**Computer Lab 2 – Sequence Editing**

**Conservation Genetics (BIOL 4174 / 5174)**

**Part I – Editing Mitochondrial DNA Sequences**

Once Sanger sequence data have been generated, a great deal of work remains before they are ready for analysis. These early steps in preparing your sequences are critical because the raw output from the sequencer contains errors that must be removed manually to minimize their influence on downstream analyses. In this lab you will focus on sequence editing.

Many software solutions exist for editing sequences, although the best are available only at a premium. Sequencher and Geneious provide excellent, but expensive, options. For the purposes of this lab we will rely on the full-featured 2-week trial version of Geneious. This program is available for Windows, Mac, and Linux.

**Before starting today’s lab:**

1. Download the updated materials from the lab github page (<https://github.com/tkchafin/ConsGen2018)>. To do this, open the Terminal and **cd** to your local repository (which should be located in (/Users/ConsGen/Desktop/<username>/ConsGen2018). Type **git pull** to fetch new changes.
2. Launch Geneious and initialize your free trial using the key that you received from Biomatters. If you have not yet requested a free trial key, this can be accomplished at: <http://geneious.com/request-trial>. You may be asked for administrative access- ask your TA if this happens.

Sequence editing instructions:

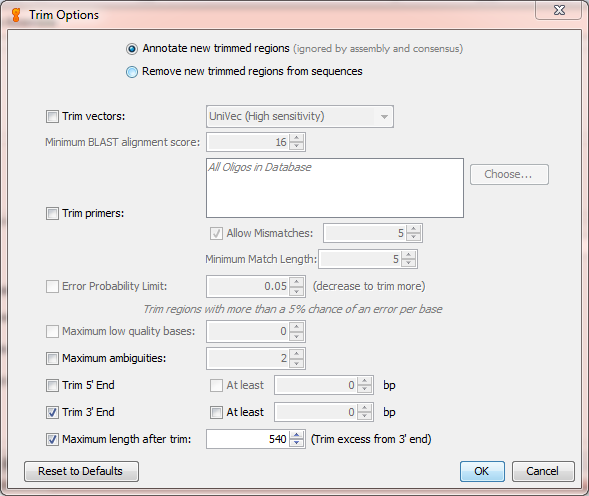
*You can use these instructions as a baseline for editing the first two sets of mitochondrial sequences. Parts II and III will deviate from the workflow outlined below, however the basic instructions for importing and exporting sequences will remain the same.*

1. **Importing sequences**
   1. Create a new folder
      1. Highlight the “Local” folder in the left-hand pane of Geneious by clicking it once
      2. Create a new folder (File 🡪 New Folder…)
      3. Enter “sequences” as the folder name and select “OK”
   2. Click your newly created folder to make it active
   3. File 🡪 Import 🡪 From Multiple Files…
   4. Navigate to the location where your sequences for Part I are located. Select all four files in the “part\_1” folder and click “Import.”
   5. The files should now appear in your sequences folder.
2. **Assembling *contigs* (sets of overlapping DNA sequences)**
   1. Select all four of the sequences you imported into Geneious.
   2. Click the “Align/Assemble” button at the top of the Geneious window and choose “De Novo Assemble…”
   3. Leave all options set at default, except select the “Do not trim” radio button.
   4. Select “OK”
   5. The contig will be in your “Sequences” folder named as “Reads Assembly”
3. **Cleaning sequences**
   1. Double-click the “4 documents Assembly” file to open it.
   2. To better view the chromatograms in the contig you may need to zoom in (top right corner of the assembly window).
   3. Click the “Allow Editing” button at the top of the window.
      1. If prompted to “Edit as linear” click the “Edit as linear” button.
   4. Make sure the forward and reverse sequences are identified properly
      1. On the left hand side of the screen, the sequence names should be identified as forward (FWD) and reverse (REV)
      2. The forward sequence is identified by a lower-case “a” in its name, (e.g. 62S8W01a) and the reverse is identified by a lower-case “b” in the same position).
      3. If the two are not identified correctly by Geneious, click the R.C. button at the top of the screen. Make sure “Reverse complement entire alignment” is selected in the resulting window, and click “ok.”
   5. Above the four sequences you will see a consensus sequence – any edits made to this will affect all four sequences.
   6. Find the start codon (ATG). The beginning of our target sequence in Part I will begin with ATGCCTCAA. Delete everything to the left of this point. The best way to accomplish this is by editing the consensus sequence.
   7. Now find the stop codon (TAA). The end of our target sequence in Part I should be AACGTCTAA. Delete everything to the right of this point. Again, do this by editing the consensus sequence.
   8. At this point if you have followed all directions correctly you should have sequences that are a little longer than 842bp in length.
   9. Clean the sequences by eliminating any gaps erroneously inserted by the program. Also, check any differences that exist across the sequences to make sure they have been correctly identified. The following are important details to keep in mind while editing:
      1. The sequences in parts I and II of this exercise code for proteins. Insertions and deletions that will induce a frameshift mutation (i.e., 1 or 2 bases) will be **extremely** rare.
      2. The sequences you are working with have been sequenced bi-directionally. For example, the files 62S8W01a and 62S8W01b represent the same sequence from the same individual organism. The difference is one sequence was generated using the “forward” primer, and the other was generated with the “reverse” primer. This is done to ensure full coverage of the DNA sequence you have amplified, and also to ensure accuracy (i.e., you can identify genotyping errors).
      3. At times one sequence will be of much higher quality than the other (e.g., the forward sequence may have less background noise or more clearly defined peaks in the chromatogram when compared to the reverse). If two sequences for the same individual give you conflicting results, rely on the sequence that appears to be of higher quality.
   10. The sequences should be 842 bp in length when you are done editing. If you have something longer or shorter than this, try to locate your error.
   11. **Important**: **Save your edits.**
4. **The amino acid translation can aid you in editing your sequences**
   1. To view the amino acid sequence, click the “Translation” checkbox in the right-hand pane. Then click “Options.”
   2. For “Genetic Code” select “Vertebrate Mitochondrial.”
   3. The sequence you are editing contains two genes that exist in two different reading frames. The first gene (“Gene A”) should be in Frame 1. To view the translation of the second gene (“Gene B”) you will have to change the reading frame. The reading frame is set in the “translation options” box.
   4. Rely upon the start and stop codons to determine where “Gene A” begins and ends; then change the reading frame to identify the location of “Gene B.”
      1. Hints:
         1. The two genes overlap one another.
         2. The stop codon for “Gene A” should be represented by nucleotides 166-168.
5. **Exporting Sequences**
   1. Make sure you have saved your edits (File 🡪 Save)
   2. In the main Geneious program window, highlight the “4 documents Assembly” and go to File 🡪 Export 🡪 Selected Documents…
   3. Navigate to the “exported\_sequences” folder in your Lab2 folder, and save the sequences as “part\_1.fasta”. Make sure you select the file type FASTA from the drop-down menu.

**Part II – Editing the Mitochondrial Control Region**

For this part of the exercise you will edit sequences from the mitochondrial control region of a species of bird. This region of the mitochondrial genome does not code for an amino acid sequence, so you cannot use the translation to help you out. Additionally, it is very possible for insertions or deletions (“indels”) to be present.

It will also be necessary to trim these sequences in order to place them into a contig. This time, after selecting “De Novo Assemble…” check the “Trim Ends” radio button and use the settings as shown below:



Also, manually delete the beginning portion of the sequences which is unreadable (this should be approximately the first 25-30 bp of the resulting contig).

**Hint**: One sequence from this part of the exercise contains a single base deletion. It should be around the 240 bp mark once you have finished trimming the sequences. Be sure to save your edits and export the sequences as the file “part\_2.fasta” when you are finished.

**Part III – Nuclear Introns**

Edit the set of intron sequences found in your part\_iii folder. Because this is non-protein coding nuclear DNA, you will not be identifying start and stop codons. Introns have a much greater chance for presence of indels, although none are present in the intron sequences you have been given today. However, there will be heterozygous bases in these sequences. When you find a heterozygous base, change it to the proper base ambiguity code shown below.

Base Ambiguity Codes:

G or T = K

A or C = M

A or G = R

C or T = Y

C or G = S

A or T = W

Only edit the portions of the sequences covered by both the forward and reverse strands. Delete the remaining portions of the sequences by editing the consensus sequence. Once you are finished editing, your sequences should be a little less than 340 bp in length.



**Hint**: In the above chromatogram, notice the presence of 2 true peaks (“A” and “G”) at position 84, indicating that this is a heterozygous base. In this case, the “N” bases should be replaced with the ambiguity code “R”. Be careful when calling heterozygotes though- it won’t always be so clear cut!

Be sure to save your edits and export the sequences as the file “part\_3.fasta” when you are finished.

**Homework Assignment**

Any portion of this lab not completed during lab hours will be considered homework. **Please answer the questions from the homework document in your lab\_2 folder and submit it to Blackboard**.

Also, make sure you export your sequences as FASTA files when you finish each section. You do not have to submit these to me, but you will need them for next week’s lab. We will be using some of these sequences next week as we begin to explore Genbank and BLAST.