

Phred, Phrap, and Consed

UNIX COMMANDS

ls

This command is used for listing the files and directories in any directory.
ls at the command prompt and you get the list of all the files and folders in that directory.

mkdir

This command is used for creating directories.
mkdir abc, the directory *abc* will be created in the current directory.

cd

This command is used for changing from one directory to another. If you are working in the directory *abc* and you want to change to the directory *cba*, you need to type the command *cd cba*. If the directory *cba* is not present within the directory *abc*, you have to mention the entire path location to the directory *cba* starting from the root directory

pwd

If you have reached a directory through various other directories and you are not aware of the path to the directory, you can simply type in *pwd* and get the current location path.

cp

Basic syntax of this command is *cp file1 file2*. Here, *file1* is the name of the file you want to copy and *file2* will be the name of the file when it is copied to the current directory. Remember, the original file is never hampered in this case.

mv

Can be used not only for moving the file but also for renaming it.
Suppose you want to move the file *inventory.txt* created in the directory *abc* to your current directory: *mv /abc/inventory.txt ..* The dot (*.*) at the end of the command indicates current directory. When using this command, you will be left with just one file. This command does not copy the file, but moves it from one location to another.

Now suppose that we have the file *abc.txt*, in your current directory.
mv abc.txt def.txt, the file *abc.txt* will be renamed to *def.txt*.

pwd

Can be used to determine where you are in the directory structure at any time.
pwd

rm

This command is used to remove or delete a file.
rm abc.txt, the file will be permanently deleted from the system.

less

Using this command you can see the information in a file, one page at a time.
less abc.txt, will display the contents of the file *abc.txt*, one page at a time.

head

This command is used to display the first ten lines of any file on your screen.
head abc.txt, then what you get on the screen are the first ten lines of the file *abc.txt*.

tail

tail abc.txt, will present you with the last ten lines of the file *abc.txt*.

A. Xming

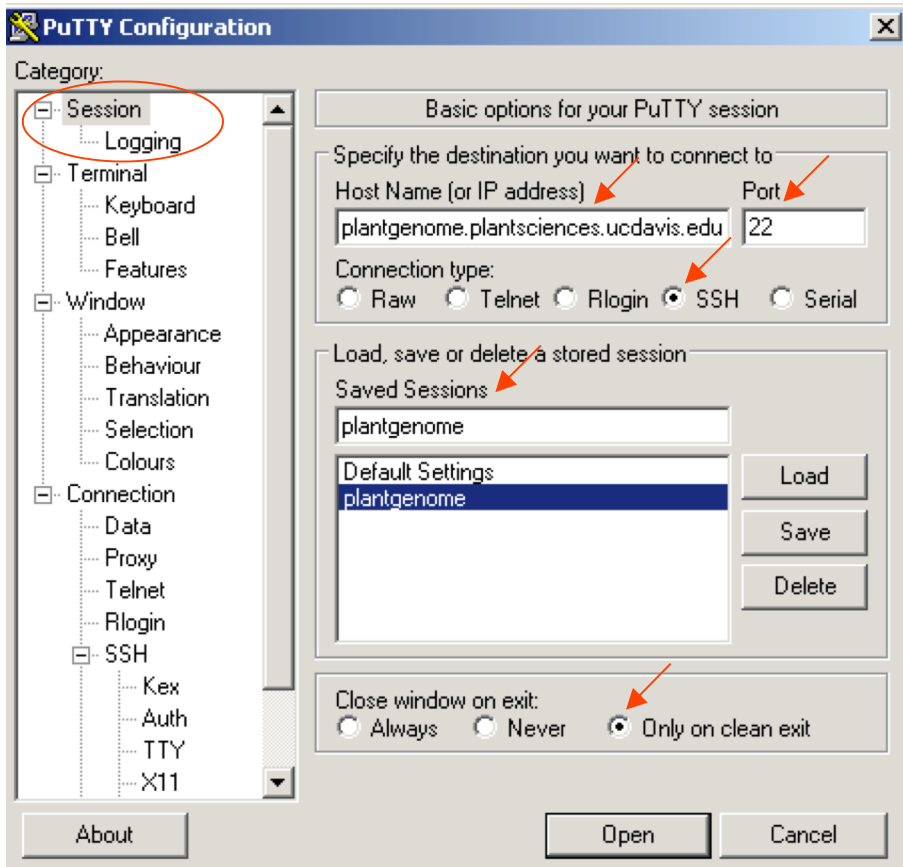
Xming is preinstalled on your lab computer (Shortcut on the desktop or in Start/Programs/BioInformatics). This software allows you to run and display UNIX applications (X clients) from the familiar Microsoft Windows environment, integrating the Windows desktop with environments such as UNIX, VMS, X Window System, IBM mainframes, and the Internet.

By using Xming to run remote applications on your local PC, you can:

- access applications and information running on networked hosts;
- establish simultaneous connections to different computers running X clients
- use a window manager to preserve your familiarity with the PC or X environment.

Start **Xming**.

Login to 'plantgenome' using **PuTTY** (be sure to have X11 forwarding enabled)



Logging in to the Plant Sciences Server

B. UNIX server connection

PuTTY is a free implementation of Telnet and SSH for Win32 and UNIX platforms. The software can be downloaded from: <http://www.chiark.greenend.org.uk/~sgtatham/putty/>.

Steps to connect to the UNIX server using PuTTY:

B.1. Open **PuTTY** by double-clicking on the shortcut on the desktop or go to Start/Programs/BioInformatics/putty. The 'PuTTY Configuration' window will be displayed:

In **Session/Logging**:

Host Name (or IP address):

plantgenome.plantsciences.ucdavis.edu

Port: 22

Connection type: SSH

Saved Sessions: plantgenome

Close window on exit: Only on clean exit

In **SSH/X11**:

Mark 'Enable X11 forwarding'

Go back to Session/Logging, click on SAVE (next time you want to connect, you will

Load 'plantgenome', your saved session, and then click on Open), and then click on OPEN.

B.2. The PuTTY' window will be displayed, and a login name and a password will be asked to you:

Username: your Kerberos username

Password: your Kerberos password

PHRED

Phred observes the base trace, makes base calls, and assigns quality values (qv) of bases in the sequence.

- a. Writes base calls and qv to output files for Phrap assembly.
- b. Quality values (phred scores) range from 0 to 60. 20 and above is considered a confident base call (1 in 100 chance that it has been called wrong ~99% accuracy)
- c. Useful for consensus sequence construction
- d. Formula: $q = -10 \times \log_{10}(p)$

where

q - quality value

p - estimated probability error for a base call

- b. Why use Phred?

Output sequence might contain errors.

- a. Vector contamination
- b. Dye-terminator reaction might not occur.
- c. Weak or variable signal strength of peak corresponding to a base.

PHRAP

Phrap constructs the contig sequence as a mosaic of the highest quality parts of the reads rather than a statistically computed "consensus".

- a. Avoids the complex algorithm issues associated with multiple alignment methods
- b. Speed and accuracy
- c. Phrap is an assembler NOT an aligner
- d. The sequence produced by Phrap is quite accurate (less than 1 error per 10 kb in typical datasets)
- e. Phrap considers sequence quality at a given position (determined by Phred)

CONSED

Consed is a program for viewing and editing assemblies produced by Phrap. Currently supports Solexa, 454, and Sanger Reads and allows mixtures of these read types

- a. Assembly viewer - allows for visualization of contigs, assembly (aligned reads), quality values of reads and final sequence.
- b. Trace file viewer - of a given sequence or several reads.
- c. Navigation – identify and list regions which are below a given quality threshold, contain high quality discrepancies, single-strand coverage, etc.
- d. Autofinish – automatic set of functions for: gap closure, improvement of sequence quality, determination of relative orientation of contigs, identification of regions covered by a single read or by reads of a single strand. The program automatically performs primer picking and chooses the templates.

INTRODUCTION TO CONSED

OVERVIEW: Consed is a program that can be used to visually assemble and analyze sequence data. This introduction will take you through the basics of opening and operating within Consed, as well as provide examples of three common actions that may be taken during the finishing process. This is meant to be a short introduction to a way that a finisher might approach a new project. This introduction requires a basic knowledge of Unix commands, basic computer skills, and the knowledge of process used to finish a project to high quality from an unassembled set of reads.

OPENING A CONSED FILE:

Change directory (cd) to the Consed file (.ace) location: Enter in a command that will change directory to the location of the Consed file you will be using. This file is in a sub-directory of **Lab2**. The **edit_dir** is a sub-directory of **Lab2** and holds the data to be used in this exercise.

cd Lab2/edit_dir

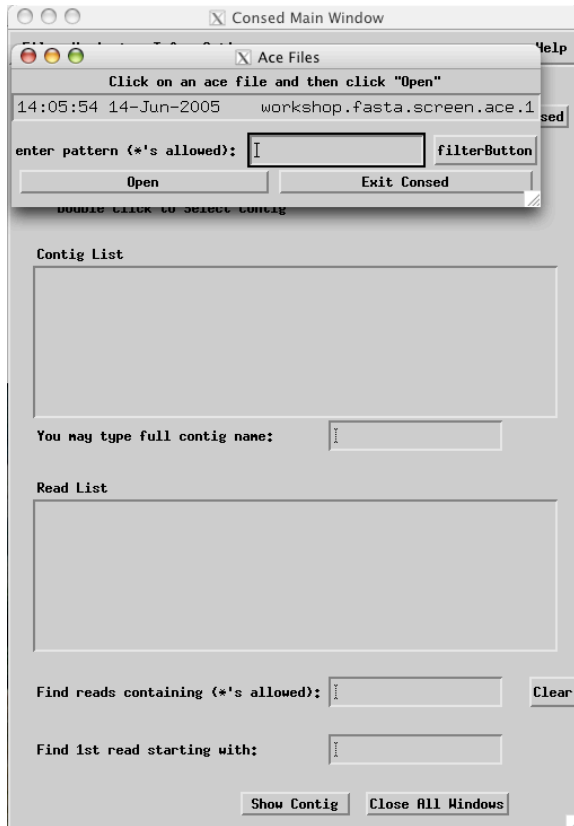
Run Consed: Start Consed by typing the appropriate command given below:

/share/apps/consed19

Two windows will appear. One of these will have the list of .ace files and will say **Click on an ace file and then click Open**.

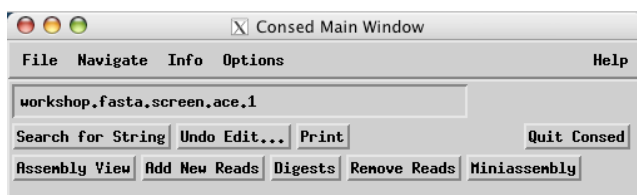
Open a Consed file: Double click on **workshop.fasta.screen.ace.1**. The .ace files (which contain the DNA sequence reads from one clone- here a fosmid) are listed chronologically with the most recently saved file listed first. After clicking, the first window goes away. You will now see a list containing two contigs and a list of reads. This is the Consed Main Window. Contigs are listed in the Consed Main Window under Contig List from the smallest to the largest. Note that there are two contigs:

Contig2 and Contig3



USING ASSEMBLY VIEW:

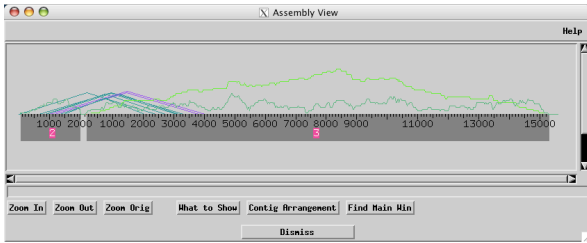
Open Assembly View: Click on the **Assembly View** button that can be found in the Main Window of Consed. Assembly View compiles a graphical representation of the major contigs and puts them in spatial relationship to one another using the initial forward/reverse pairs. Assembly View does not show contigs shorter than a set number of base pairs, and this number can be changed. Assembly View is a valuable tool used to quickly scan for problems within your assembly. If a small window appears describing a problem with sequence matches, **Dismiss** the box and continue. This will not affect any actions we will do.



The gray bar represents a single contig. The white number within the pink box indicates the contig number (which can be very useful when dealing with projects that have more than one contig at any point in time):

- Tick marks represent the number of base pairs
- Dark green line is an indication of high quality read depth at that particular location in the contig
- Light green line is an indication of total read depth (high and low quality reads)
- Green and purple triangular lines represent forward/reverse pairs that span the gap between the two contigs

Modifying Assembly View: Click on the **What to Show** button. A list of options drops down which can add detail to Assembly View allowing the finisher to examine sequence matches or more detail of forward/reverse pair locations within the assembly. Select the second option down from the top: **Fwd/Rev Pairs**. Another box will open with viewing options.



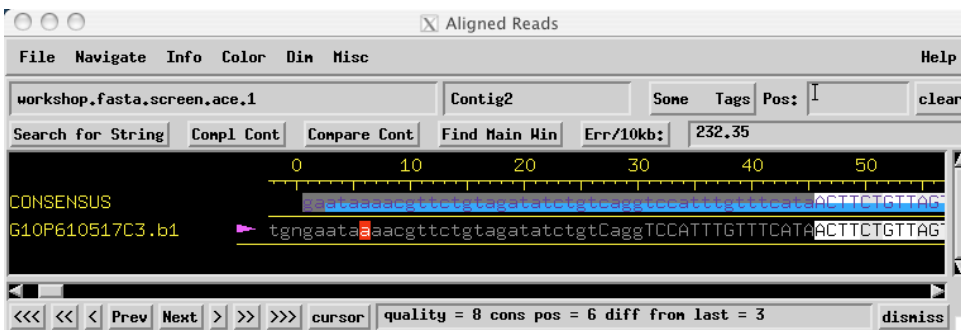
Change the following options: **Show Inconsistent Forward/Reverse Pairs**: click to select **all**. Click to select **Show each consistent fwd/rev pair within contigs**. Also, click to select **Show legs on squares for consistent fwd/rev pairs**. The buttons are **on** if the top of the diamond is shaded, and **off** if the bottom of the diamond is shaded.

Click the **Apply** button, and then the **Dismiss** button. This will return you to the modified Assembly View Window. How does the Assembly View screen change? It should show blue and purple Forward-Reverse pairs above the contigs.

Navigating from Assembly View: Right click on the gray bar at a location of interest. Select the first option: **Go to Aligned Reads Window**. This will open the Aligned Reads Window of the contig at the location specified. The Aligned Reads Window shows the alignment of all reads and the consensus sequence for a contig. This is a very common navigation tool that assemblers use to go to areas in need of more work.

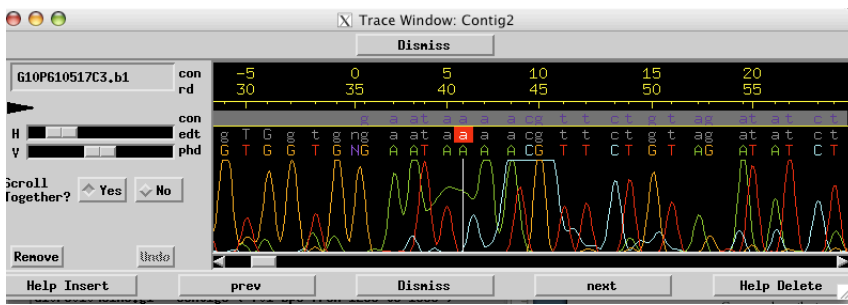
LOW QUALITY REGIONS:

Identify low quality regions: Low consensus quality means an area in which the bases have a significant probability of being wrong. The navigator saves you from having to look through large amounts of high quality data trying to find problem areas. Quality is a numerical measure of confidence made by the finishing programs. A high quality value corresponds to a high confidence in base-calling, while a low quality value corresponds to a low confidence. Numerical measures of quality are reported based on a logarithmic scale. A base-call is determined to be of 'high quality' if the quality value is equal or above 30 (often referred to as Phred 30, after the base-calling program Phred). This value can be observed in the bottom Aligned Reads Window when a base is selected by clicking with the mouse. The default visualizing parameters in Consed dim the bases in the Aligned Reads Window to indicate general quality levels. A bright white background represents a high quality base-call, fading to a black background representing low quality bases. Consed also capitalizes bases of high quality, and represents low quality bases in lowercase.



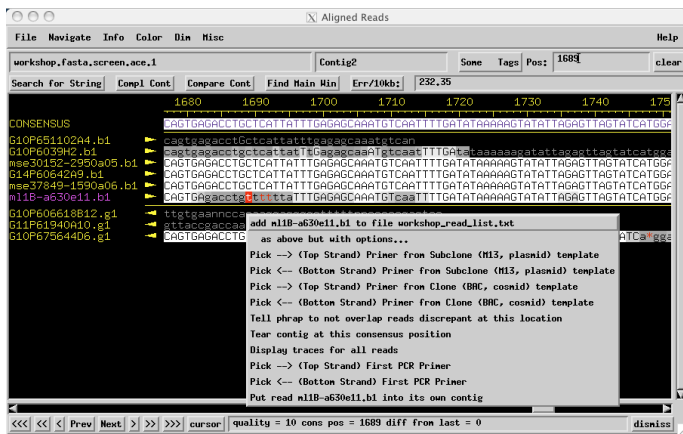
To observe how 'quality' appears in Consed, open the Aligned Reads Window for Contig2 (or by double clicking on the contig in the Consed Main Window) and scroll to the far left end of the contig.

You will notice that the first 45 bps are dimmed, indicating low quality, and that the sequence then becomes highlighted in white, indicating high quality. Open the trace for the read in this area by (center clicking) on a base from a read (not in the consensus strand); this will call up the **Trace Window**. Compare how the traces of low quality bases compare to those of high quality bases.

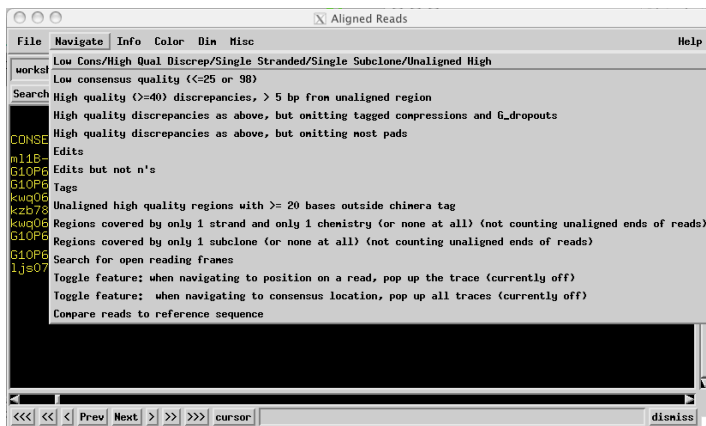


Editing a low quality region: Sometimes it is possible to resolve a low quality discrepancy or region by observing the traces for the reads. Scroll the Aligned Reads Window to **bp 1689 in Contig2**, or enter 1689 into the **Pos** (position) box and click. Open the traces for all reads at this position by clicking (right click) on bp 1689 on template **ml1ba630e11**.

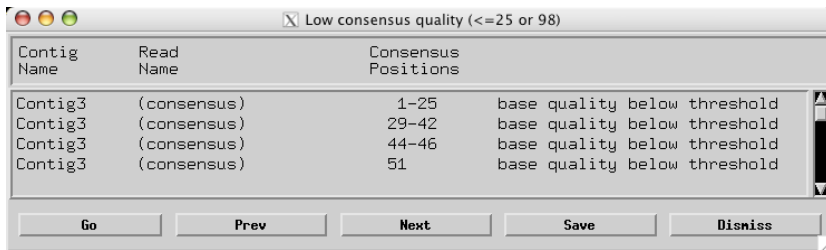
This will call up a list of options, click on an option near the bottom of the list: **Display traces for all reads**. This will open a trace window containing the traces for all of the reads at this position. Note that in template **ml1b-a630e11.b1** there are bases (visualized as red) that are discrepant with the consensus sequence. Is it possible that a sequencing error (dye blob) caused these discrepancies? (yes!) Therefore, it is reasonable to edit this base. Highlight the base by clicking either in the aligned reads window or the trace window and then overwrite the base-call with your own by simply typing the base (A, C, G, or T) that you wish it now to be. This will change the base to that which you have typed and will tag the base as **edited**. When you are finished, click the **Dismiss** button to close the window.



Navigate low quality areas: In the Aligned Reads Window, pull down the **Navigate** menu and release on the second option from the top: **Low consensus quality (<=25 or 98)**.



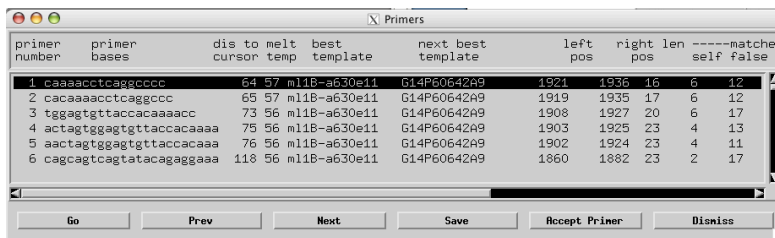
You will see a list of locations. Move the 'Low consensus quality' window down so you can see the Aligned Reads window.



Repeatedly click on **Next** until you reach the end of the list. There are 2 'Next' buttons--one on the Aligned Reads Window and one on the Low Consensus Quality Window. You can click on either, but it is probably more convenient to use the 'Next' button on the Aligned Reads Window. Thus you can keep the Aligned Reads Window in front with input focus and keep the Low Consensus Quality Window pushed out of the way. In our experience, this will be the most important **Navigate** list you will use. In fact, a major part of finishing consists of calling reads and reassembling (“re-phrapping”) with additional data.

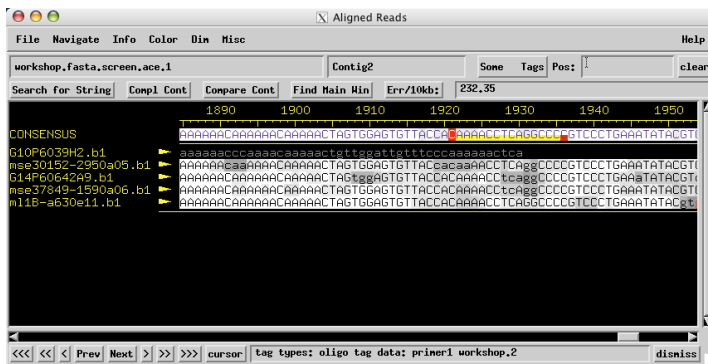
Call primers (oligos) to resolve the low quality region: Most low quality regions will need to be resolved by adding additional sequence data. Note that the low quality areas tend to occur near the beginning and the end of this contig. Normally, the high quality sequences would continue in both directions, but as we have extracted a portion of sequence for this exercise, the ends do not consist of high quality sequence. To resolve these areas, we will design new primers to be used to add data to this area, hopefully resulting in higher quality sequence. **Navigate** to the low quality area in **Contig2 at (187-195 bps)**, the second hit on the list, either by double clicking on the line containing the match, or by single clicking and hitting the **Go** button.

Right click on a base in the consensus to the right of the low quality area. Since you have chosen a primer to the right (3' end) of the area you wish to sequence, we want to choose a bottom strand primer to read sequence to the left of your selected position. This choice (bottom or top strand) is arbitrary and depends upon the position of your selected base in relation to the area you wish to sequence; however this choice can be affected by the specific sequences in each situation. (For this reaction you would likely want to avoid sequencing through the repetitious area to the left of the low quality region.) A list of options will open; click on third option: **Pick (Bottom Strand) Primer from Subclone (M13, plasmid) template.**



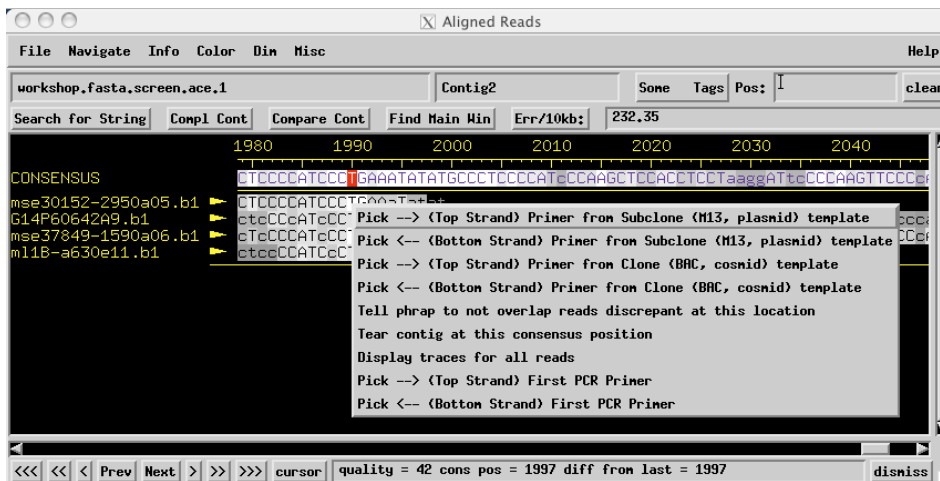
A new window will open, listing possible primers. Note that you will want to choose a primer at least 70 bps from the low quality region, as the first few bases in a typical sequencing reaction tend to be error prone. You will also want to make note of the template you will be using for this sequencing reaction. Always check to be sure that the suggested template is appropriate for the primer you choose. An appropriate template can be determined by estimating the location of the entire clone within the assembly according to the locations of the end reads in relation to one another within the entire assembly. Click on the primer you choose, highlighting the primer. Then click on the **Accept Primer** box.

A window will open asking you to “Enter Comment.” Normally you would enter a primer name here, for example **primer1** and click **ok**. The Window will close and return you to the Aligned Reads Window. Note that the primer has been tagged with colored highlight on the consensus sequence. (The 3' end is red.) If you move your mouse over (don't click) the tagged region, and look at the bottom of the Aligned Reads Window, information will appear identifying the tag as an oligo tag, and showing your comment information (“primer1”). Identify an appropriate template. To finish this contig, we will examine all the low quality areas and pick primers and templates for additional sequencing.



CLOSING A GAP:

Identify a gap: The easiest way to get a broad view of the finishing project is to open **Assembly View** from the Main Window. When two or more contigs exist in an assembly, reactions can be called to close the gap. Additional data can be added to the assembly by designing a primer (oligo) to be used in a new sequencing reaction. The goal of this reaction is to provide enough new data off the end of both sides of the gap to allow the assembly program phrap to combine the old and new data into a single contig. Primers can be designed to run off of a spanning subclone as a template for the reaction. A triangular purple line in Assembly View indicates a spanning subclone. The forward reaction must start at the end of one contig and the corresponding reverse reaction starts at the beginning of the other contig. By calling a read using one of the spanning subclones as the template, it is likely that a read will be able to close the gap. Notice that between our two contigs there are subclones that are spanning the gap.



Call an oligo: Click (right click) on the right end of **Contig2**. Then click to **Go to Aligned Reads Window**. Return to Assembly View by clicking on the window, and click (right click) on the left end of **Contig3** and click **Go to Aligned Reads Window**. Now the Aligned Reads Windows for both contigs should be open. (You may **Dismiss** the Assembly View Window if that eases viewing.) Scroll right on **Contig2** until the sequence is no longer high quality.

Click (right click) on one of the last high consensus bases you see (~1900-2000 bp) and select the first option **Pick (top strand) Primer from Subclone (M13, plasmid) template**. A "Primer" Window will open, which includes the data for suggested primers to call for the forward reaction. Double-click on a primer to navigate the Aligned Reads Window to that position. Choose a primer by clicking on the **Accept Primer** box.

Note that in many cases the best template can be one that does not currently have a read (sequence) in the area of interest. This is due to the fact that a plasmid can be identified to span a larger region than the current sequence data from that plasmid, given proper alignment of its forward-reverse pair reads.

primer number	primer bases	dis to cursor	melt temp	best template	next best template	left pos	right pos	len	----match	self	false
1	caaaacctcaggcccc	64	57	m11B-a630e11	G14P60642A9	1921	1936	16	6	12	
2	cacaaaacctcaggcccc	65	57	m11B-a630e11	G14P60642A9	1919	1935	17	6	12	
3	tggagtgttaccacaaaacc	73	56	m11B-a630e11	G14P60642A9	1908	1927	20	6	17	
4	actagtggagtgttaccacaaa	75	56	m11B-a630e11	G14P60642A9	1903	1925	23	4	13	
5	aactagtggagtgttaccacaaa	76	56	m11B-a630e11	G14P60642A9	1902	1924	23	4	11	
6	cagcagtcagtatacagaggaaa	118	56	m11B-a630e11	G14P60642A9	1860	1882	23	2	17	

A window will open asking you to “Enter Comment.” Enter a primer name here, for example **primer2** and click **ok**. The Window will close and return you to the Aligned Reads Window. Note that the primer has been tagged as before, but with a the primer name you have indicated.

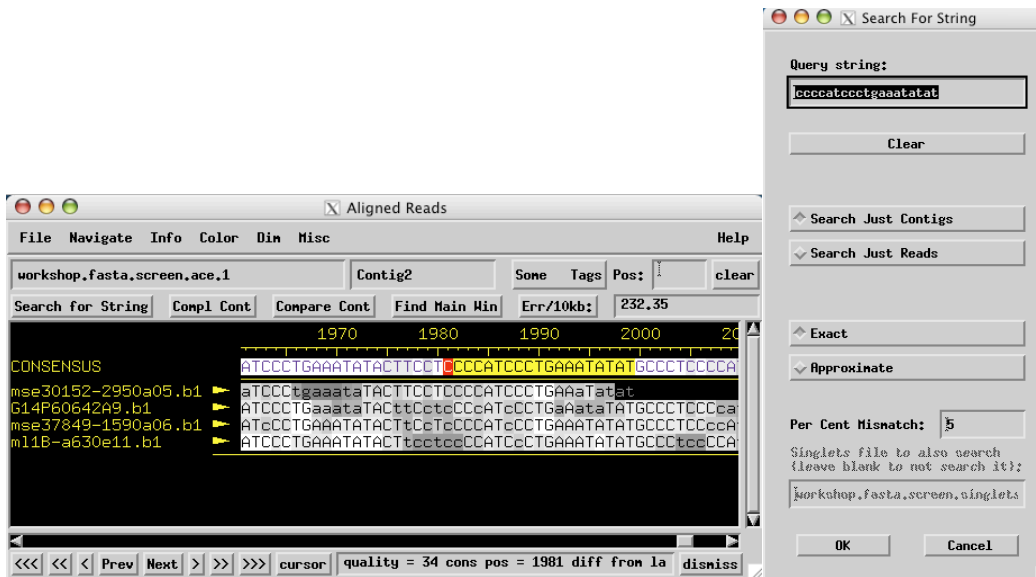
Call the reverse oligo: Click on the open Aligned Reads Window for Contig3. Scroll to the far left end of the contig and highlight one of the first high quality bases you see in the consensus (~160-200 bp) by clicking. Click (right click) on this base and select the second option down from the top: **Pick (Bottom Strand) Primer from Subclone (M13, plasmid) template**. Another Primer Window will appear; choose a primer and label it (for example: “primer3”).

If you were calling these reactions, you would order the primers you chose and request a sequencing reaction on the chosen template (usually a spanning subclone). If you are running a PCR reaction to amplify from available template, you will want to ensure that your oligos are being called for the same template, that they are a proper distance from the gap, and that the two primers have the same melt temperature. However, we will stop at this point.

MAKING A JOIN:

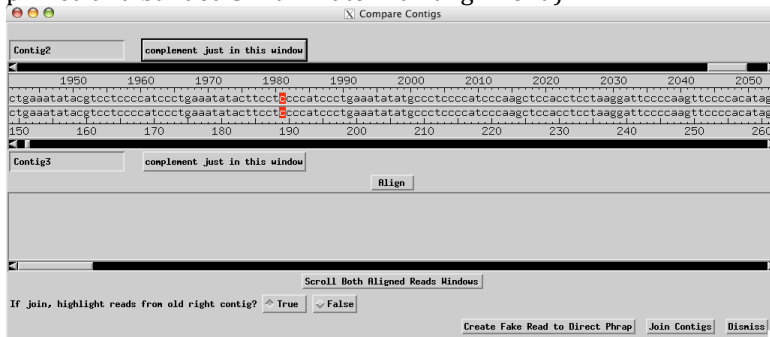
Search for String: Return briefly to **Assembly View**, note again that many subclones span the gap between the two contigs. This suggests that the gap is not large, and that the contigs may be joined without calling for any new reactions by performing a **Join**. When trying to join two contigs together, a finisher tries to find overlapping end regions. This is done by using a **Search for String** (left side of the Aligned Reads Window). Open the **Aligned Reads Window** for Contig2. Scroll to the far right end of the contig (the end that Assembly View suggests may join with the left end of Contig3). Stop scrolling when you reach the end of high quality bases on the right side of the contig (~1950-2000 bp). Click on the **Search for String** button and type in a sequence (a string) from the right end of Contig2. Alternatively, you can easily copy the consensus sequence you want by clicking and then dragging the mouse over the bases you wish to copy (this will highlight the bases). (Your query should be about 20 bases.) Now click on **Query string**; then center-mouse click to paste the sequence into the **Query string** box. You can use a sequence of your choice (~20 bases) or 1981-1999 bp: **ccccatccctgaatatat**. Click **OK** in the Search for String box. Note that there are options that can be changed that affect the search parameters if desired.

The **Searching Contigs** box will open in a new window listing the results of the search. Double click on the hit on **Contig2** (or single click on it and click on **go**). Notice that the Aligned Reads Window scrolls to that position and has the cursor on the found string. (It might be complemented, but in this case it is not. Do the same in **Contig3**. Note that the end of Contig2 matches to the beginning of Contig3. This suggests that a join could be made between the ends of these contigs and that these ends are actually the same location, but did not meet phrap’s assembly requirements for some reason.



Navigate to the area to be joined: In the Aligned Reads Window for Contig2, click on **Compare Cont** (a button above the consensus in the Aligned Reads Window, on the same line as “Search for String”). Now double click on the **Contig3** line in the Searching Contigs Window used above. The Aligned Reads Window for Contig3 will move to the foreground and scroll to the specified location. In that Aligned Reads Window, click on **Compare Cont**.

Utilizing the “Compare Contigs Window”: Now the Compare Contigs Window should be visible. In the Compare Contigs Window, try scrolling back and forth. You can change the cursors (blinking red), but if you do, return them to the locations 181 and 189 bp for the next step. The cursors 'pin' these bases together when doing an alignment. (The algorithm is a pinned and banded Smith-Waterman alignment.)



Align: Click on **Align**. Try scrolling the alignment by dragging the scroll bar in the lower half of the Compare Contigs Window. Symbols between the two sequences indicate the quality of the match at each location. An 'X' means there is a discrepancy between the two contigs. There is also a 'P' (see if you can find it!) The P indicates the bases that you pinned together (the bases you highlighted in red). You will also notice that some bases are lighter and some are darker. This indicates quality, just as in the Aligned Reads Window. In this case, discrepancies only occur on low quality bases This is your cue that the discrepancy is just a base calling error rather than a genuine difference between the two contigs.

Make the join: Click on either contig in the bottom alignment. You will notice that both contigs will have the red blinking cursor in the same position. Click on **Scroll Both Aligned Reads Windows** and look at the Aligned Reads Windows to see that they scroll to the corresponding positions. You can pull up traces for the contigs and scroll by clicking (center clicking) on a non-consensus base in the Aligned Reads Window. Try experimenting with this. Then click **Join Contigs**. The two previous Aligned Reads Windows will disappear and will be replaced by a new window that has a new contig entitled: **Contig4**. You have made a join!

Examine the join made: Scroll left and right. You will notice that many of the reads are highlighted. These are the reads that came from the previous "right" contig.

Concluding and closing:

Examine Assembly View: Note that you now have an assembly comprised of one large contig, one of the goals of a finisher!

Save your file: From the Main Window, click on the **File** menu and select the first option: **Save Assembly**. Another window will open. Consed automatically suggests numbering the .ace file sequentially after the last saved file, but you can name the file whatever you wish. Enter the file name into the box below **Save assembly to file**, and click **OK**.



Quit Consed: Click on the **Quit Consed** button in the top right hand corner of the Main Window. If it asks you some questions, answer **Quit Without Saving and Discard .wrk File**. Consed keeps a log of all changes you make to an assembly: adding new reads, putting reads into their own contigs, making joins and tears, adding and removing tags, and changing bases. This log is kept in a file ending with .wrk. However, because you just saved a new .ace file, you can discard the .wrk file without losing any data.

****Export the consensus sequence from this exercise. First, run this sequence against the local instance of Swiss-Prot. Depending on your results, proceed with more comparison to determine something about this sequence.**
