Bio4342 – Robert Fulton, Lecture Notes (1/30/06):

Introduction to Computational Tools for Finishing: Characteristics of Phred/Phrap/Consed

1. *Introduction:*
   a. Tools for Finishing (Phred/Phrap/Consed):
      - Phred is the base-caller (confidence/quality of base calls)
      - Phrap is the assembler (putting the pieces back together)
      - Consed is the viewing tool
   b. Consed and other programs:
      - While Consed is packaged with Phred, Phrap and Crossmatch, it can be used independently from Phred and Phrap
        ⇒ For example, the GSC uses KB as the base caller
      - Most of our interactions with the sequencing data will be accomplished using Consed.
        ⇒ As we obtain additional data, we may need to use Phred and Phrap.

2. *More on Phred Scores:*
   a. Phred Scores Basics:
      - Phred scores can be found at the phd_dir directory and are called phd files.
      - Phred scores range from 1-99 and use a $\log_{10}$ scale.
        ⇒ Score=10 indicates an error rate of 1 in 10 bases.
        ⇒ Score=20 indicates an error rate of 1 in 100 bases, etc.
        ⇒ Score=99 indicates artificially high quality (i.e. edited bases)
        ⇒ Score=98 indicates artificially low quality
   b. Changing Phred Scores:
      - Changing Phred scores to indicate artificially high/low qualities is particularly useful in cases of misassemblies.
      - For example, if there are regions (i.e. repetitive regions) that Phrap has trouble with, artificially boosting the Phred scores in the unique regions may allow Phrap to correctly reassemble the region.
      - Increasing the Phred score also allows us to break apart misassemblies
      - Similarly, if regions have relatively low quality, artificially decreasing the Phred scores would result in Phrap putting less weight on the discrepancies in the region and may allow Phrap to correctly reassemble the region.

3. *More on Phrap:*
   a. Comparison of Phrap with other Assemblers
      - One major problem with assemblers is their inability to resolve repetitious regions (i.e they tend to stack all the reads together).
      - Phrap is better than other assemblers because it uses the entire read in addition to the Phred (quality) scores.
      - Using the entire read is particularly useful in cases of repetitive regions.
      - Other assemblers typically remove low quality regions prior to assembly. This means they are throwing away substantial amounts of data.
      - Even though the regions near the edge may be low quality, they nonetheless may provide unique data that could help in resolving multiple repetitive regions in the assembly.
      - Naturally, for this to work, Phrap must also take into account the quality of the bases in the region to prevent misassemblies.

- ♦ Phrap evaluation of quality is based on multiple factors. In addition to the quality of the signal for that particular base (i.e signal strength; width, height of signal), Phrap also considers the quality of the region surrounding the base in question.
- ♦ New assemblers (PCAP) uses information on forward and reverse read pairs to further improve the assembly (Phrap does not use this information)

b. Additional options in Phrap:
- ♦ Changing parameters for minmatch and minscore
- ♦ minmatch indicates the number of bases that must be aligned for the reads to be assembled
  ⇒ For example, for repetitive clones, we may consider boosting the minmatch value to 150
- ♦ Minscore indicates the minimum score the alignment must achieve before the reads are assembled together
- ♦ Changing the force level in Phrap will affect how hard Phrap will try to put the reads together.
  ⇒ If the force level is high enough, Phrap will put everything together.

4. *Common Problems in Consed*

a. Navigate Function:
- ♦ We can use the "Navigate" function in Consed to identify potentially problematic regions in our assembly.
- ♦ The option "Dim low quality" can be used to identify most of the problematic regions in the assembly and is the recommended setting.
- ♦ The option "Dim Nothing" may reveal additional information that we may miss when using the other Dim options. There is useful information within low quality regions.
- ♦ Green tags indicate the presence of repeats
- ♦ Unaligned high quality reads, high quality discrepancies

b. Edits, *edit_dir* and *phd_dir*:
- ♦ Note that edits that are made in Consed are stored in the phd file and NOT in the ace file
- ♦ Hence any edits you have made will propagate into any newly created ace files.
  ⇒ This is particularly important if we decide to re-*phredphrap* the assembly
  ⇒ When you reassemble the sequence, the edited sequence will be used instead of your original sequence
  ⇒ However, ace files that were saved prior to the edits will remain unchanged (i.e. still contain the unedited sequence)

5. *Other Useful Information:*
- ♦ P0 and Pfos1 are the files that contain the vector sequences used. They are usually masked as X in the assembly.
- ♦ The *.contigs file within *edit_dir* represents the consensus sequence, the *.fasta.qual contains information on the quality of each base.
- ♦ The *status* file can be found in the *edit_dir*. It contains information about the assembly and is useful as an initial check on the quality of the assembly (i.e. compare predicted size versus known size).