Bracken-Grissom Lab:
Cleaning Sequence Data with Geneious

1. Retrieve your sequence data.
    a. Genewiz submits the results of your sequencing plates on their website. To retrieve your data, navigate to their website and login using the account information that you used when you submitted the order.

    b. At the bottom of the page, in the "MY ORDERS" tab, your plate results should be displayed. The results will be available as two different tracking numbers; one for the forward strand results, and one for the reverse strand results

    c. Select one of the sequence results by clicking on its tracking number.

    d. Click "View Results" and select every row by checking the boxes in the far left column. At the top of the page, select ".ab1" and click "Download."

    e. The files will be downloaded as a ZIP file. Move the ZIP file to a folder named after the name of the sequencing plate, and unzip it there. A new folder should appear containing 96 raw read files. These reads represent the sequences for one strand of every sample you submitted, either the forward sequences, or the reverse sequences.

    f. Repeat the above steps to download the sequence data for the other strand sequence data.

2. Rename your raw read files, if necessary
    a. If the order was submitted correctly, the raw reads should have the same names as what you submitted. If not, they will only have the plate well numbers for names (i.e. only A1, A2, A3…H12).

    b. It will save an immense amount of time in the long run if you rename the files of at least one folder before you move forward. If you're working with multiple genes, it's advised that you rename the files in both folders. Refer to your Reference Plate Map spreadsheet to do so. The names should be formatted as follows:
    A1_G_species_Gene
    A2_G_species_Gene
    A3_G_species_Gene
    …
    H12_G_species_Gene

3. Upload the sequence data to Geneious
    a. Open Geneious

    b. If not already done, create a new folder where you can place your sequence data.
        i. On the left is a series of folders. Right click on the top folder (Local), and select "New Folder." Name the folder after yourself, or your project. Click "Ok."

    c. Right click on your folder, create a new folder named after the sequencing plate that your sequences come from, and then create a new folder inside that folder for every gene you sequenced on this plate.
        i. This may seem a bit excessive, but it's important to keep your sequences in order, especially if you're working with multiple genes

    d. Open your folder of forward sequences that you downloaded earlier, and select every file for one gene contained therein. Either copy the files and paste them into Geneious into the gene folder they correspond to, or simply drag and drop them. Do the same for the reverse sequences. Finally do the same for every gene on the plate
        i. Each gene folder in Geneious should contain a complementary forward and reverse sequence for each specimen.

4. Assemble your sequence reads in Geneious
    a. Highlight every sequence in the gene folder

    b. Right click, or click the Align/Assemble button at the top of the window. Select De Novo Assembly.
        i. Assemble by 1st part of name, separated by _ (Underscore).
        ii. This will pair your sequence reads based on their names, looking exclusively at the part of the name before the first underscore. This should be the Sequence Plate well # if you named the reads as suggested above. If not, make sure you select an option that reliably pairs your reads together, or rename your reads.
            1. Assembling the sequence reads pairs the forward and reverse strand sequences with their complementary strand, and aligns them so their complementary base pairs are paired together.

c. The assemblies will have their name reduced to the part of the sequence read name that you assembled them by. It will be helpful in the long run to rename the assemblies to include their Voucher#, Genus, and Species.
    i. If the raw read files were named correctly before you imported them to Geneious, the description column for each specimen will show the file names that they came from, which will include all of this information, making it very easy to rename the assemblies.

d. It's possible not all of your reads will assemble. This is likely because at least one of the reads in the assembly was too low in quality. If that's the case, you can check the "Unused Reads" report in your assemblies' folder. If one strand looks high enough in quality, you can use the single strand read as the sequence, but this is not ideal, and can be risky. Unless, you're desperate, you should most likely try to re-amplify the specimen and re-sequence on another plate.

5. Trim the messy ends, and the primers from the ends of your assembly. Trimmed regions of a strand do not contribute to identifying the consensus sequence.
    a. Select an assembly and right-click. Select "Trim Ends."

    b. Next to the primer selection box, click "Choose." Search for the primer sequences you use to amplify the reads in this assembly. Select all of them that were used for this particular gene with cmd+click.
        i. If your primer sequences aren't available, you'll have to import them to Genious.
        ii. Create a folder inside your named folder, and name it "Primers," if you don't already have one
        iii. Find the sequence of your primers, copy the sequence (just the sequence, not the file), and paste it into the Primers folder you just created as an unformatted sequence.
        iv. Select the sequence file you just imported to Geneious, click the "Primers" button at the top of the window, and select "Convert to Oligo." Select the primer option, and press "OK."
        v. Rename the primer sequence
        vi. Repeat for every primer you need to import.
        vii. Repeat Step 5b, and your primers should now be available to select

    c. Click "Select"

    d. Click "OK"

e. The assembly should now be annotated to show where primers were located in the reads, and what regions were trimmed, either because they were after the primer region, or they were too low quality.

f. Double-click the assembly file to open it in a separate window.
    i. You should see two strands that are near identical (hopefully), with a series of different colored peaks. These peaks represent where a base was sequenced, and the color of the peak represents which base it was.
    ii. A high quality sequence will have very defined peaks with no background "noise."
    iii. Above the read assembly strands is a third line of bases. This is the consensus sequence, which represents the combination of the two strands into a single strand. The consensus sequence is usually what is used for any bioinformatics analysis. Whatever changes you make to the assembly strands will change the consensus sequence

g. Manually check the assembly to ensure there are no disagreements, and that the Trim Ends function worked correctly.
    i. If there is a disagreement between the two strands at a particular base:
        1. First make sure that Geneious called the bases correctly. If one of the bases is called incorrectly, manually change it to what the colored peak shows it as.
        2. If one strand is too messy to call at a particular base, then use the other clean strand. Make sure the consensus sequence agrees with the cleaner strand.
        3. If both bases are called correctly, but they still disagree, then you must change the consensus sequence to show there is an ambiguity. Use the IUPAC Nucleotide Nomenclature Table to ensure the consensus sequence shows the correct ambiguity.
        4. If both strands are too messy at a particular base, that base in the consensus sequence must be labeled as an "N."
        5. Finally, make sure there's not been a frameshift in one of the strands. This will be obvious if there's a sudden increase in disagreement between the two otherwise complementary strands. To resolve the frameshift, find the region where the shift occurs. Either insert a gap, or delete a base from one of the strands.
    ii. It may happen that Geneious trims off portions of a strand that seem perfectly acceptable, quality-wise. To resolve this issue:
        1. select the annotation (The grey bar running the length of the trimmed region), and delete it.

2. Locate the actual region of the assembly that needs to be trimmed (either the primer region, or where the strand starts to become too messy)
3. Highlight from where it starts, all the way to the end of the strand (just on one strand, not both, and not the consensus sequence)
4. Click "Add Annotation" at the top of the window.
5. Make sure the "Type" is Trimmed. You don't need to change anything else. Click "Ok."
6. The correct region of the strand should now be trimmed.

6. Repeat Step 5 until all assemblies are cleaned.

7. Verify there was no contamination by BLASTing your assemblies.
    a. Select all of the assemblies

    b. Right click and select BLAST

    c. Click "Search"

    d. Each Assembly will have a separate BLAST folder with the top 5 BLAST hits from the GenBank database. Go through each folder and make sure the results make sense. Each hit should be a species close to your specimen (Doesn't have to be exact, but reasonably close), and should be of the same gene. If you you're sequencing a COI fish gene, and the top hit is a 16S human gene, there was definitely a mistake somewhere during the process.

8. Repeat Steps 3 – 7 for every set of reads for each gene on your sequence plate

Update any spreadsheets you have to show which sequences were successful, and which sequences failed.

END