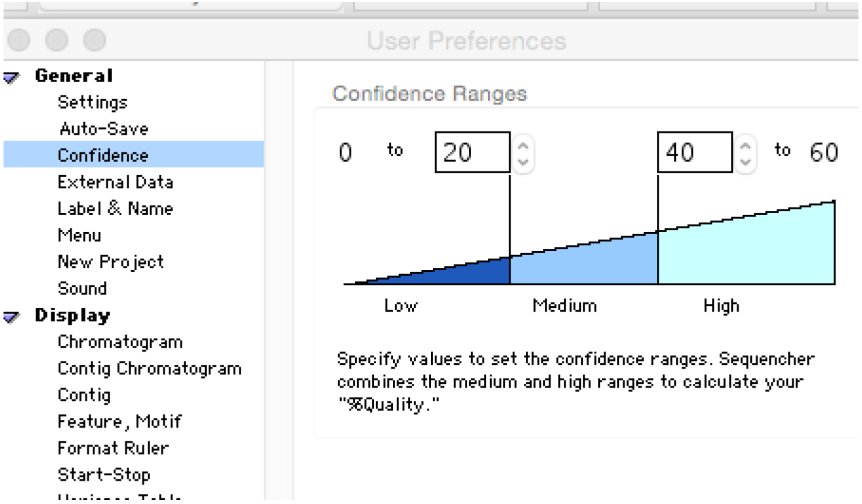
February 2016

Processing Raw Data with Sequencher

1. First set “**Confidence**” in the “**User Preferences**” box (**WINDOW** menu) so that “Medium” confidence ranges between 20 and 40 (these should be the numbers in the boxes). You should also choose “**Display Base Confidences**” (**VIEW** menu).

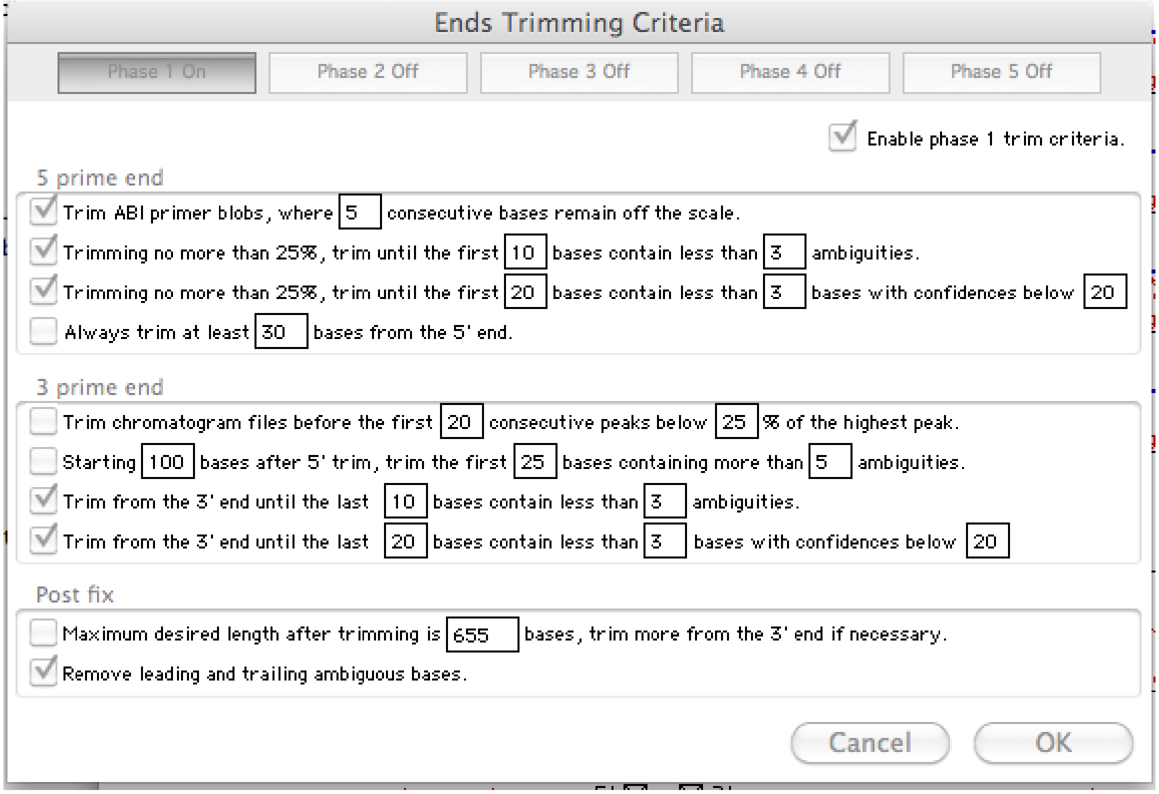


Confidence Ranges

0 to 20 … 40 to 60

Specify values to set the confidence ranges. Sequencher combines the medium and high ranges to calculate your "%Quality."

2. Trim sequences using the “barcode parameters.” Set parameters as shown in the picture below. Use “**Change Trim Criteria**” from the “**Trim Ends**” window. (**SEQUENCE** menu)



☑ Enable phase 1 trim criteria.

5 prime end

☑ Trim ABI primer blobs, where 5 consecutive bases remain off the scale.

☑ Trimming no more than 25%, trim until the first 10 bases contain less than 3 ambiguities.

☑ Trimming no more than 25%, trim until the first 20 bases contain less than 3 bases with confidences below 20

□ Always trim at least 30 bases from the 5' end.

3 prime end

□ Trim chromatogram files before the first 20 consecutive peaks below 25 % of the highest peak.

□ Starting 100 bases after 5' trim, trim the first 25 bases containing more than 5 ambiguities.

☑ Trim from the 3' end until the last 10 bases contain less than 3 ambiguities.

☑ Trim from the 3' end until the last 20 bases contain less than 3 bases with confidences below 20

Post fix

□ Maximum desired length after trimming is 655 bases, trim more from the 3' end if necessary.

☑ Remove leading and trailing ambiguous bases.

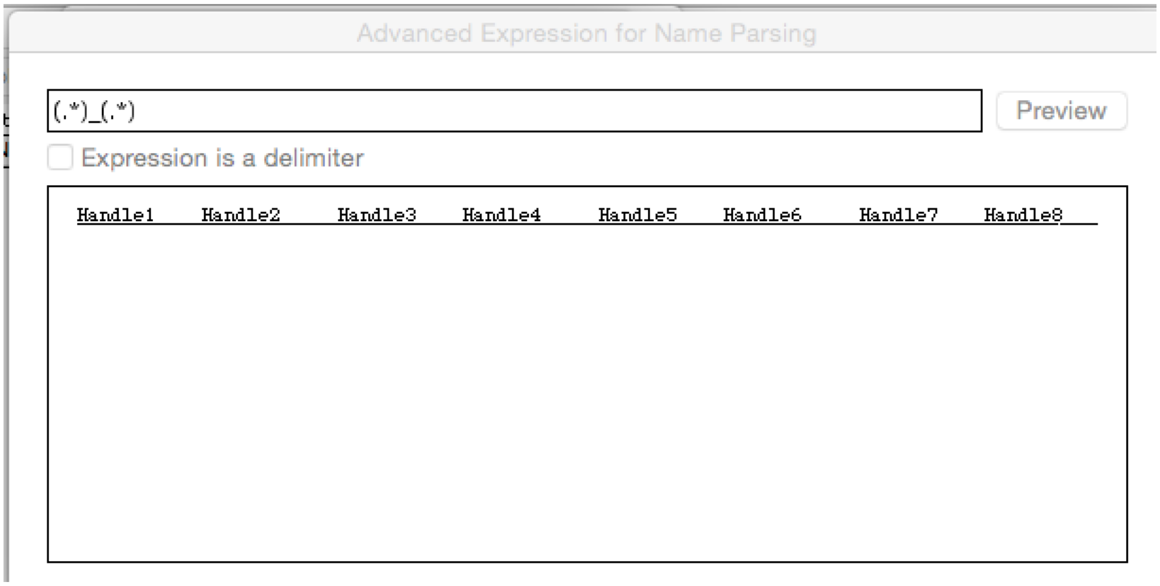
3. Sort seqs by size, **delete all < 450 bp** long (for CO1, this will be different for other genes, of course)

4. Sort seqs by quality, **delete all < 85%.** Be very critical of remaining sequences with <95% quality. They are generally still pretty ugly (often with high background). They may well be discarded later on

5. Set the contig criterion to "**Contig by Confidence**" (**CONTIG** menu). If you forget to do this upfront "Consensus" is the default, but you can change this later and all the contigs will be automatically updated. If you find a ton of ambiguities in the contig, this is a clue you may not have chosen "Contig by Confidence."

6. Choose "**Assemble by Name**" from the pull-down in the top part of the project window.

7. To get consistent names for assembly, you must change the "**Name Delimiter**" in the "**Name Settings**" window under "**ASSEMBLY PARAMETERS**" (on the upper region of the main project window). In the "**Name Settings**" window, choose "**Advanced Expression**" in the pull-down and click "**Define**." We generally use a regular expression e.g. (.*)_(.*) so that the final contig name will contain field_id (or voucher number), taxon name, AND ESPECIALLY the barcode. **The BARCODE MUST BE IN THE FILENAME**. Also UNCHECK the "Expression is a delimiter" box.



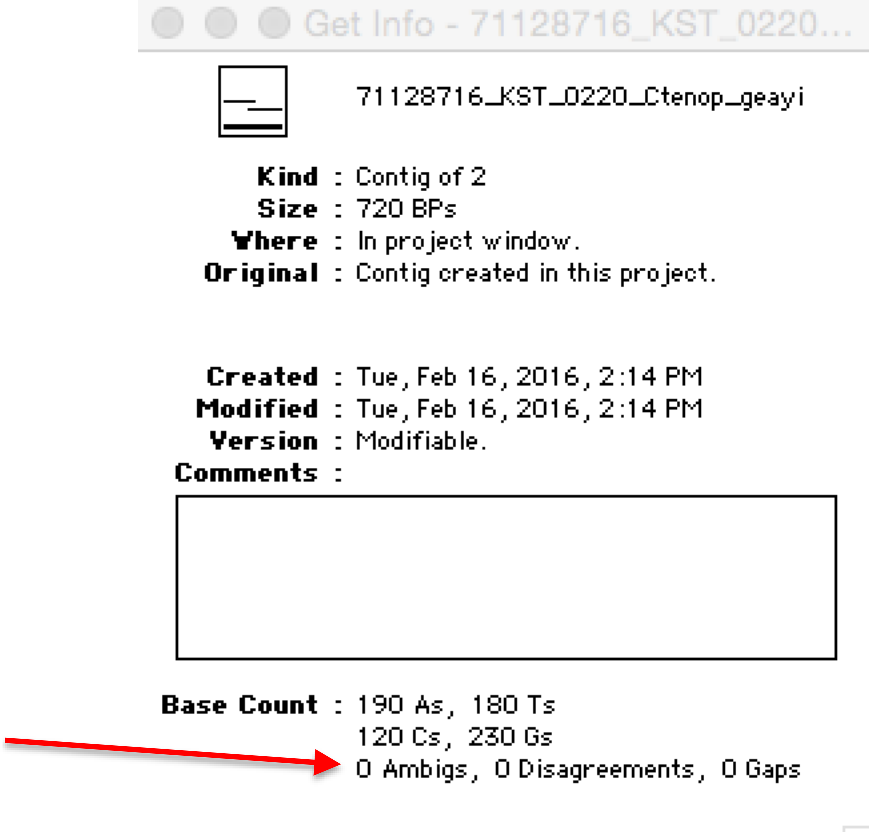
(.*)_(.*)

□ Expression is a delimiter

8. Assemble all sequences by name using **97-98%** sequence similarity.

9. Open windows for all contigs (when all are highlighted, hit enter). Use "Get Info" (command-I) to check # ambiguities, # disagreements, # gaps

a. If *0 ambigs, <4 disagreements, 0 gaps* → *contig passes*. (This info is at the bottom of the "Get Info" box (see below). You will want to quickly glance at the quality of the bases on the 5' and 3' ends, where there is no overlap with the other strands. If all are med-light blue, they should be fine.



71128716_KST_0220_Ctenop_geayi

Kind : Contig of 2
Size : 720 BPs
Where : In project window.
Original : Contig created in this project.

Created : Tue, Feb 16, 2016, 2:14 PM
Modified : Tue, Feb 16, 2016, 2:14 PM
Version : Modifiable.
Comments :

Base Count : 190 As, 180 Ts
120 Cs, 230 Gs
0 Ambigs, 0 Disagreements, 0 Gaps

b. Edit remaining contigs manually → only make changes that everyone would agree with. Consider the most unpleasant person you know – only make changes he would not disagree with.

c. Generally there will be *no N*'s or other ambiguous bases in the final consensus sequence, especially when dealing with mitochondrial data. If N's, etc. are present, something is wrong and the sequence should be checked carefully. If there is not better base call than N, then it can remain. But there should be no more than 3 N's in the final sequence.

d. *Check ambiguities, fix gaps*. If one or the other sequences looks like junk (high background, double peaks, etc.) throw them both out.

e. Check overview window to see that contig has an open reading frame. (For CO1). You may have to change the translation table ("Genetic Code", WINDOW menu).

f. Once all of these have been checked, and if there are sequences remaining to be contig'ed, dissolve contigs and Refrigerate (**EDIT menu**) them.

10. Contig remaining sequences at 96%. A lot of these will be quite bad and have >10 disagreements.

a. Check these out manually – anything that's remotely horrible throw out. Often these have some *double-peaks*. There should be none in mitochondrial data, obviously. Sometimes double peaks may arise from sequencing error and these might need to be changed to ambiguity codes in the consensus sequence. But if there is more than one of these, assume there is a problem and throw out the sequence.

b. You can check any sequences that won't contig at 96% to see if there's a large dye blob or something similar -- in sequence text this will look like a region of very low phred scores (dark blue) surrounding by sequence of very high quality. You can choose to contig these at < 96% and see if the other strand has good calls in this region.

11. To remove primer sequences, I throw out all singletons and everything else that's not in a good contig. Then contig together every sequence that passed ("Standard" not "By Name" for assembly method) into one big contig if you can – usually I use 60% similarity here, and 30 bp overlap. You may have to adjust the overlap to get things to contig. OR if you're working on inverts, they might not contig at all.

12. Import sequencher project with the primers used to sequence or a reference sequence (if available). (Or make the primer sequences yourself ("Create New Sequence" (SEQUENCE menu). Contig these with the big multi-species contig, usually at 60% similarity and about 9 bp overlap. Trim off the primer sequence region (you will have to remove the primer files from the contig first, then trim remaining sequences). This can be a bit tricky with non-coding DNA, as often the 3' end of the sequences aren't aligned correctly. This can be dealt with in one of two ways: manual editing of the alignment before trim, or contig'ing multiple specimens together in less broad taxonomic groups (by genus for example).

13. Dissolve the giant contig and again contig sequences by name. Export (FILE menu) "Consensus" as "Fasta - Concatenated." Use the "Options" button to make certain "Strand to Export" is "Current Orientation."