similar? Does this align with your expectations? How are you assessing similarity?

6. **Mutagenesis changes to gene of interest.** Compare the sequence of your original and mutated *RFP* genes. On the BLAST homepage, locate the "Global Align" tool.
   a. Set your ancestral gene sequence as the reference, and a derived gene as the Query.

   b. Repeat this for each derived sample from that ancestor.

c. How do the sequences compare? What hypotheses can you form based on the differences between the two? Are they more or less similar than you expected? What kind of mutation occurred, if any?

d. Make a table of DNA changes for your notebook, and report any other important evidence for your research question.

7. **Mutagenesis impacts on the protein sequence.** We will use the program ExPASy to determine the most likely AA sequence given our DNA sequence data.

   a. Navigate to web.expasy.org/translate/, paste a nucleotide sequence, and click translate. The program will provide several options with different codon starting locations. The best option is the one with the largest open reading frame, that is, the largest number of amino acids (AAs) found between a start and stop codon.

   b. Keep these amino acid sequences for your records and navigate back to the BLAST Global Align program. Alternatively, you can use MUSCLE (https://www.ebi.ac.uk/Tools/msa/muscle/). Click the protein tab and set your original RFP sequence as the reference, and your mutant gene as the Query.

   c. Make a table of AA changes. Are some more likely than others to have contributed to functional differences between the proteins? Were there more or fewer AA changes than nucleotide changes between the sequences? Why do you think these numbers are different?

8. **Interpreting your results.** Discuss your results with your partner, peers, and TAs. Do you have the information you want to address your research question or hypothesis? Consider other comparisons you want or need to make and identify the data and applications you'll need. Then make those comparisons.