

An Introduction to Finishing Using the Consed Platform (Part I)

Developed by Andrew Nylander from notes by Chad Tomlinson.

Revised by William Barshop and Wilson Leung.

Files for this Tutorial:

All files for this tutorial are contained within 16B18.tgz

Introduction:

Advances in genetic sequencing technology have allowed for highly automated generation of raw sequence data from a wide variety of organisms ranging from the simple yeast to humans, but despite these advances human labor and skill are still required to insure the quality of the data in the finishing phase. Finishing is the final step in the sequencing process in which human finishers identify and resolve low quality data by editing sequences and calling further sequence reactions in order to generate more sequence data for the regions in question. Of course, no sequence is ever truly “finished,” but we can bring the data set up to a high standard, estimated as one error per 1000 bases.

This tutorial will examine basic finishing techniques that you will use later on in the course to finish the assigned fosmids. For finishing we will be using a program called Consed that allows the user to view and edit the sequence assembly.

Note: This tutorial is based on the use of a 3-button mouse.

For a one-button mouse:

Left click: one-button mouse click

Center click: ‘alt option’ + one-button mouse click

Right click: ‘⌘’ + one-button mouse click

Viewing the Assembly:

For this part of the tutorial we will be using the 16B18 project. Open a new xterm and navigate to the ‘edit_dir’ of the 16B18 directory. In order to do this simply type ‘cd ~/16B18/edit_dir’ in to the xterm window. Be sure to include a space between cd and ~ when you type the command. Type ‘consed &’ to launch the program and open the **16B18.fasta.screen.ace.1** file to view the initial assembly. Ace files are always listed in chronological order with the most recently saved version listed first and the others in order below. When you click on the .ace file you wish to open a dialog box will pop up with a history of the work which you can recover if you desire. In this case we will simply select no. After this we should see the Consed Main Window appear with all the contigs in the assembly of that file listed from largest to smallest under the Contig List box in the window (Figure 1).

When you are done with your work and wish to end a session simply go to the ‘File’ menu on the Consed Main Window and select ‘Save Assembly.’ A window will pop up and suggest a name for the file and it is highly recommended that you use the name that is suggested. Consed tries to name the files in chronological order so that (project name).fasta.screen.ace.1 is older than (project name).fasta.screen.ace.2 and so

forth. To exit Consed go to 'File' in the Aligned Reads Window and select 'Quit Consed.' To start, we will need to examine the entire fosmid assembly in order to see how many contigs are in the assembly, and to provide us with a general picture of the problem areas within the sequenced fosmid. To do this, just click on the 'Assembly View' button in the Consed Main Window.

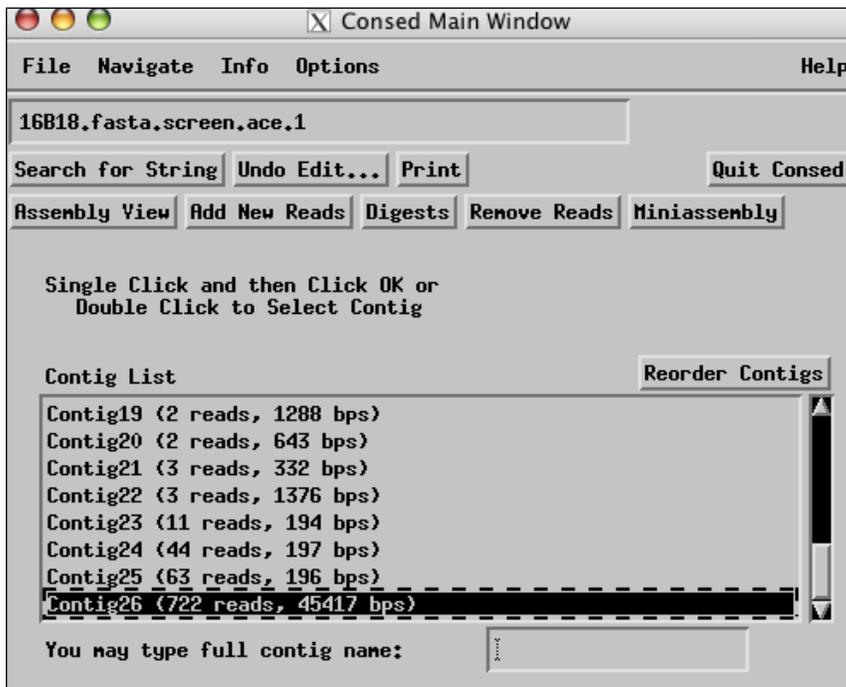


Figure 1. View of the Consed Main Window

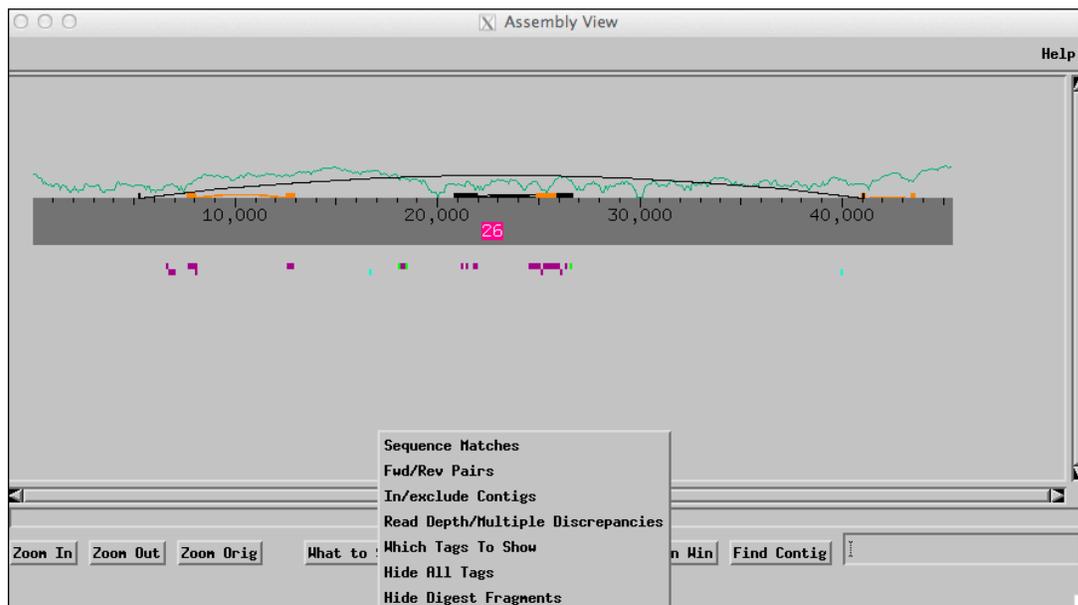


Figure 2. View of entire assembly

From this view we can see that this project is already assembled into one large contig. The dark green line corresponds approximately to the number of high quality (phred 20 or higher) reads for each base. As you edit the sequence and order more reads to improve the quality of the sequence, the high quality line should increase in height. In order to view tandem repeats and inverted repeats we click on the ‘What to Show’ button at the bottom of the Assembly View window and select ‘Sequence Matches’ from the menu. A dialog box should appear and then click ‘Run crossmatch’ using the default settings shown below (Figure 3).

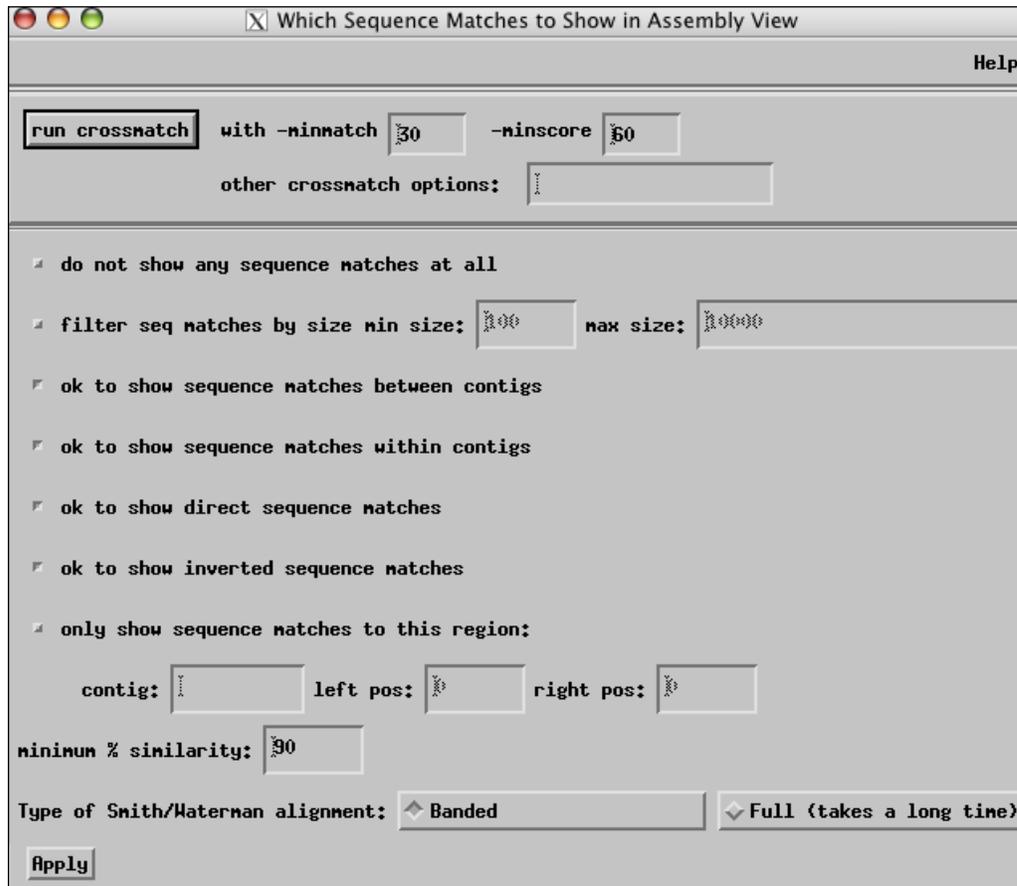


Figure 3. Configuration dialog box for crossmatch

The tandem repeats will appear as orange blocks and connecting lines, as seen in Figure 2, while the inverted repeats are represented as black blocks and lines. To view a particular repeat in the contig simply left click on an orange or black box and a ‘Sequence Match’ box will pop up with positions of each copy of the repeat as well as repeat size and percent similarity between the two repeats.

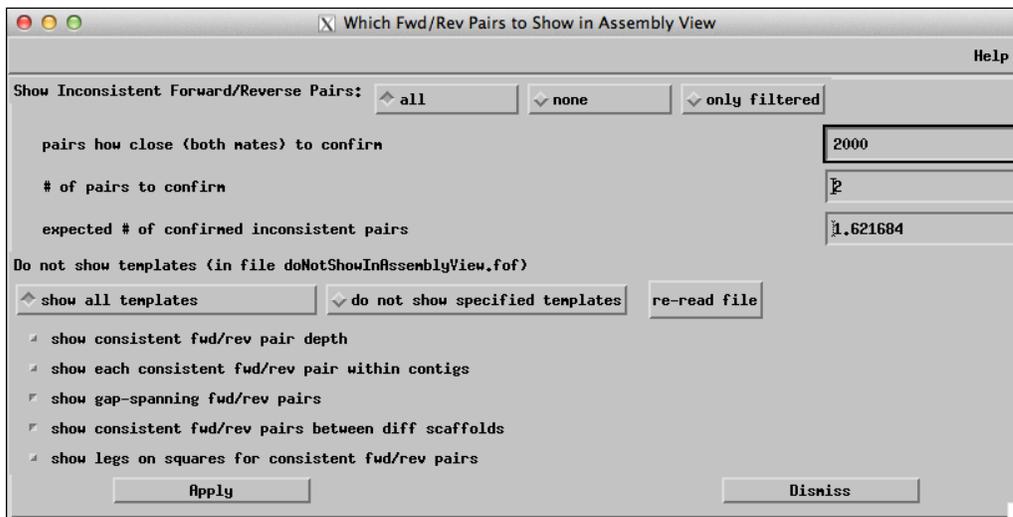


Figure 6. Fwd/Rev Pairs dialogue box

The inconsistent forward/reverse pairs will appear as red lines or purple lines underneath the contig in the assembly view (Figure 7).

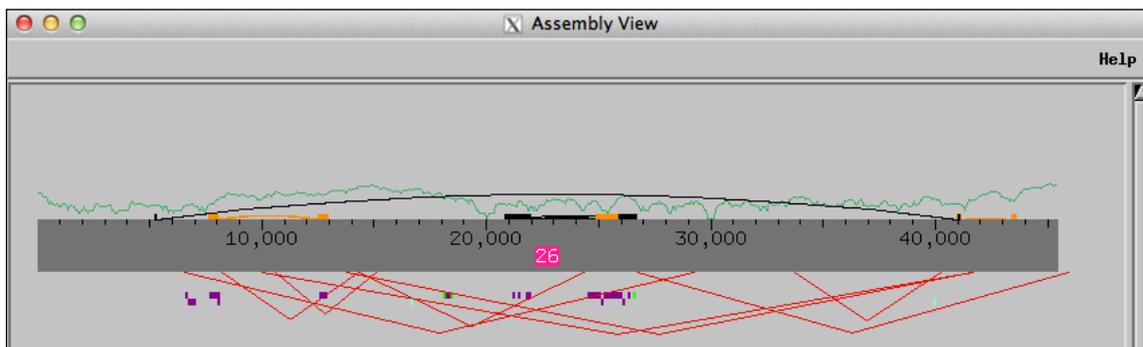


Figure 7. Assembly View with inconsistent forward/reverse pairs

Navigating through the Assembly

To further assess the project we will next look at the regions with low quality sequence, high quality discrepancies, single strands or single chemistries. Right click on any region of interest within the gray bar of the contig and select the 'Go to Aligned Reads' option from the drop down menu. The 'Aligned Reads' window will pop up and we will click the 'Navigate' button on the top bar. For initial assessments of the fosmid it is best to select 'Low Cons/High Qual Discrep/Single Stranded/Single Subclone/Unaligned High' option from the drop down menu (Figure 8). This will allow us to see a list of all the problems within the assembly (Figure 9). To view any of these areas in the Aligned Reads window simply select the problem that you wish to see and click 'go' at the bottom of the window.

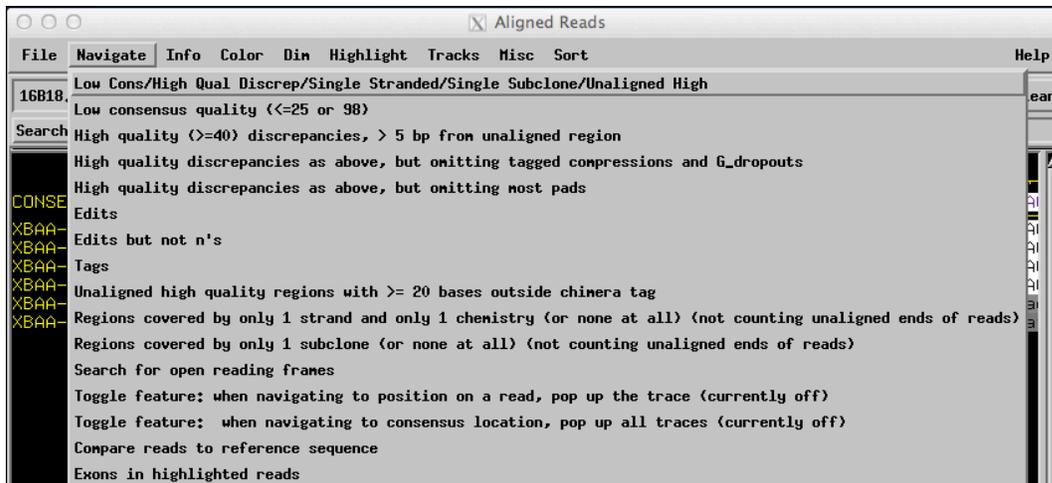


Figure 8. Navigating in the Aligned Reads window

Contig Name	Read Name	Consensus Positions	Description
Contig26	(consensus)	560-1399	854 bp single strand/chem
Contig26	XBAA-aaf38f06.b1	1847-1969	123 unaligned high quality
Contig26	(consensus)	3273-3481	210 bp single strand/chem
Contig26	(consensus)	5292-5319	28 bp single strand/chem
Contig26	XBAA-aaf39b02.b1	12006	high quality base disagrees with consensus
Contig26	XBAA-aaf40g02.b1	12565	high quality base disagrees with consensus
Contig26	XBAA-aaf40g02.b1	12741	high quality base disagrees with consensus
Contig26	XBAA-aaf49a07.g1	16333	high quality base disagrees with consensus
Contig26	XBAA-aaf46f03.g1	17597	high quality base disagrees with consensus
Contig26	(consensus)	19559-20399	842 bp single strand/chem
Contig26	(consensus)	20049	base quality below threshold
Contig26	(consensus)	20051-20060	base quality below threshold
Contig26	(consensus)	20051-20090	40 bp single subclone
Contig26	(consensus)	20065	base quality below threshold
Contig26	(consensus)	20067-20074	base quality below threshold
Contig26	(consensus)	20082-20083	base quality below threshold
Contig26	(consensus)	22172-22515	345 bp single strand/chem
Contig26	(consensus)	29861-30044	191 bp single strand/chem
Contig26	(consensus)	29951-29953	base quality below threshold
Contig26	(consensus)	29970-29971	base quality below threshold
Contig26	(consensus)	29977	base quality below threshold
Contig26	(consensus)	29979-29983	base quality below threshold
Contig26	(consensus)	29987-29992	base quality below threshold

Figure 9. List of all low consensus quality areas, high quality discrepancies, single stranded, and single chemistry regions in the assembly

If you want to navigate to problems of only a certain type, such as low quality or high quality discrepancies, then you may select a more specific navigator from the list seen in Figure 8. *Note: To view only high quality discrepancies, be sure to select the 'High quality discrepancies as above, but omitting tagged compressions and G_dropouts' option. Also, when viewing high quality discrepancies, be sure to take note of any regions with a large number of high quality discrepancies, this could be an indication of a possible misassembly.*

We can also scroll back and forth through the ‘Aligned Reads Window’ in order to simply view stretches of sequence data. To do this you can either drag the thumb of the scrollbar near the bottom of the ‘Aligned Reads Window’ or use the ‘<<,’ ‘>>,’ ‘<,’ ‘>,’ buttons underneath the scroll bar (Figure 10).

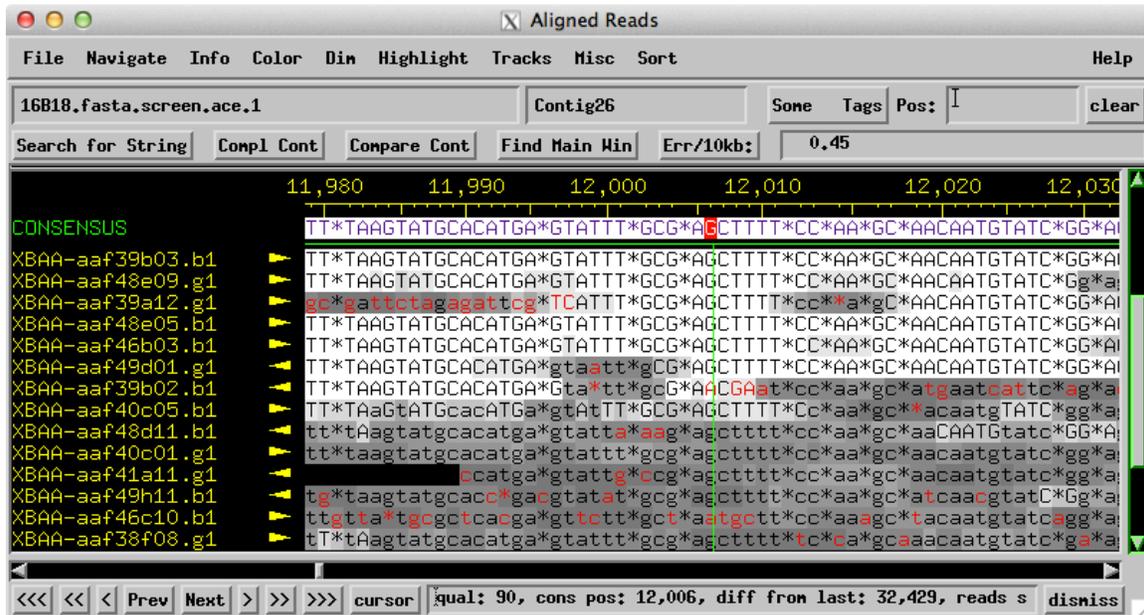


Figure 10. Aligned Reads Window with scroll bar and scroll buttons in bottom left hand corner

The ‘<,’ ‘>,’ buttons only scroll small amounts (25 bp) while the ‘<<,’ ‘>>,’ scroll larger amounts (50 bp). The ‘<<<,’ ‘>>>,’ buttons will take you to either end of the contig. When you are scrolling you can also click on a read name on the far left end of the screen and it will turn magenta. This highlighting of the reads can be very helpful in visually keeping track of the read as you are scrolling through the sequence. *Note: It is better to use these buttons to find the ends of the contig as opposed to using the scrollbar because in most assemblies there are reads which extend beyond the length of the actual contig. By scrolling with the scrollbar you will often over shoot the contig and instead go to the end of the farthest read, while using the ‘<<<,’ ‘>>>,’ buttons ensures that you go to ends of the actual contig.*

In the Aligned Reads Window you can also move around by using the ‘Pos:’ box in the upper right corner. To use this function simply type in the base number you wish to view and the window will shift to that position. This is very handy when you wish to quickly show a part of the sequence to another student or instructor.

Editing Bases:

Next, go through the errors listed in the navigator and look at each area in the Aligned Reads window. To do this, simply select the error listed in the navigator that you wish to view and then click on the ‘go’ button in the lower left-hand corner (see Figure 9). An aligned reads window should appear, positioned over the error. Red bases in the reads indicate that the base in that particular read disagrees with the consensus

sequence. Also the different shades of background for the reads serves as an indicator for the quality of a read at a particular position. White is the highest quality for a read. Lower quality regions of read sequence data are indicated a darker background. As the background goes from light grey to dark grey to black, the quality of the read drops with black being the lowest level of quality. Reads usually have black backgrounds at their ends where the sequencing quality is extremely poor. In the 'Aligned Reads Window' we have the option to change the background of the low quality regions depending on your viewing preference. Click on the 'Dim' option at the top of the 'Aligned Reads Window' and select 'Dim Nothing.' You should see that the discrepant bases in the low quality areas become red. This however can be a somewhat distracting and so we normally choose to view the reads with 'Dim Low Quality or Unaligned' option selected from the 'Dim' menu (Figure 11).

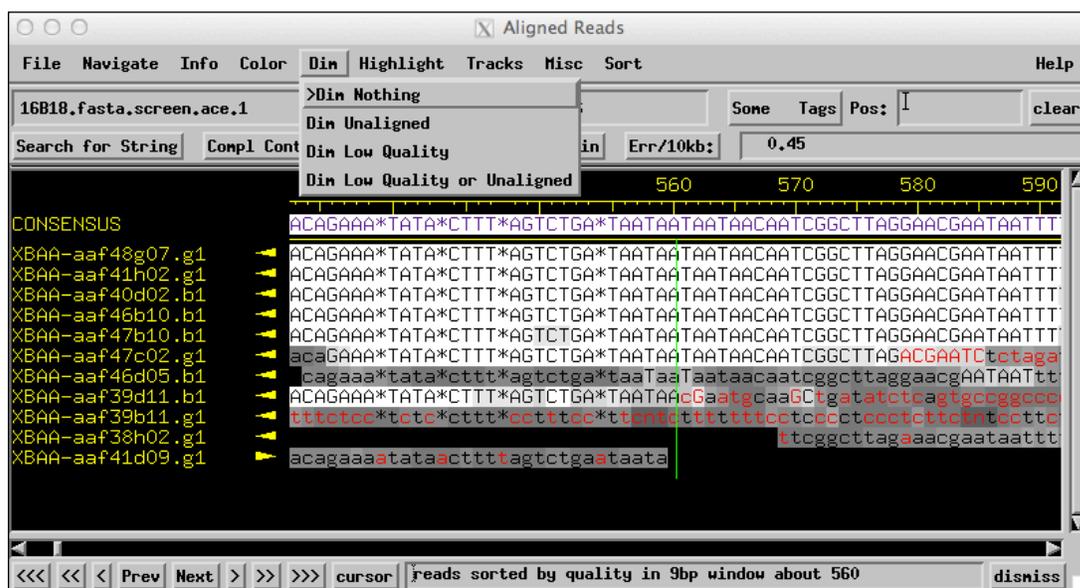


Figure 11. 'Aligned Reads Window' the 'Dim Nothing' option selected

For some errors, particularly high quality discrepancies, we can often manually correct the problem by making edits to the discrepant read. To edit the read, right-click on the questionable nucleotide in the Aligned Reads window and select 'Display traces for all reads' from the menu; a window containing the traces should appear (Figure 12).

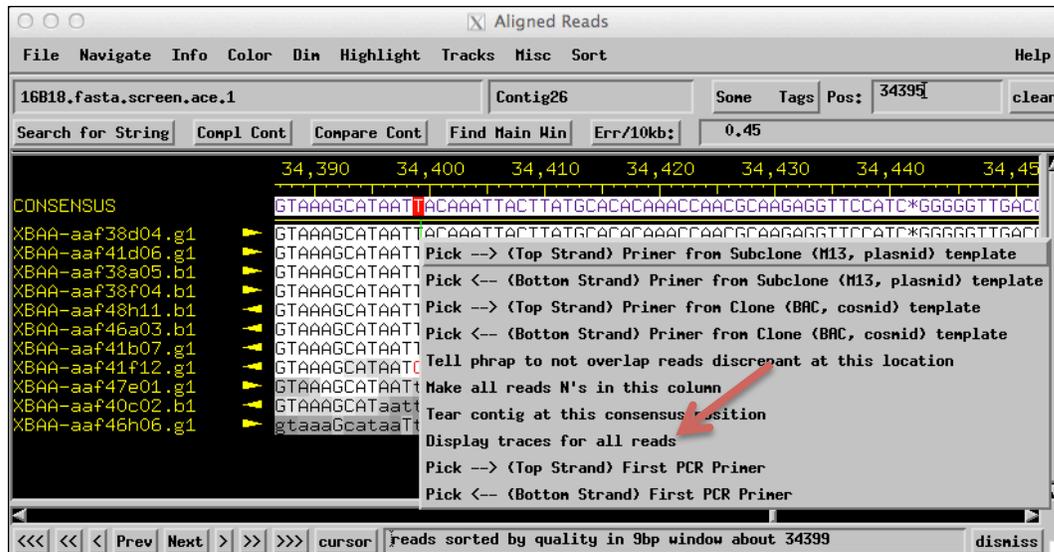


Figure 12. Aligned Reads window

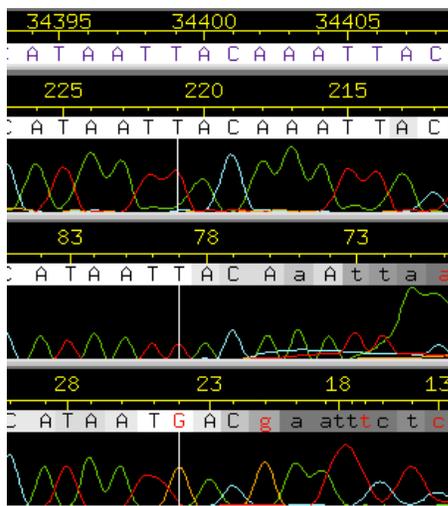


Figure 13. Traces Window

In the Traces Window we can compare the quality of peaks from each read and see if we can confidently make an edit to the discrepant read (Figure 13). From the example in Figure 13 there are well-defined red-colored peaks in most of the reads that indicate a T nucleotide at that position. However, when we examine the read XBAA-aaf41f12.g1 there is a G peak instead of a T. When dealing with a discrepancy like this it is important to look at the traces both up and down stream of the nucleotide of interests. If the quality of the trace is low in either direction around the position of the discrepant nucleotide, then it is highly probable that the base was miscalled and mislabeled as high quality. This situation often arises near the beginning and end of reads. If this is the case, you can edit the incorrect base so that it matches the consensus. To make an edit, left click on the base you wish to edit and type the correct base. In most cases, you should use lowercase letters when making edits. This is because uppercase edits indicate that the base has the highest quality and phred likely miscalled the base because of

ambiguity in the trace. A yellow edit tag will be added to the base you have edited. The edited base will appear in red if it disagrees with the consensus and in black if it matches the consensus (Figures 14a, 14b).

You may be able to resolve many high quality discrepancies by simply editing the discrepant base. However, you should only edit a base if the trace supports such an edit. One should not edit a base simply to resolve a discrepancy and match the read to the consensus sequence. Genuine high quality discrepancies can indicate either the presence of polymorphisms in the genome or the presence of inexact repeats that have been mistakenly assembled together.

The edit made in Figure 14a is an example of an incorrect edit. In this example, the read was edited to show an “a” at position 24; however, the trace clearly shows a “g” base (g bases appear in orange within the trace window). Hence, this edit should not have been made. The example in Fig. 14b shows a case where the consensus base is a “t,” but the read has been called “g.” In this case, the trace data shows a broad G peak that spans multiple bases but there are two distinct high quality T peaks at approximately the same height as the G. Consequently, we would edit the “G” base call to a “T”.

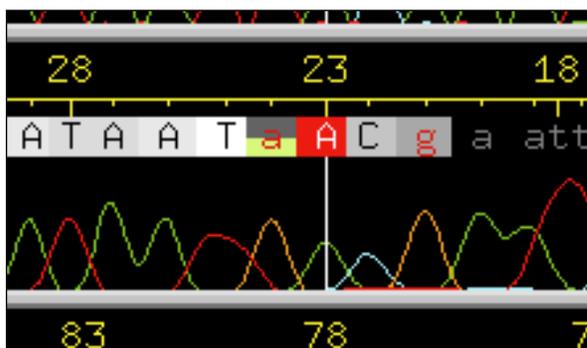


Figure 14a. Example of an incorrectly edited base. The trace data shows no evidence for an edit to “a.”

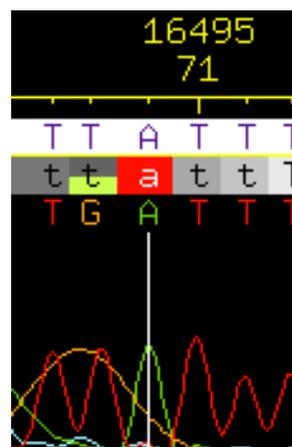


Figure 14b. Example of a correctly edited base. The trace data shows sufficient evidence for an edit to “t.”

We can also insert pads (represented by the “*” character) into the reads if we found that two reads are identical except for one or more missing bases. We add pads to fill in for missing bases in order help improve the consensus quality by aligning it properly with the rest of the reads. For example, if we have two sequences such as:

```
GTTGACAGTAATCTA
GTTGACATAATCTA
```

We can insert a pad where the g should be in the second sequence in order to have the two sequences align properly:

```
GTTGACAGTAATCTA
GTTGACA*TAATCTA
```

In order to insert a pad into a read we simply push the space bar to add a * into the selected read just like when we edited bases in the 'Traces Window' (Figure 15). If you need to undo either a pad or a base edit, click on the 'Undo' button on the left side of the window below the names of each of the reads.

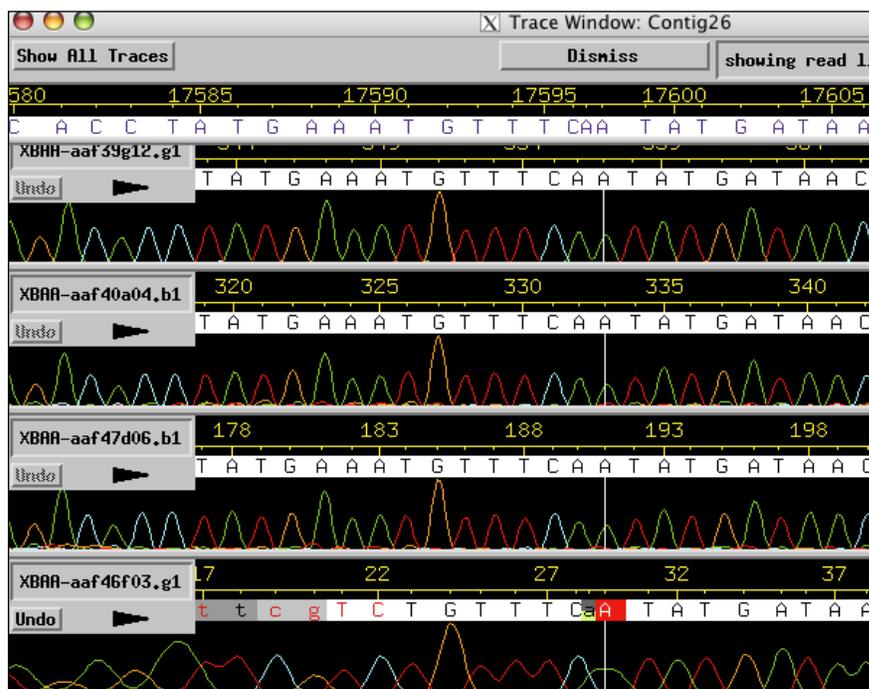


Figure 15. Traces Window

If you would like to delete a base then you can simply overstrike the base with a '*' just as if you were making an edit in the 'Trace Window.' Phrap ignores the '*', which means that the base has been effectively deleted but the '*' provides the finisher with a visual indication of a base deletion (Figure 16). Consed will also automatically remove the pads when you export the consensus.

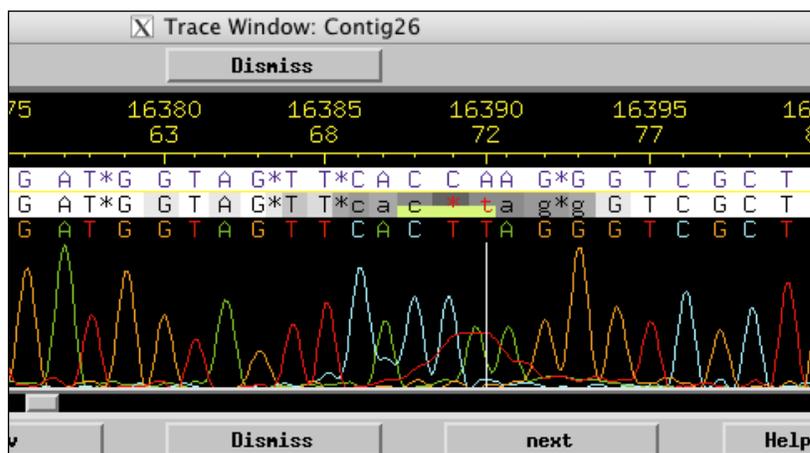


Figure 16. Example of a deleted base in the 'Traces Window'

Tags:

In Consed we can apply different tags to regions of the read sequences and these can have a variety of functions. To apply a tag to a region within a read, middle mouse click on the read you wish to tag to open a trace window for just that single read. We can then highlight this region by clicking with the middle mouse button, holding down the button and swipe over the span of bases you wish to tag (Figure 17). At the end of this region, release the middle mouse button and a window will pop up with a variety of different tags to choose from.

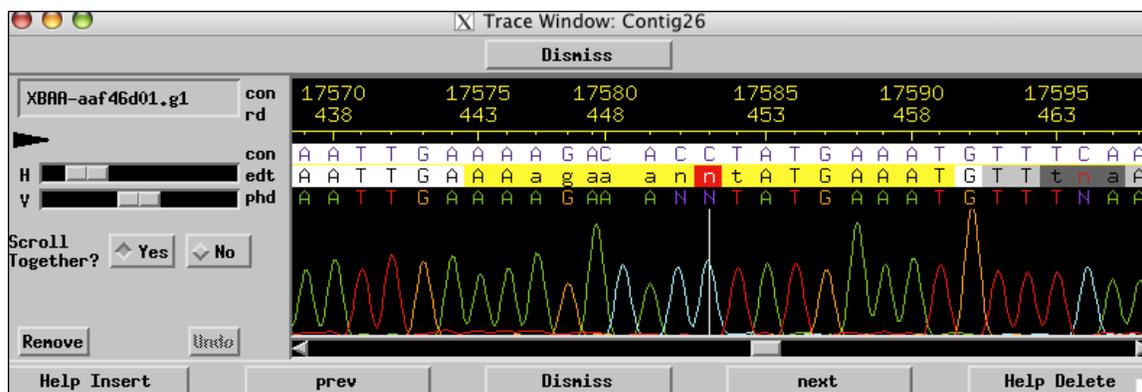


Figure 17. Traces window for single read

These options include a variety of functions that we may or may not need use in finishing a fosmid. After a tag has been added, the tagged region will be partially colored. The color of the tag will vary depending on the tag type. Please see Appendix A for a complete description of all the tags and their functions.

Calling Oligos:

For some problem areas, such as low quality regions and single-stranded regions of sequence, you have the option to mark these areas with a 'dataNeeded' tag. This will allow you to find these regions easily when you are designing oligos. To do this, go to the region that you wish to mark in the Aligned Reads window and move your pointer to the position in the consensus where you wish to start the tag; then hold down the middle mouse button and drag to where ever you want the tag to end. When you let go of the middle mouse button the 'select tag type' window and click on 'dataNeeded.' Later you can right click on the tag to view more detailed information about the region and add any comments you would like about the region covered by the tag (Figure 18).

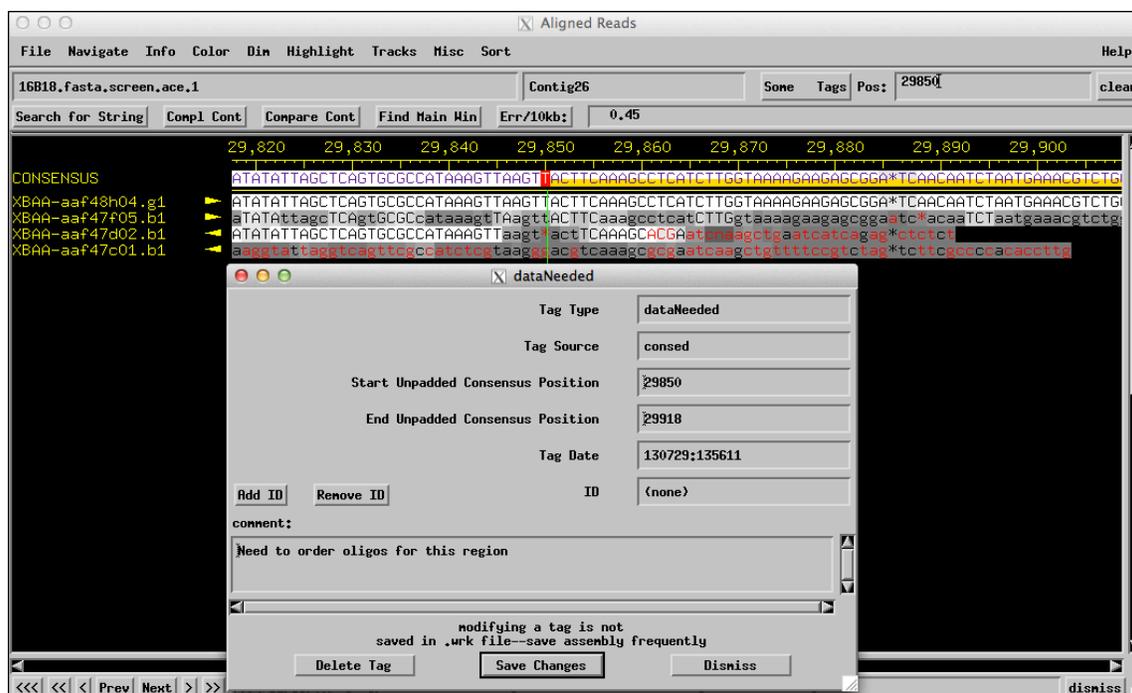


Figure 18. Tagged region with details window

Once you have identified all of the regions that will need additional data the next step is to call oligos for those areas. To select an oligo, go to the closest end of a high quality sequence near the region you want to re-sequence and right-click on a read. Select either 'Pick -->' or 'Pick <--' depending on which direction you wish to sequence (Figure 19).

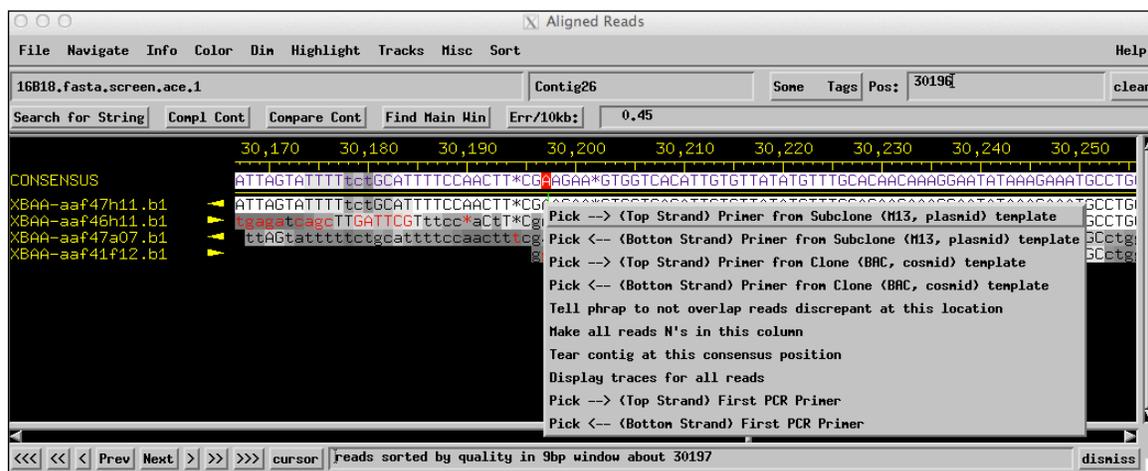


Figure 19. Selecting a sequencing orientation

Because we are using the fosmid DNA as the template for sequencing, you should use the 'Primer from Clone (BAC, cosmid) template' options. This option uses more stringent primer selection parameters. Once you select an option from the menu a

window with a list of possible primers will appear. The first 50 bases of a sequencing read tends to be low quality, so when choosing a primer be sure to select one that is about 70 bases away from where you want to be able to read sequence data and that has a melting temperature of 55° C or above. When you click on a primer the Align Reads window will move to the position of the primer's complementary sequence. To ensure that this primer sequence is unique to just this location, and will not bind to other similar sequences in the fosmid, you should do a 'Search for String'. Select about 10 to 12 bases from the complementary primer sequence by clicking and swiping the region with the left mouse button. Then click on 'Search for String' in the upper left corner of the Consed Main Window. In the 'Search for String' box, middle click in the query box to paste the sequence that you highlighted, select the 'exact' option from the search parameters and click 'OK' (Figure 20).

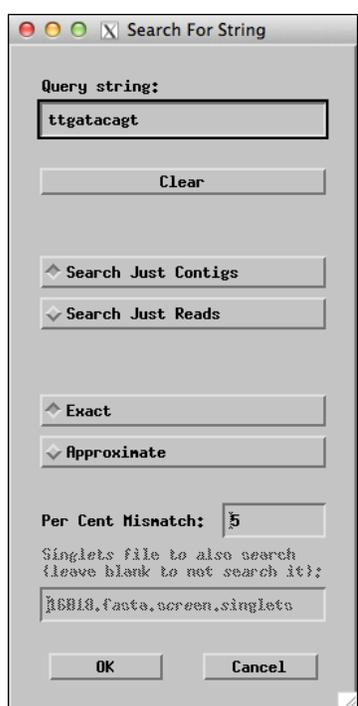


Figure 20. Search for String box

If your search returns only sequence at the position specified for your proposed primer then go back to the primers window and click on 'accept primer' (Figure 21). A comments box will appear and you can tag the primer with a note to remind yourself later what the purpose of the primer is (what sequence problem you aim to resolve) and which chemistry you want to use. Because you will likely be calling many reactions, especially for the first round of finishing, taking the time to make careful notes in the comments box will save you both time and frustration when it comes time to submit your reaction orders; this procedure is highly recommended.



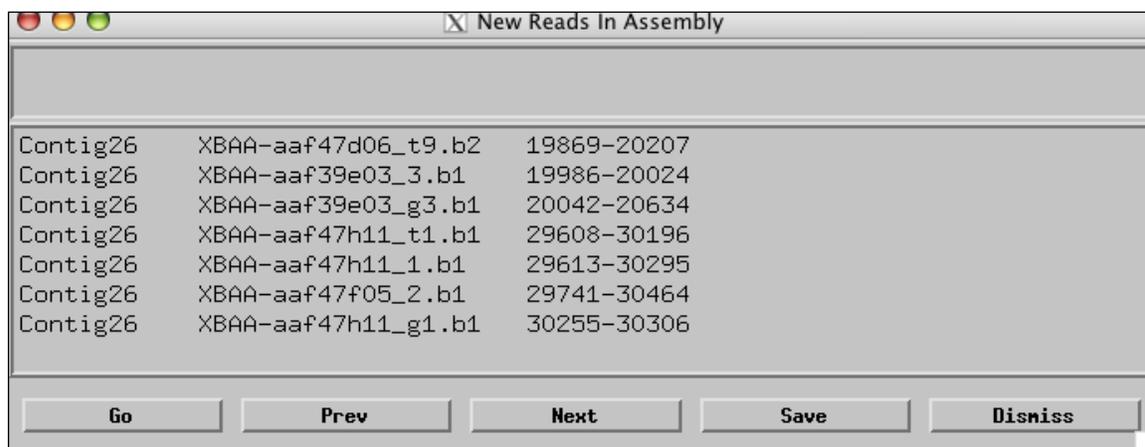
Figure 21. Region tagged for an oligo in Aligned Reads window

Note: You may want to stop following the text here and put together your own list of problem areas and the oligos you would pick to resolve the problems. Keep this as a table in a word document to compare with those the finisher designed in figure 22.

Contig Name	Read Name	Consensus Positions	Comment	Oligo Name	Oligo Templates
Contig26	(consensus)	3738-3753	reverse primer	16B18.4	XBAA-aaf46a07...
Contig26	(consensus)	5105-5129	forward direction	16B18.5	XBAA-aaf41d08...
Contig26	(consensus)	19932-19954	forward direction	16B18.3	XBAA-aaf39e03...
Contig26	(consensus)	20211-20234	reverse direction	16B18.6	XBAA-aaf40a04...
Contig26	(consensus)	29700-29715	Forward reaction	16B18.2	XBAA-aaf47f05...
Contig26	(consensus)	30327-30348	reverse direction	16B18.1	XBAA-aaf47h11...
Contig26	(consensus)	32177-32195	reverse direction	16B18.7	XBAA-aaf46h11...
Contig26	(consensus)	38314-38336	forward direction	16B18.8	XBAA-aaf39h05...

Figure 22. Oligos called by the original finisher

After you have ordered the reactions and the new sequencing results have come back from the Genome Institute you will need to generate a text file of the reactions that can then be used to add the new sequence into the project. For this example project go to the edit_dir file and find the file called 'reads_to_add.fof'. This file contains list of new sequence reads we wish to incorporate into our assembly. To add the reactions into the current assembly simply click on 'Add New Reads' in the Consed Main Window. A window should appear with a list of files in edit_dir; select the reads_to_add.fof file. Two windows will appear in succession; click yes on both of them. After this a new box called 'New Reads in Assembly' will appear (Figure 23).



Contig	Read File	Read Range
Contig26	XBAA-aaf47d06_t9.b2	19869-20207
Contig26	XBAA-aaf39e03_3.b1	19986-20024
Contig26	XBAA-aaf39e03_g3.b1	20042-20634
Contig26	XBAA-aaf47h11_t1.b1	29608-30196
Contig26	XBAA-aaf47h11_1.b1	29613-30295
Contig26	XBAA-aaf47f05_2.b1	29741-30464
Contig26	XBAA-aaf47h11_g1.b1	30255-30306

Buttons: Go, Prev, Next, Save, Dismiss

Figure 23. New Reads in Assembly box

Return to Aligned Reads Window and check the new list of problem areas using the ‘Low Cons/High Qual Discrep/Single Stranded/Single Subclone/Unaligned High’ option. Have we made progress in generating a high quality assembly?

Checking for contamination

In your final check of your consensus sequence, you will want to confirm that there is no vector or bacterial contamination. In order to do this we will need to export the sequence for use in a BLAST search. Select ‘Export consensus sequence’ from the ‘File’ menu in the Aligned Reads Window. The sequence of the contig that you are viewing in the Aligned Reads Window will be saved as Contig#.fasta where # corresponds to the contig number in the Consed database. For example, the sequence from contig 26 will be saved as Contig26.fasta. Once this is done, there will be a file in the edit_dir directory with that name. This sequence file is just a plain text file so you can open it in any text editor (e.g. TextEdit on the Mac, gedit, Emacs, vim in Linux). You can also view the sequence within the xterm by issuing the following commands:

```
cd ~/16B18/edit_dir
```

```
more Contig26.fasta
```

You can now copy this sequence and perform a BLAST search (against the Microbes RefSeq Genomes database) to see if any of the sequence comes from the vector.

You can also export part of the consensus sequence. Go to the ‘File’ menu in the Aligned Reads Window and select ‘Export consensus sequence (with options)...’ A window will pop up which will allow you to specify the exact region of the consensus sequence you wish to export (Figure 24).

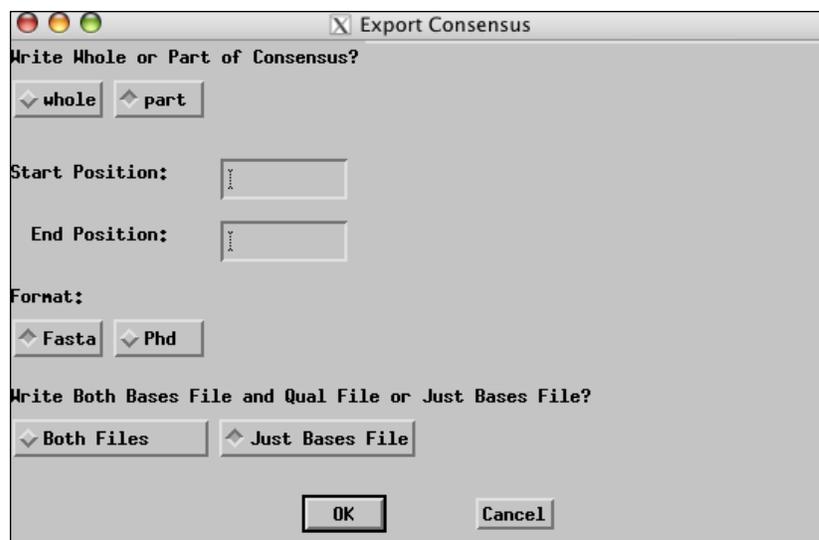


Figure 24. Dialog box for controlling which region and data formats of the consensus sequence to export.

Once you have specified the region you want to export select the ‘Both Files’ option and click ‘OK.’ By default, this file will be saved as ‘Contig#.fasta’ as before. If the file already exists, a dialog box will appear that asks if you want to overwrite the existing file. You can examine the contents of the new files using any text editor. You can also inspect the file using the xterm by typing the following commands:

```
more Contig26.fasta
```

```
more Contig26.fasta.qual
```

Last Update: 07/29/2013