

SEQUENCE ANALYSIS/ PHRED PHRAP CONSED BNFO520

Sequence reaction: dye terminator and dye primer
Gel image and ABI PRISM 377 chromatography
BAC based and Shotgun based sequencing

Phred/Phrap/Consed Sequence assembling and finishing program:

Phred

Phred is a base-calling program for DNA sequence traces. The program was developed by Drs. Phil Green and Brent Ewing. It is widely used by the largest academic and commercial sequencing laboratories.

In automated sequencing with fluorescent dyes attached either to the primer (dye primer chemistry) or to the dideoxy chain-terminating nucleotide (dye terminator chemistry). Typically a different dye is used for each of the four reactions. The four reactions are run in one lane in the dye primer method or one reaction is run in one lane in dye terminator reaction. At the bottom of the gel, a laser excites the fluorescent dyes in the fragments as they pass, and detectors collect the emission intensities at four different wavelengths. The laser and detectors scan the bottom of the gel continuously during electrophoresis in order to build a gel image in which each lane has a ladder-like pattern of bands of four different colors, each band corresponding to the fragments of a particular length.

Computer analysis is then used to convert the slab gel image to an inferred base sequence (or read) for each template. Typically this analysis consists of four distinct steps:

1. **Lane tracking** the gel lane boundaries are identified;
2. **Lane profiling** each of the four signals is summed across the lane width to create a profile, or trace, consisting of a set of four arrays indicating signal intensities at several thousand uniformly spaced time points during the gel run;
3. **Trace processing** signal processing methods are used determine the signal estimates, reduce noise, and correct for dye effects on fragment mobility and for long-range electrophoretic trends;
4. **Base-calling** the processed trace is translated into a sequence of bases.

The processed traces are usually displayed in the form of chromatograms consisting of four curves of different colors, each curve representing the signal for one of the four bases and drawn left to right in the direction of increasing time to detection (increasing fragment size). An idealized trace would consist of evenly spaced, non-overlapping peaks, each corresponding to the labeled fragments that terminate at a particular base in the sequenced strand. Real traces deviate from this ideal for a variety of reasons having to do with imperfections of the sequencing reactions, of gel electrophoresis, and of trace processing. Because of unusual migration of very short fragments (caused by relatively greater effects of the dye and specific base sequence on mobility) and unreacted dye-primer or dye-terminator molecules, the first 50 or so peaks of a trace are noisy and unevenly spaced. Toward the end of the trace, the peaks become progressively less evenly spaced as a result of less accurate trace processing, less well resolved as diffusion effects increase and the relative mass difference between successive fragments decreases, and more difficult to distinguish from noise as the number of labeled fragment molecules of a given size decreases. In particular, poorly resolved peaks for the same base may yield a single broad, often lumpy peak.

There are many problem to effect the bases, such as compressions, hairpin-like structure, GC-rich sequence, AT-rich sequence. Other frequently seen problems include weak or variable signal strength and noise peaks not corresponding to a base. The signal downstream of a run of mononucleotide or dinucleotide repeats frequently is degraded, possibly because of strand slippage during copying by the polymerase.

The goal of base-calling software is to produce a sequence as accurate as possible in the face of the above data problems. The phred base-caller uses a four-phase procedure to determine a sequence of base-calls from the processed trace.

Phase 1: Locating Predicted Peaks: Idealized peak locations (predicted peaks) are determined. The idea is to use the fact that fragments are locally relatively evenly spaced in most regions of the gel. Its height exceeds 10% of the height of the previous peak and is greater than the heights of the other three arrays at the same position.

Phase 2: Locating Observed Peaks: observed peaks are identified in the trace.

Phase 3: Matching Observed and Predicted Peaks: observed peaks are matched to the predicted peak locations, omitting some peaks and splitting others; as each observed peak comes from a specific array and is thus associated with 1 of the 4 bases, the ordered list of matched observed peaks determines a base sequence for the trace.

Phase 4: Finding Missed Peaks: the uncalled (i.e., unmatched) observed peaks are checked for any peak that appears to represent a base but could not be assigned to a predicted peak in the third phase, and, if found, the corresponding base is inserted into the read sequence. The entire procedure is rapid, taking less than half a second per trace on typical workstations.

Error probability has been established in Phred. Four parameters are particularly effective at discriminating errors from correct base-calls.

1. Peak spacing. The ratio of the largest peak-to-peak spacing, in a window of seven peaks centered on the current one, to the smallest peak-to-peak spacing. The minimum possible value of one corresponds to evenly spaced peaks.
2. Uncalled/called ratio. The ratio of the amplitude of the largest uncalled peak, in a window of seven peaks around the current one, to the smallest called peak; if there is no uncalled peak, the largest of the three uncalled trace array values at the location of the called base peak is used instead. If the called base is an N, phred assigns a large value of 100.0. Note that this is not what is sometimes called the signal to noise ratio, as uncalled peaks may be true peaks missed by the base-calling program rather than noise in the conventional sense. The minimum parameter value is 0 for traces with no uncalled peaks.
3. Same as 2, but using a window of three peaks.
4. Peak resolution. The number of bases between the current base and the nearest unresolved base, times -1 (to force the parameter to have the right direction). (A base is unresolved if it is called as N or if for at least one of its neighboring bases, there is no point between the two corresponding peaks at which the signal is less than the signal at each peak). The minimum possible parameter value is half the number of bases in the trace, times -1, and the maximum value is 0.

(Please see listed references for more detailed of error probability).

Phred scores:

Phred 10 means 1 error in ten to the 1.0 power

Phred 20 means 1 error in ten to the 2.0 power

Phred 30 means 1 error in ten to the 3.0 power

Phred 40 means 1 error in ten to the 4.0 power

Several reasons why Phred is used by leading sequencers are:

1. Better base calling accuracy. In a recent study, Phred achieved a 40-50% lower error rates than ABI software on large test data sets.
2. Error probabilities for each base call. The highly accurate error probabilities Phred calculates for each base enable increase automation of the sequencing process, for example: More accurate consensus sequences.
3. Automatic identification of areas that require "finishing" efforts.
4. Dramatically lower false positive error rates in mutation detection.
5. Effective quality control immediately after sequence production.
6. Quantitative benchmarking of different sequencing methods and protocol changes.
7. Identification of repeat sequences in during assembly.

Genome Sequence accuracy

The genome sequence accuracy is important. The high accuracy genome sequence can be used for diagnostics, forensics, predicting protein coding regions, finding repeat sequences, utilization of SNPs, etc.

Phrap

Phrap ("phragment assembly program", or "phil's revised assembly program"; a homonym of "frappe" = French for "swat") – a program for assembling shotgun DNA sequence data.

Phrap is a leading program for DNA sequence assembly. Phrap is routinely used in some of the largest sequencing projects in the Human Genome Sequencing Project and in the biotech industry. Some of Phrap's features include:

1. **Allows use of entire read** (not just trimmed high quality part); Phrap can construct contig sequence as a mosaic of the highest quality parts of reads (rather than a consensus) that utilize all sequence information.
2. **Fast assemblies.** Assemblies of cosmid- to BAC sized projects with several hundred to two thousand reads typically take only minutes to complete on high-powered workstations or personal computers.
3. **Accurate consensus sequences.** Phrap uses Phred's quality scores to determine highly accurate consensus sequences. Phrap examines all individual sequences at a given position, and generally uses the highest quality sequence to build the consensus - similar to the way scientists would correct consensus sequences during "contig editing". Compared to simple majority rules use in older sequence assembly programs, Phrap's approach can give significantly more accurate consensus sequences, especially in regions of low coverage or regions of systematic errors like compressions.
4. **Consensus quality estimates.** Phrap uses the quality information of individual sequences to estimate the quality of the consensus sequence. In addition, Phrap uses available information about sequencing chemistry (dye terminator or dye primer) and confirmation by "other strand" reads in estimating the consensus quality. This often allows scientists to ignore random errors, and to focus finishing efforts exclusively onto regions where the data quality is insufficient. Consensus quality estimates can also be very helpful in mutation detection by DNA sequencing (see Rieder, Taylor, Tobe & Nickerson (1998), *Nucleic Acids Research* 26: 967-973).
5. **Ability to assemble very large projects.** Phrap has been used routinely to assembly bacterial genomes sequenced by the "shotgun" approach, where each project contained tens of thousands of reads. Smaller bacterial genomes (2 million bases or less) could often be assembled in less than three hours.
6. **Improved identification and handling of repeats.** Phrap uses quality scores to estimate whether discrepancies between two overlapping sequences are more likely to arise from random errors, or from different copies of a repeated sequence. For repeats with 95 to 98% identity (like human Alu sequences) and high quality sequence data, this typically yields correct assemblies.

Cross_match

Fast DNA Sequence Comparisons and Vector Screening

Cross_match is a program for fast comparisons of DNA sequences that uses the same algorithms as Phrap. For example, the comparison of several hundred thousand bases of "raw" sequence to the sequence of an entire BAC typically takes less than one minute. Within the Phred - Phrap system,

1. Cross_match is typically used for vector screening.
2. Identification of overlaps between contig ends after assembly with Phrap or other assembly programs.
3. Identification of potential repeat sequences in assemblies.
4. Generation of error summaries and lists after completion of sequencing projects.
5. Estimation of vector contamination in newly created libraries.

Consed

Consed is a graphical tool for sequence finishing. It is a program for editing sequence assemblies created with Phrap assembly program. It was written specifically to go with Phrap -- it takes advantage of quality values assigned by Phred and Phrap and the consensus sequence created by Phrap. In addition to a full set of standard features (view traces, edit reads by inserting a base, deleting a base, substituting a base, etc.), it supports an efficient editing procedure designed for use by Phrap in subsequent reassemblies of the same data set.

QUICK TOUR OF CONSED

(Modified for class, from CONSED 13.0 DOCUMENTATION)

1) Log on your Watson account, open the X-window.

2) Run Phred and Phrap

Create three fold for your sequence analysis:

```
> mkdir chromat_dir
```

```
> mkdir edit_dir
```

```
> mkdir phd_dir
```

3) Copy the ABI sequencer format sequences from mi653SequenceFold into chromat_dir by

```
> cp /home/mic653/SequenceAnalysis/standard/chromat_dir/* chromat_dir/
```

```
> cd edit_dir
```

Run Phred and Phrap by typing

```
> phredPhrap
```

The perl script phredPhrap will run phred for basecalling and run Phrap for sequence assembly. You will find file with name "YourAccountName.fasta.screen.ace.1" after the program.

4) Start Consed by type

```
> consed -ace YourAccountName.fasta.screen.ace.1
```

or

```
> consed
```

(To be consistent in the class, copy the same previously assembled fold in you fold.

```
>cp -r /home/mic653/SequenceAnalysis/standard .
```

```
>cd edit_dir
```

```
>consed)
```

Two windows will appear, Consed Main Window and Ace Files. The Ace Files window will have the list of .ace files and say 'select assembly file to open' and 'standard.fasta.screen.ace.6'. Double click on that name. Click "Yes". The Ace Files window goes away.

You will now see a list of one contig and a list of reads. This is the 'Main Consed Window'. Double click on 'Contig1'.

The 'Aligned Reads Window' will appear. Try scrolling back and forth. Try scrolling by dragging the thumb of the scrollbar. Also try scrolling by clicking on the 4 << < > >> buttons for scrolling by small amounts. For scrolling by tiny amounts, click on the arrows at either end of the scrollbar. For scrolling by huge amounts, use the middle mouse button and just click on some location on the scrollbar. For scrolling to the beginning or end of the contig, use the <<< or >>> buttons.

(Question: why can't you just move the scrollbar to the extreme right in order to go to the beginning of the contig? Answer: in typical assemblies, there are reads that protrude beyond the beginning of the contig and reads that protrude beyond the end of the contig. Moving the scrollbar to the extreme right will scroll the contig to the end of the rightmost read--typically far to the right of the end beginning of the contig. Thus you should get in the habit of using the <<< and >>> buttons.)

Notice the colors. Scroll to position 937 and notice the 'a' in sequence djs74-423.s1. The red bases are the ones that disagree with the consensus.

Notice the different shades of grey background (around the bases). They have the following meanings, but first, you need to understand the meaning of the quality values:

A quality value of 10 means 1 error in ten to the 1.0 power

A quality value of 20 means 1 error in ten to the 2.0 power

A quality value of 30 means 1 error in ten to the 3.0 power

A quality value of 40 means 1 error in ten to the 4.0 power

and for quality values in between:

A quality value of 25 means 1 error in ten to the 2.5 power

(These have actually been empirically verified--if you are interested in the gory details, read the phred papers below)

Also notice the upper and lowercase. This is just a cruder indication of the quality of the bases.

5) To see the quality value of a particular base, point at it and click with the left mouse button. You will see the quality displayed in the Info Box on the Aligned Reads Window.

These quality values are shown in grey scales:

Quality 0 through 4 is given by dark grey

Quality 5 through 9 is given by a shade lighter

Quality 10 through 14 is given by a shade still lighter

Quality of 40 through 97 is given by white (the brightest shade)

A quality value of 99 is reserved for bases that have been edited and the user is absolutely sure of the base ('high quality edited').

A quality value of 98 is reserved for bases that have been edited and the user is not sure of the base ('low quality edit').

The ends of the reads show bases that are grey and have a black background. These are the low quality ends of the reads or the unaligned ends of reads, as determined by Phrap.

6) Click on a base on a read. Then hold down the control key and type 'a'. You will move to the beginning of the read. Hold down the control key and type 'e'. You will move to the end of the read. (Emacs users will recognize these commands.)

7) Scroll so that location 490 is about in the middle of the aligned reads window. Push the left mouse button down on the menu item 'Dim'. There will be a list of choices that will appear. Drag the cursor down to 'Dim Nothing' and release. Now look what happened to the color of the bases. The ends of the reads that used to be with a black background now appear red with a grey background. You are seeing the clipped-off bases with all the same information as any other base. Since there is a huge amount of red (discrepant) bases, the screen becomes distracting and busy. Thus by default the low quality clipped-off bases are made with a black background and a grey foreground so they don't distract you.

Notice there is a distinction here between 'low quality ends of reads' and 'unaligned ends of reads'. Unaligned ends of reads can be low quality as well, or they can be high quality, as in the case of chimeric reads.

Point with the mouse to a read name and hold down the right mouse button. You will notice there is a line that says "high quality from nnn to nnn; aligned from nnn to nnn; chem: prim". This is giving the same information in number form. Check that the numbers agree with the dimming.

You can play with the dimming options a bit. Then return it to 'Dim Low Quality' for the rest of this tour.

Traces and editing

8) Point with the mouse at a base of one of the reads and click with the middle mouse button. The Trace Window showing the traces for that stretch of read should popup.

There are 2 rows of numbers:

'con' are the consensus positions

'rd' are the read positions

There are 3 rows of bases in the trace window:
'con' is the consensus
'edt' is where you can edit the base calls of the read
'phd' is the original phred base calls

Notice that a red rectangle blinks (the 'cursor') in the corresponding positions of the Aligned Reads Window and the Trace Window.

9) Try editing in the Trace Window. You can click the left mouse button on a base in the 'edt' line to set the cursor (a blinking red rectangle). You can directly overstrike a base by typing a letter. Try this. Try undoing it (by clicking on 'undo'). If you want to undo more than one edit, you will have to go back to the main Consed window and click on the button labeled 'Undo Edit...'--you will learn that later.

You can move left and right with the arrow keys.

We believe that the user should change a base call only while examining the traces. That is why editing is done here--not in the Aligned Reads Window.

10) You can insert a column of pads by pushing the space bar. Try this. (You may need to click on a base on the 'edt' line first.)

(For those of you new to editing assemblies, a 'pad', which in Consed and Phrap is represented by the '*' character, is used to align two or more sequences such as these:

```
gttgacagtaatcta
gttgacataatcta
```

in which one sequence has an inserted or deleted base with respect to the other. By inserting the pad character, it is possible to get a good alignment:

```
gttgacagtaatcta
gttgaca*taatcta
```

This is the purpose of pad character--it is just a placeholder.)

You can then overstrike a pad with a base. In this way you can insert a base, and still preserve the alignment.

11) Try highlighting a stretch of a read on the edt line by holding down the middle mouse button and dragging the cursor over some bases. They will turn yellow as you drag. Then release the mouse button. A window will pop up giving you some choices of what to do with those (yellow) bases.:

Make High Quality--makes the highlighted bases edited high quality (99). This tells Phrap (when it reassembles) that you are sure of the sequence here.

Change Consensus--make the highlighted bases edited high quality and change the consensus to agree with that stretch of the read. This is a directive to Phrap (upon reassembly) to use that stretch of that read to be the consensus.

Make low quality--makes the highlighted bases edited low quality. This tells Phrap (when it reassembles) that you are not sure of the bases here and Phrap can go ahead and make a join even if the bases in this region don't match perfectly.

Make Low Quality to Left End--same as above, but all the way to the left end of the read.

Make Low Quality to Right End--same as above, but all the way to the right end of the read.

Change to n's--Change the highlighted bases to n's which means they are unknown bases. This tells Phrap (when it reassembles) to not make any join based on these bases. It is useful when you believe the bases may be in the chimeric portion of a read.

Change to n's to left--same as above but to left end.

Change to n's to right--same as above but to right end.

Add Tag--allows user to add any tag to a stretch of read bases.

Dismiss--you decided you don't really want to do anything with this stretch of bases.

This popup is made so that nothing else works until you choose something. Try each of these choices, except for tags, which you'll try below.

'Change Consensus' has an additional function--if a read extends out on the right beyond the end of the consensus, you can extend the consensus by using this function. You might want to do this, for example, if crossmatch did not correctly find the cloning site and thus clipped too much. You can add these bases to the consensus by using 'Change Consensus'. Typically, the quality of these bases in the read and in the consensus is 99. That is so that next time Phrap runs, it will correctly extend the consensus.

However, if you aren't going to reassemble, you might want to just leave the quality values the way phred originally called them. You can do this by using a Consed resource (consed.extendConsensusWithHighQuality).

12) To delete a base, overstrike it with a '*' character. (Phrap ignores '*', so this is the same as deleting the character.) If you overstrike all bases in a column with * characters so the entire column consists of *'s (including the consensus base), there is no way to remove the column. This is OK since when you export the consensus (try the exercise on EXPORTING THE CONSENSUS), the *'s are not exported. While you are editing in Consed, we believe there should be a visual indication that a base was deleted.

Saving the assembly

13) To save the assembly, pull down the 'File' menu on the Aligned Reads Window, and release on 'Save assembly'. A box will pop up with a suggested name. I suggest you always use the one it suggests. The idea is that the ace files:

```
(project).fasta.screen.ace.1  
(project).fasta.screen.ace.2  
(project).fasta.screen.ace.3  
(project).fasta.screen.ace.4  
(project).fasta.screen.ace.5
```

are in order of how old they are. If you feel you are taking up too much disk space, then start deleting the ace files starting at the oldest. I do not recommend that you overwrite existing ace files. The version numbers just keep growing, and that is not a problem.

Exporting the consensus

14) Exporting the consensus. Bring the Aligned Reads Window into view again. Hold down the left mouse button on the 'File' menu and release the button on 'Export consensus sequence'. Notice that the consensus will be stored (in this case) in a file called 'Contig1.fasta'. Click 'OK'. There is now a file in your edit_dir directory called 'Contig1.fasta' that has the consensus sequence in it. If you want to see the file, bring up another X-term (if you are UNIX literate), and type:

```
cd standard/edit_dir  
more Contig1.fasta
```

15) Fancier exporting the consensus. Bring the Aligned Reads Window into view again. Hold down the left mouse button on the 'File' menu but this time release on 'Export consensus sequence (with options)...'. Just export a little snip of the consensus, from 400 to 410. (You will notice this contains a pad * character.) Ask for both the bases file and the quality file. Click 'OK'. Consed will want to call this file 'Contig1.fasta' again. You can overwrite the existing file.

Look in your other X-term at these files:

```
more Contig1.fasta  
more Contig1.fasta.qual
```

The one file contains the bases (but no * pads) and the other contains the corresponding qualities of those bases.

16) Exporting the consensus of all contigs at once: Go to the Main Consed Window. Point to 'File', hold down the left mouse button, and release on 'Write all contigs to fasta file'. You then can choose a filename for all contigs to be written to. (In this project there is only 1 contig, so there is no difference between this option and just exporting a contig at a time.)

17) (For this step, first click on the 'Dim' menu and release on 'Dim Nothing'.) Point to the 'Color' menu, hold down the left mouse button and release on 'Color Means Edited and Tags'. Notice that the bases that you have edited (make sure you have edited some bases) will stand out in either white or grey (depending on whether the base was made high quality or low quality). Observe this both in the Trace Window and the Aligned Reads window. This color mode is useful if you are interested in easily spotting which bases are edited.

Return to the 'Color Means Quality and Tags' color mode by the following: point to the 'Color' menu, hold down the left mouse button and release on 'Color Means Quality and Tags'.

Multiple undo edit

18) Now that the Consed Main Window is visible, click the 'Undo Edit...' button. There will be a popup indicating the most recent edit. (If it says "no edits so far", then bring up a trace and make several edits. Then click on 'Undo Edit...' again.) Click 'undo'. Then you will see the edit that was done before that. Click 'undo'. You can continue undoing if you like. You now know how to undo more than one edit. You cannot choose which edits to undo and which to not undo--edits can only be undone in precisely reverse order from the order you made them. Once you save the assembly, you cannot undo prior edits.

Scrolling traces and aligned reads together

19) In the Aligned Reads window, scroll along the contig to a different point. Click the left mouse button on a read whose trace is already up. Notice that the existing trace instantly scrolls to the corresponding location. Now go to the Trace Window and scroll the traces to a new location. Click on the edit line with the left mouse button. You will notice that the Aligned Reads window will instantly scroll to the corresponding location. Thus you can keep the Aligned Reads window and the traces scrolled to the same location.

Examining all traces

20) Go to a region where there are lots of reads, say base 1660. Push down the right mouse button and release on 'Display traces for all reads'. You will see all traces displayed in a scrolling window. You can drag the scrollbar on the right down and up to see all the traces. This feature is particularly useful for polymorphism/mutation detection work. This feature was added to work in cooperation with polyphred.

21) Alphabetical ordering of reads

The reads can be ordered in two ways:

a) alphabetically

b) first all the top strand reads and then all the bottom

strand reads. The top strand reads are then ordered

by the left end of the reads. Same with the bottom

strand reads.

Try changing between a) and b). In the Main Consed Window (click on 'Main Consed Window' on the Aligned Reads Window if you can't find the Main Consed Window because it is covered up with other windows), pull down the 'Options' menu, and release on 'General Preferences'. Scroll down until you find 'Display reads sorted alphabetically or by strand/left end of read.' Switch it between 'alpha' and 'strand'. Then click 'Apply and Dismiss'. Notice the effect in the Aligned Reads Window. Many polymorphism and mutation detection labs find that alphabetically sorting is most useful, while many genomic sequencing labs find that sorting by strand/left end of read is most useful.

After you are done playing with these features, exit Consed and go back to the previous database:

```
cd ../../standard/edit_dir
```

ls ../../consed
Double click on standard.fasta.screen.ace.1

When it says "There is an edit history file (a .wrk file)...Do you want to apply those edits?", click on "no".

Double click on Contig1 to bring up the Aligned Reads Window again in preparation for the next step.

Navigating

22) In the Aligned Reads window, pull down the Navigate menu and release on 'Low consensus quality'. 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. (Low consensus quality means an area in which the bases each have too high probability of being wrong.) This saves you from having to look through large amounts of high quality data trying to find problem areas.

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.

You may want to click on the 'Save' button in the Low consensus quality Window to save to a file a copy of this list of problem areas as you work through them.

In our experience, this will be the most important navigate list you will use. In fact, finishing consists mainly of adding reads and rephrapping until this list is reduced to nothing.

23) Dismiss the Low consensus quality window. Pull down the 'Navigate' menu again and release on 'High quality discrepancies as above, but omitting tagged compressions and G_dropouts'. You will probably notice there are no entries (unless you created some yourself by editing). That is because there are no high quality discrepancies with this dataset. So let's force there to be some by lowering the quality threshold. First, dismiss the High quality discrepancies window.

Click on 'Main Consed Window'. In the Main Consed Window, pulldown the 'Options' menu and release on 'General Preferences'. Notice that the default for 'Threshold for High Quality Discrepancy' is 40. Change it to 15 and click 'Apply & Dismiss'.

Then follow the steps above to bring up the High quality discrepancies menu. Now you will see several entries. Click 'next' repeatedly to go successively to the next high quality discrepancy in the Aligned Reads Window.

You can also double click on a particular line in the High quality discrepancies window to go to that location. Alternatively, you can single click on a line and then click the 'Go' button.

Dismiss the High quality discrepancies window.

24) Similarly, try the other navigate lists: Unaligned high quality regions (this list will be empty with this data set), Edits, Regions covered by only 1 strand and only 1 chemistry, and Regions covered by only 1 subclone.

Unaligned high quality regions are regions in which the traces are high quality so there is no question of the bases, but the region differs so much from other reads that Phrap has given up trying to align the region with the consensus. This could be due to a chimeric read, or perhaps the read belongs somewhere else.

We believe that regions covered by only 1 subclone should be covered by a 2nd subclone to prevent the possibility of there being a deletion in the single subclone.

There are so many different problem lists that you may forget to check one of them and thus miss a serious problem. Thus we combined them all into a single list. This is the first menu item: 'Low Cons/High Qual Discrep/Single Stranded/Single Subclone/Unaligned High'. We suggest you use this list.

25) Also try navigate by tags by selecting 'tags' under navigate: when the Select Tag Type Window appears, double click on 'compression'. (Note that you can't do anything else until you deal with this window.) This gives a list of a particular tag type in a particular contig.

26) There is also a way of getting a list of a particular tag type in all contigs: Click on 'Main Consed Window'. In the Main Consed Window, point to the 'Navigate' menu, hold down the left mouse button, and release on 'Tags in all contigs'. Continue as in the previous step. (Since there is only one contig, this list will not be any different than the corresponding list for Contig1.)

Primer-picking

To do this step, you must have first completed the INSTALLING CONSED (below). So, if you haven't done that yet, please complete that first.

27) Go to some location near the right end of the contig, say base 2470. Click with the right mouse button on the consensus and click on either one of the top strand primer choices (either from subclone template or from clone template). Consed will pause a moment, and then there will appear a selection of primers that pass all of Consed's requirements. Templates are also chosen for each primer. You may have to scroll the primer list to the right to see the templates. Consed lists these templates in order of quality--all of them will cover the read you want to make.

Double click on one of the primers in the Primers Window. That will cause the Aligned Reads Window to scroll to show that oligo in context. Click on 'Accept Primer'. A comment box will pop up. Enter some comment and click 'OK'. Notice that a yellow oligo tag, with a little red end, is created on the consensus for that primer. The red end points in the direction of the oligo. The tag contains all the information you need to order that oligo and do the reaction--you will learn how to pop it up below under 'tags'.

What is the difference between 'Pick Primer from Subclone Template' and 'Pick Primer from Clone Template'?

There are 3 differences:

1. which vector file the primers are screened against? In the former case, the primer is screened against the file primerSubcloneScreen.seq and in the latter case against the file primerCloneScreen.seq
2. In checking for false matches elsewhere in the assembly, if the template is the whole clone, then Consed must check for false matches in the *entire* assembly, including all other contigs. But if the template is just going to be a subclone, Consed only needs to check elsewhere in that subclone. Actually, to be conservative, Consed checks for false matches +/- the maximum insert size of a subclone.
3. If you are picking primers for subclone template, then the primer picker can also pick the subclone templates. If it doesn't find any suitable subclone template, it will reject the primer. (By default, picking of subclone templates is turned on. If you prefer to pick your own primers, and want Consed's primer picker to be much faster, you can turn it off temporarily or permanently. To turn it off temporarily, go to the Consed Main Window, point to the Options menu, hold down the left mouse button and release on 'Primer Picking Preferences'. Scroll down to 'Pick Subclone Templates for Primers' and click 'False'. Click on 'Apply and Dismiss'. To change this permanently, see CONSED CUSTOMIZATION below. Beware: you must correctly customize determineReadTypes.perl for template picking to work. See INSTALLING CONSED below.)

When you are done editing and have saved the assembly and exited Consed, run ace2Oligos.perl (supplied with this distribution—make sure your system administration installed it) which will extract all the oligos you just created. This is handy for email ordering of oligos.

In the X-term, type:

```
ace2Oligos.perl standard.fasta.screen.ace.2 oligos.txt
```

where standard.fasta.screen.ace.2 is whatever the name is of the ace file you just saved.

28) Picking PCR primer pairs

In the Aligned Reads Window, go to the location where you want to pick the first PCR primer, say base 500. Point to the consensus, hold down the right mouse button and release on "Top Strand PCR Primer". Then scroll to the location where you want to pick the second PCR primer, say base 2200. Point to the consensus, hold down the right mouse button and release on "Bottom Strand PCR Primer". There will be a pause and then there will be a list of PCR primer pairs. Click on the pair you want and click "Accept Pair".

You can modify the parameters for choosing PCR primer pairs by going to the Main Consed Window, pointing to "Options", holding down the left mouse button, and releasing on "Primer Picking Preferences." For example, by default Consed does not display all PCR primer pairs—this would take too long and give you too many. However, you can ask it to show you all such pairs. In the Primer Picking Preferences, scroll down to "Check All PCR Pairs (huge) or Just Sample?" and click on "All". Then click on "Apply and Dismiss". Then pick PCR primers again, as above. Don't be surprised if you get 10,000 or more pairs of primers!

Search for string

29) Try the 'Search for String' button (left side of the Aligned Reads Window). Type in a string (such as aaaca), and click 'ok'. There should be a list of 'hits'. Double click on one of the hits (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.)

Dismiss this window. Try this again, only this time in the Search For String Window select 'Search Just Reads'. Then click 'OK'. You will notice there are many more hits. This is because this shows hits in each read, even if they are at the same consensus position.

You can also try the approximate match search for string by clicking on 'Approximate' instead of 'Exact'. The 'Per Cent Mismatch' only applies to the Approximate match search.

Copy and paste

30) In the Aligned Reads Window, swipe some bases by holding down the left mouse button. You should see the bases turn yellow, at least temporarily. Then click the 'Search for String' button. Use the middle mouse button to paste the bases you have just swiped into the 'Query string:' box. Notice that you can swipe bases either from the consensus or from a read.

The search for string is case-insensitive so don't worry about the pasting being upper or lowercase.

Correcting false joins made by Phrap

31) Phrap may put several reads together that you believe do not belong together. (For example, you may see several high quality discrepancies between the reads.) If you are sure these reads do not belong together, you can force a subsequent reassembly by Phrap to not assemble those reads together. You do this by finding a location where there is a high quality discrepancy. Then click on the read with the right mouse button and release on "Tell Phrap not to overlap reads discrepant at this location". There are no high quality discrepancies with this dataset so Consed won't let you do this.(Try it and see.) However, when you use your own data, you may get the chance!

Tears and joins

Just so you get the same results as I do, exit Consed and bring it up again using the original ace file standard.fasta.screen.ace.1

If it asks if you want to apply edits, just say 'no'.

32) When Phrap really screws up, you may want to just tear the contig apart in several places and then join the pieces back together in a different way. Let's try it:

Go to location 1500. Point the mouse at the consensus base at 1500 and push the right mouse button down. Release the button on 'Tear Contig at This Consensus Position'. Up will pop a list of reads with 2 little buttons next to them <- and ->. Leave everything as it is and just click 'Do Tear'. (If you want to play around with which reads goes into which contig, do that another time.)

Now you should have 2 Aligned Reads Windows on top of each other. One should contain 'Contig2' and the other 'Contig3'. Dismiss the little window that says 'Tear Complete'.

Now let's join these 2 contigs back together:

Click on 'Search for String' and type in the following bases: agctgccatc and 10> mismatch

Click 'OK'.

Search for string should find 2 locations, one in Contig2 and one in Contig3:

Contig2 (consensus) 1447-1456 (??) (uncomplemented)
Contig3 (consensus) 829-838 (??) (uncomplemented)

Double click on the first one. The Aligned Reads Window for Contig2 will scroll to location 1447 (??) and the window will rise up. In that Aligned Reads Window, click on 'Compare Cont'.

Now double click on the 'Contig3' line in the above Search for String results. The Aligned Reads Window for Contig3 will scroll to location 829 (??) and lift up. In that Aligned Reads Window, click on 'Compare Cont'.

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, please return them to the locations 1447 (??) and 829 (??) for the next step. The cursors 'pin' these bases together when doing an alignment. (The algorithm is a pinned Smith-Waterman alignment.)

Click on Align. Try scrolling the alignment by dragging the thumb in the lower half of the Compare Contigs. An 'X' means there is a discrepancy between the 2 contigs. There is also a 'P' (see if you can find it!) The P indicates the bases that you pinned together.

Click with the left mouse button 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 have traces up for the contigs, and they will scroll as well. Experiment with this. Then click 'Join Contigs'. The 2 previous Aligned Reads Windows will disappear and there will be a new one which has a new contig 'Contig4'. You have made a join!

It is possible to have more than one Compare Contigs windows up at a time. This allows you to investigate a repeat that has more than 2 copies.

Compare Contigs is one method of exploring joins of contigs that were not made by Phrap. Another method is to use phrapview, supplied with Phrap. phrapview gives a high level view of all internal joins while 'compare contigs' shows the alignment of a single internal join. Some users have found them to work well together--phrapview to find a join and, having found it, 'compare contigs' to examine it in more detail.

Removing reads

33) You can also remove individual reads and put them into their own contigs. For example, in the Aligned Reads Window, go to location 2000. Point to the read name of read djs74_2664.s1 and hold down the right mouse button. Release on 'Put read djs74_2664.s1 into its own contig.' Consed will ask you 'Are you sure...?' Answer 'yes'. Presto-chango! The read is put into its own contig and the old contig is redrawn without the read in it. At this point you should save the assembly--you should always save the assembly after removing reads.

Tags

34) Bring up a trace for a read (as above). Swipe some bases on the 'edt' line with the middle mouse button. A list of choices will pop up. Select 'Add Tag'. Type in a comment in the box at the bottom, and select 'comment' from the list of tag types. You will now see a blue box both in the Aligned Reads Window and in the Traces Window on that read.

To see the comment, you can just point to it in the Aligned Reads Window and you will see the comment in the lower right hand corner of the Aligned Reads Window. Alternatively, you can click on that blue tag in the Aligned Reads Window with the right mouse button and release on 'Tag: comment Show more info?'. Alternatively, you can click on the blue tag in the Traces Window with the right mouse button.

Try creating some other kinds of tags: again swipe some bases in the Trace Window by selecting a different tag type. You will notice that different tags are in different colors. You can always use the methods above to see what kind of tag it is if you forget what a particular color means.

35) You can create really, really long tags as follows: Just create a short version of the tag as above for where you want the tag to start. Then figure out the consensus position of where you want the tag to end. In the Aligned Reads Window, click on the short tag with the right mouse button and release on 'tag: show more info?' (as above). A Tag Window will appear for that tag. In the Tag Window, simply change the End Unpadded Consensus Position to the place you want it to end. Then click 'OK'. You will now notice that the tag will be as long as you wanted.

36) You can create tags on the consensus in the same way. In the Aligned Reads Window, use the middle mouse button to swipe some bases on the consensus in the Aligned Reads Window. Up will pop a list of tag types. Click on one of them. Try it again somewhere else. Try it with the tag type being 'comment'. In this case, you must enter a comment. Notice the pretty colors! If you forget what a particular color means, you can click on the colored tag with the right mouse button and it will tell you.

37) Try creating some tags that overlap each other. You will notice that the overlapping region will be purple. If you want to know which tags overlap, you can click with the right mouse button on the purple and you will be told all tags that are on that base.

38) If you have many tags that overlap and thus are purple, you can hide some less relevant tag types so there is less purple and there is less distraction. Make sure you have a few tags visible. Then click on 'Main Consed Window'. In the Main Window, open the Options menu, and release on 'Hide Some Tag Types'. A list of tag types will pop up. Select the type that you have visible (above). Then click 'OK'. Go back to the Aligned Reads Window. That tag should still be visible. Click on the button 'Some Tags' in the upper right part of the Aligned Reads Window. Your tag should disappear. The 'Some Tags' button should have changed to 'Show All Tags'. Click on it again. Your tags should have reappeared.

39) Normally, when you re-assemble, Phrap will name the contigs differently--what was Contig31 before may become Contig32. To help you know which contig is which, Consed allows you to give a name (e.g., "A") to a contig which will persist after re-assembling. To do this, swipe some consensus bases with the middle mouse button (as above). When the "Select Tag Type" box pops up, click on "contigName" and also type a name into the "Contig Name:" field and then click "OK". The next time you re-assemble, the name "A" will appear in the list of contigs on the Main Consed Window.

Search for read name

40) Restart Consed using the original ace file
standard.fasta.screen.ace.1

If it asks if you want to apply edits, just say 'no'.

Instead of clicking on a read or contig name, type a read name into the 'Find read:' box. Try typing djs74-2. You will notice that as you type each letter, the first item in the list that matches the letters typed will be highlighted. Experiment with deleting a few letters and typing others. This is a powerful method of quickly getting to the read name you are interested in. When you get to the name in the list, you do not have to type the rest of the name--just type carriage return or else click on 'OK'.

Even more powerful is the "Find read (with *'s)":. In this case you can just type "2689" and then push the "Enter" key and Consed will immediately bring up the Aligned Reads Window with the cursor on read djs74-2689.s1. Suppose that there were more than one read that matched? For example, suppose you type: "26" and then push the "Enter" key. This matches 3 reads:

```
djs74-2689.s1  
djs74-2679.s1  
djs74-2664.s1
```

Try it and see what happens...

Try entering "26*9" and see what happens. What does the "*" mean?

Online documentation

41) On the Aligned Reads Window, click on the 'Help' menu and release on 'Show Documentation'. You will see this document. You can search for keywords in it.

Goto position

42) In the Aligned Reads Window, click in the 'Pos:' box in the upper right-hand corner. Type in a number, such as 540, and push the 'Return' or 'Enter' key. The Aligned Reads Window will scroll to position 540. We find this feature is particularly useful when one person wants another person to look at something in the sequence.

Highlighting read names

43) In the Aligned Reads Window, click on a read name with the left mouse button. The name will turn magenta. Click again and it will turn yellow again. Try turning it magenta and then scrolling. This feature is helpful in keeping track of a particular read as you scroll.

If you have an emacs window open (or any editor window), you can paste the read name in by just clicking with the middle mouse button. When you clicked on the read name in the Aligned Reads Window with the left mouse button, the read name was loaded into the paste buffer.

Complementing the contig

44) Push 'Comp Contig' in the Aligned Reads Window to complement the contig. This displays the opposite strand of the contig including the consensus and all reads. Push this button again to uncomplement it.

Recovery from crashes

45) It is important to feel that your data are safe, even if the computer (or Consed) were to crash. Consed will recover your data from such a crash.

Make an edit (remember, edits are made in the Trace Window) and jot down its location. Also note the name of the ace file which is displayed in the upper left box in the Aligned Reads Window. Then simulate a crash by going to the X-term where you started Consed and typing control-C. Restart Consed and select the same ace file you noted (above). A box will come up saying "There is an edit history (a.wrk file)

Consed may have crashed during a previous session with this same file. Do you want to apply those edits?' Click on 'yes'. Go and find the edits you made before Consed crashed--you will find them.

This is the purpose of the .wrk files--they are a log file of your edits and they are added to as you make edits.

Protein translation and open reading frames

46) If you would like, you can see the amino acid translation of the consensus in all reading frames. In the Aligned Reads Window, push down the left mouse button on the 'Misc' menu and release on 'Show Top Strand Protein Translation'. Try again but this time release on 'Show Bottom Strand Protein Translation'. Notice that there are 2 characters that are in magenta color. What are those characters? Why are they made in a different color? To not show the protein translation, push down the left mouse button on the 'Misc' menu and release on 'Don't show protein translation'.

47) You can search for open reading frames within a contig. In the Aligned Reads Window, push the left mouse button on 'Navigate' and release on 'Search for Open Reading Frames'. Notice that the open reading frames are shown for all 6 reading frames and are sorted by length.

Error rate

48) In the Aligned Reads Window is a box (upper right) labeled 'Err/10kb'. This is the estimated error rate for this contig, and it is a good indicator of when you are done (or not done) finishing. In addition, you can find the error rate for a particular region of contig as follows: Point at 'Misc' menu, hold down the left mouse button, pull down and release on 'Show Error Info For Region'. Fill in the boxes for left and right consensus position, click on 'Calculate' and you will be given the error and single subclone data for that region.

Maximum number of traces displayed

49) Bring up dataset standard. Scroll to position 1162. Bring up 4 reads and then try bringing up additional reads. You will notice that new reads are put at the top of the stack of traces and, once there are 4 traces displayed, traces are automatically removed from the bottom of the stack. If you want to change this maximum number of traces to something besides 4, you can do that: In the Main Consed Window (click on 'Find Main Win' on the Aligned Reads window), pull down the 'Options' menu, and release on 'General Preferences'. Try changing the 'Max Number of Traces Shown' to 3. Then click 'Apply and Dismiss'. Now dismiss the Trace Window and again start adding additional traces to the Trace Window. You will notice that now the number of traces shown will not exceed 3.

Hotkeys for editing

50) If you do a lot of editing, you will want to have a faster method of doing these edits than having the popup and selecting an option. Thus the following hot keys exist:

- < and > (less than and greater than) to make n's to the left and the right (respectively) of the cursor
- control-l and control-r to make low quality to the left and the right (respectively) of the cursor
- over striking with a capital letter (e.g., C instead of c) causes the base to become high quality rather than low quality
- over striking with a lower case letter causes the base to become low quality

51) Now go to the menu labeled 'color', and pulldown and release on 'color means match'.

Now you notice different colors: The colors have the following meaning:

- Blue: agrees with consensus
- Orange: disagrees with consensus
- Yellow: this stretch of this read was used to form the consensus
- Grey: Low quality or unaligned ends of reads

Now go back to the color mode 'color means quality and tags' (the default) for the next exercise.

52) Long, long, long read names

If you have very long read names, you might not be able to see the whole name in the Aligned Reads Window. To fix this, go to the Main Consed Window, pulldown the 'Options' menu and release on

'General Preferences'. Scroll down until you see "Max Chars for Read Names in Aligned Reads Window". Increase the number and click on "Apply". When you are satisfied with how the read names look in the Aligned Reads Window, click on "Cancel" in the General Preferences Window.

GLOSSARY:

Alignment: refers to the procedure of comparing two or more sequences by looking for a sequence of individual characters or character patterns that are in the same order in the sequence. There are two types of alignment, local and global alignments.

Annotation: The prediction of genes in a genome, including predictions of:

1. the location of protein encoding genes
2. the sequences of the encoded proteins, introns, exons and splicing junctions
3. and significant to other proteins of known functions,
4. the location of RNA-encoding genes.

Assembly: The process of alignment and building a consensus (contig from overlapping short sequence reads determined by DNA sequencing)

Contig: A set of clones (sequences) that can be assembled into a linear order.

Consensus: A single sequence that represents, at each subsequent position, the variation found within corresponding columns of a multiple sequence alignment.

Finishing: The process of preparing a sequence for submission to GenBank, the public database where all data is submitted for the public to view. There are five primary finishing goals:

1. Filling gaps. Join contigs into a single, large contig that represents the sequence of the original clone (BAC, cosmid, etc.)
2. Address regions of low sequence quality that may contain sequence errors.
3. Add coverage to single-clone regions (areas spanned by only one M13 or plasmid subclone). This is important because sometimes mutations can occur in a subclone; therefore we want to have at least 2 different subclones covering every base pair of a finished project.
4. Examine high quality discrepancies. These are regions where two high-quality sequencing reads have differences that cannot be explained by sequencing errors. There are many reasons why this can happen: one subclone could have mutated during the shotgun sequencing process; a read could be misassembled (put in the wrong place), especially if it falls inside a repeat region.
5. Confirm that the sequence is assembled correctly by comparing the sequence-derived restriction map to a fingerprint that is obtained by digesting the original clone with restriction enzymes and running these fragments on an gel.

Gap: (In sequence alignment) mismatch in the alignment of two sequences caused by either an insertion in one sequence or a deletion in the other. (In assembly) the regions without sequences.

Reference:

Gordon D, Desmarais C, Green P. Automated finishing with autofinish. *Genome Res.* 2001 Apr;11(4):614-25.

Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing. *Genome Res.* 1998 Mar;8(3):195-202.

Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 1998 Mar;8(3):186-94.

Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 1998 Mar;8(3):175-85.