GEP Hybrid Assembly Walkthrough

Developed by Christopher Shaffer, with input from GEP members Don Paetkau, Michael Rubin, Laura Reed

Prerequisites

Consed version 25 or higher
Familiarity with Consed, (For example, prior training with "Using Consed Graphically" and the "Drosophila Finishing Problem Set")

Files for this Exercise

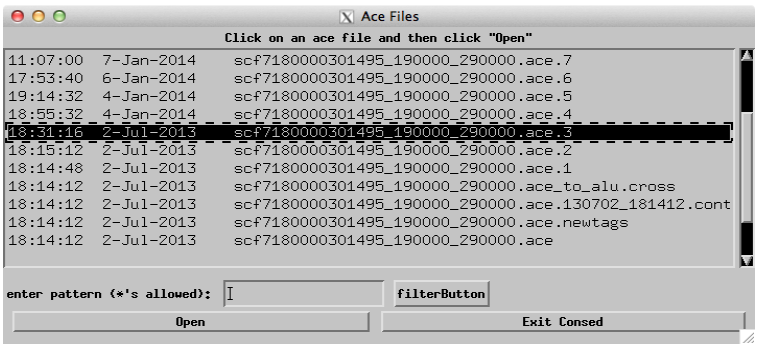
Project scf7180000301495_190000_290000

Introduction

This walkthrough will illustrate the techniques for consensus error correction as well as closing gaps using the *Drosophila biarmipes* project scf7180000301495_190000_290000. Note that this walkthrough assumes the reader is already familiar with *Consed* and the exact details on how to accomplish many of the tasks are not given. Many of the figures will not match exactly the images obtained by the user (even if they follow the protocol exactly). Users of this walkthrough are expected to have sufficient experience to interpret any differences and determine if they are significant or trivial differences. As such, users of this walkthrough should be very familiar with the techniques covered in in the "Using Consed Graphically" walkthrough and "Drosophila Finishing Problem Set" exercise (available on the GEP web site).

Set up

Launch X11 and open a new `xterm`; navigate to the `edit_dir` of the *D. biarmipes* project scf7180000301495_190000_290000 (e.g. `cd scf7180000301495_190000_290000/edit_dir`). Enter `consed&` at the `xterm` prompt. The "&" will keep your terminal active in case you need to use it later. Open `scf7180000301495_190000_290000.ace.3`. Select "No" if a prompt appears that asks if you would like to apply edits from the edit history (`.wrk`) file.



When improving hybrid assemblies, we will use custom settings that differ from the default Consed settings. You will need to verify these settings (and change these settings as necessary) each time you launch Consed.

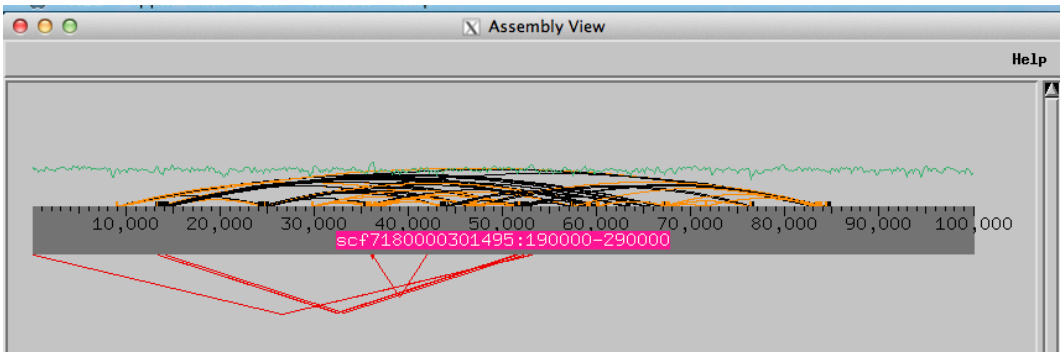
First, in the main Window, select “Options -> General Preferences”. Check that the Threshold for Low Consensus Quality (highest low)” is set to 25, and the “Threshold for High Quality Discrepancy (lowest high)” is set to 30.

Now double click on any contig (e.g. scf7180000301495:190000-290000 in this project) to open an Aligned Reads window. In the “Dim” menu, verify that the Dim option is set to “>Dim Nothing”. By default Consed will dim the unaligned regions at the ends of reads. This makes a lot of sense when working with Sanger reads that often have vector sequence at the end, it does not make sense when working with Illumina reads. In Illumina reads **ALL** data is relevant and should not be given a black background, hence the “Dim nothing” setting.

In the Sort menu, click on the item “Sort Options and Help”. In the dialog box that appears, change the “Display reads sorted alphabetically or by strand/left read end?” field to “Strand/Left End”. Change the “When you click on the consensus, how do you want reads sorted at cursor position” to “by base”. Be aware that with this setting means that the screenshots in this walkthrough may not exactly match the image on your computer screen. However, it does result in the ability work more quickly and efficiently when doing actual finishing so these settings were used throughout the walkthrough.

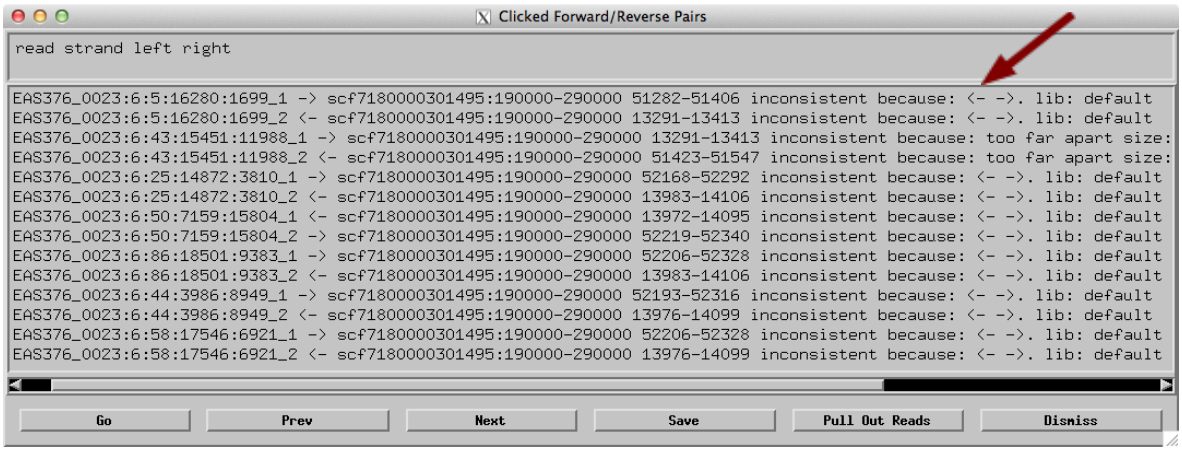
Assembly view

This project has a single contig of 100kb. Click on the “Assembly View” button on the *Consed* Main Window to open assembly view. Run cross_match to detect sequence matches within this contig and identify regions with a high density of discrepant read pairs.

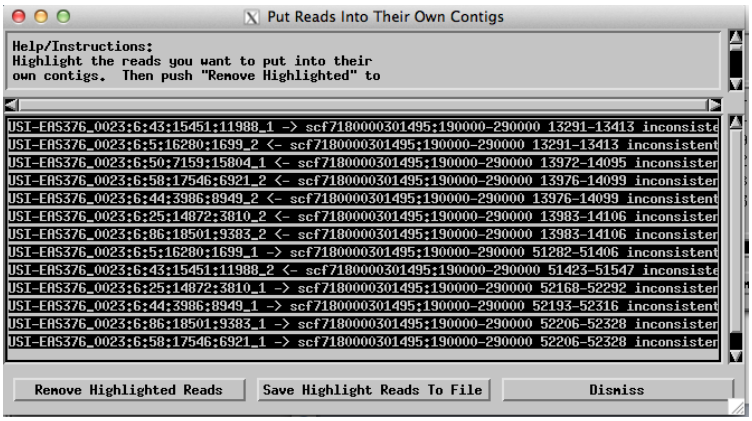


In this project, there is a cluster of 7 inconsistent mate pairs where one member of the mate pair is placed at around 13kb while the other member is placed at around 51kb. The number of discrepant reads of this type will vary with each project. While it is not necessary to remove these inconsistent reads due to the typical very high levels of coverate, it is recommended as it will likely reduce the size of the HQD list that must be analyzed.

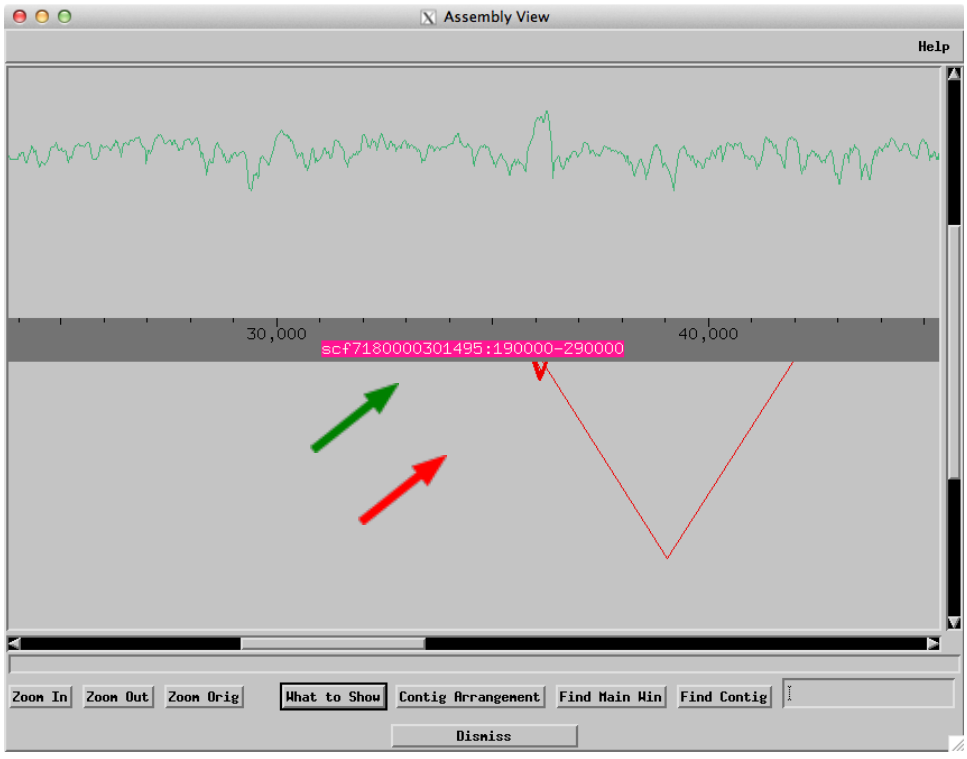
As an example of how to remove reads of this type, click on the cluster of red lines that spans from 13kb to 51kb of the contig. This will bring up a list of all the discrepant reads in that group. Note how most pairs are mapping in an inconsistent orientation (i.e. the paired end reads are pointing away from each other), hence, both distance and orientation are evidence that these reads are improperly mapped.



Click the "Pull Out Reads" button. In the next window **select all the reads** and click the "Remove Highlighted Reads" button. This will put each read into its own contig (Contig290001 through Contig290014). As with any procedure that changes the assembly (i.e. moves any reads into or out of a contig or makes any tears or joins), you must save the assembly before making any additional changes to the assembly. The project was saved as `scf7180000301495_190000_290000.ace.4`, to if you think you have made a mistake you may quit Consed and load this ace file. Throughout the walkthrough other ace file names will be given, they can be loaded to set the state of the project before continuing.



Navigate to Assembly View and notice another set of inconsistent mate pairs that spans from the beginning of the main contig to the region at around 53kb. These are also inconsistent because they are in the incorrect relative orientation and because they are too far apart from each other. Pull out these inconsistent reads from the main contig using the protocol described above (i.e. click on red lines, click to pull out reads, select all reads, remove highlighted reads, save assembly).



The set of reads that extends from 36k to 42k (red arrow in figure above), contains only two reads and for the purposes of this walkthrough will not be removed. Finishers are free to develop their own policy in regard to the number of reads in a cluster that would justify their removal from the main contig. The last set of inconsistent reads is a small cluster around 36 k (green arrow). You may need to “Zoom In” to see these. There are sufficient

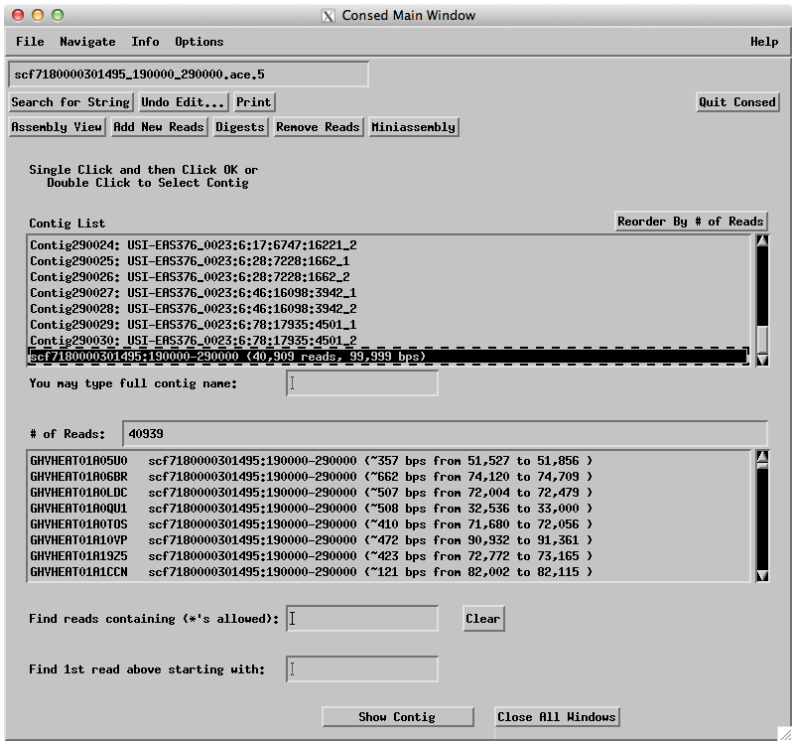
reads here that if this were a real project their removal is recommended. However for the purposes of this walkthrough they were left in the project.

Before continuing with the rest of this walkthrough, readers who wish to have an exact match to the screenshots shown in this walkthrough may wish to quit Consed, restart and open `scf7180000301495_190000_290000.ace.5`. (Remember to change the Consed settings described at the top of page 2 after you open the ace file.)

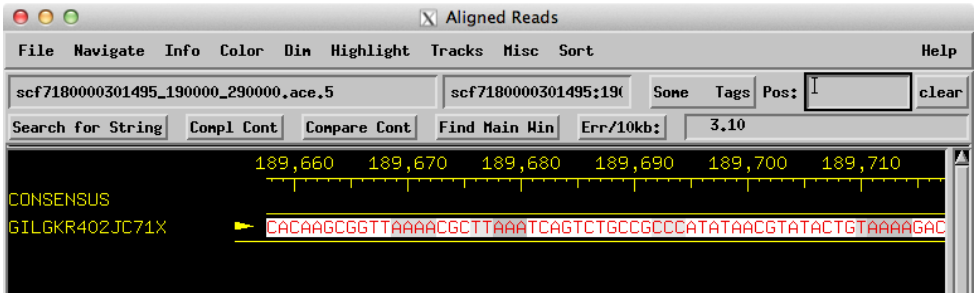
Resolving base errors at mononucleotide runs

The primary goal of the GEP sequence improvement project is to correct consensus errors within mononucleotide runs (MNR). The secondary goal of the sequence improvement project is to see if there is sufficient Illumina data to close gaps. Optionally, finishers may design primers to carry out PCR/Sanger to add new data to the project. This can be used to resolve gaps and regions with low consensus quality (discussed below). You should talk to your mentor about primer design and Sanger sequencing, this will only be done is special cases if you are doing a wet lab component to your finishing project.

This walkthrough will address the primary goal of resolving the MNR regions before addressing gaps and low consensus quality regions. However, because of the amount of time required to produce additional sequencing data, if you are planning on doing PCR/Sanger we recommend that you begin with primer design. You can work on correcting errors in MNR regions while waiting for the results from your Sanger reactions.

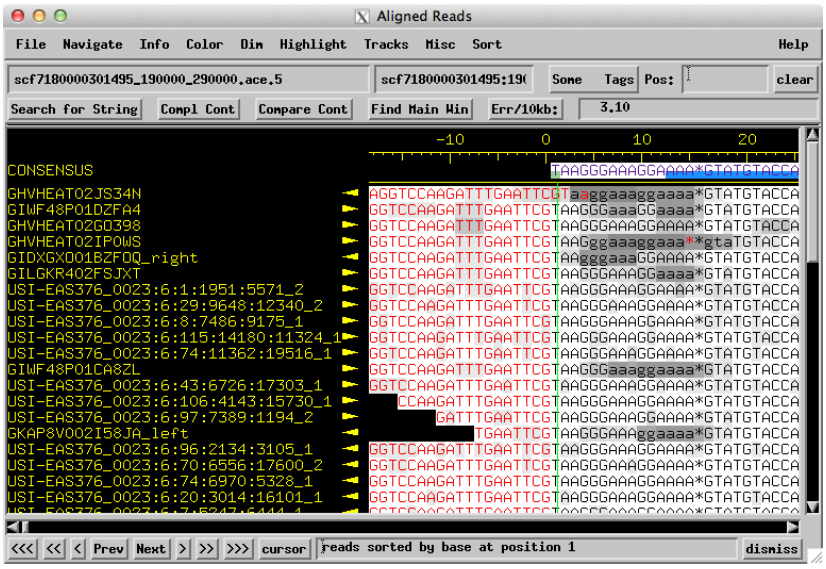


In the “Contig List” section of the Consed Main Window, you will see the project now has a long list of the individual reads (we removed them from the main contig above). The main contig we are working on is at the bottom of the list (scf7180000301495:190000-290000). Scroll down to find the main contig in the “Contig List” and double click on the main contig to open it in an Aligned Reads window. Alternativly open the align reads window directly from the assembly view.

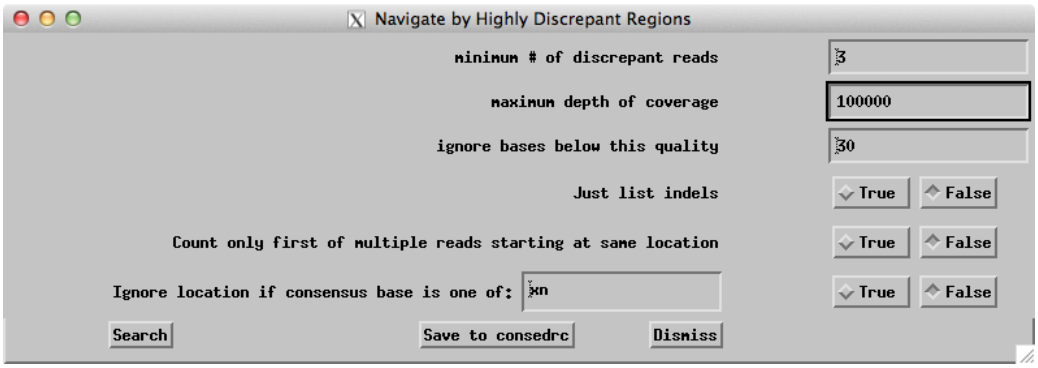


Depending on your settings you may see that the base numbering system starts with 189,660, this is a new feature available in Consed v25 that allows for proper numbering when a project is actually a subsection of a larger scaffold. This project is derived from a portion of the genomic scaffold scf7180000301495 in the *D. biarmipes* whole genome assembly and the numbering here indicates the location within that scaffold.

You can change the numbering system so that base 1 is the first base in the project. In the Aligned Reads window, click on the Misc menu and select “Turn On/Off User-Defined Consensus Scale Numbers”. This will temporarily change the numbering scheme for as long as Consed is running, If you quit and restart Consed the numbering system will change back. If you wish to permanently change the numbering scheme, you can delete the “startNumberingConsensus” consensus tag at the beginning of the contig. Click on the “<<<” button to navigate to the start of the sequence in this project.

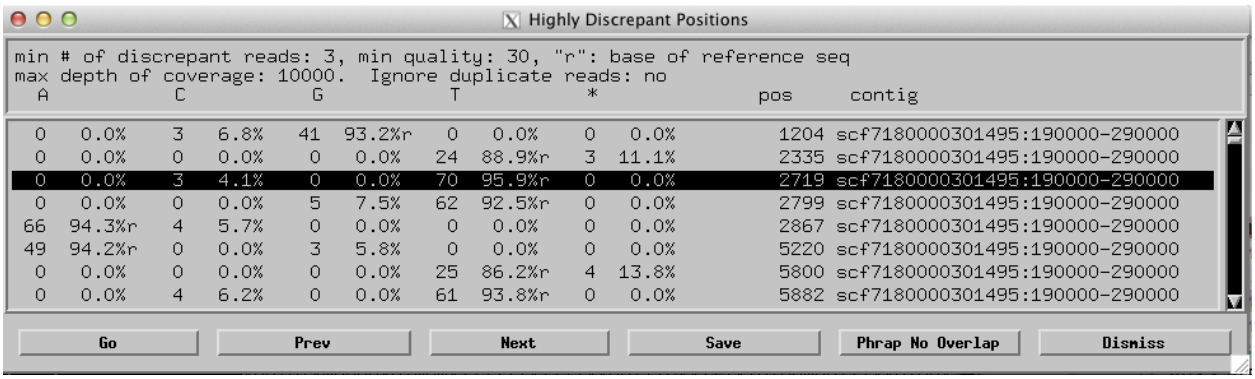


To screen for base errors, return to the *Consed* Main Window and select Navigate -> "Search for Highly Discrepant Positions". Given the high read depth and the large number of reads that are improperly placed in these assembly, sites with one or two HQD's are quite common and seldom indicate a genuine problem with the consensus. As an initial screen for problem areas in the main contig, we will set the "minimum # of discrepant reads" field to 3. This focus the list of discrepant regions where a consensus error might actually exist. To focus on regions where the Illumina data does not support the consensus set the "ignore bases below this quality" field to 30. Click "Search".

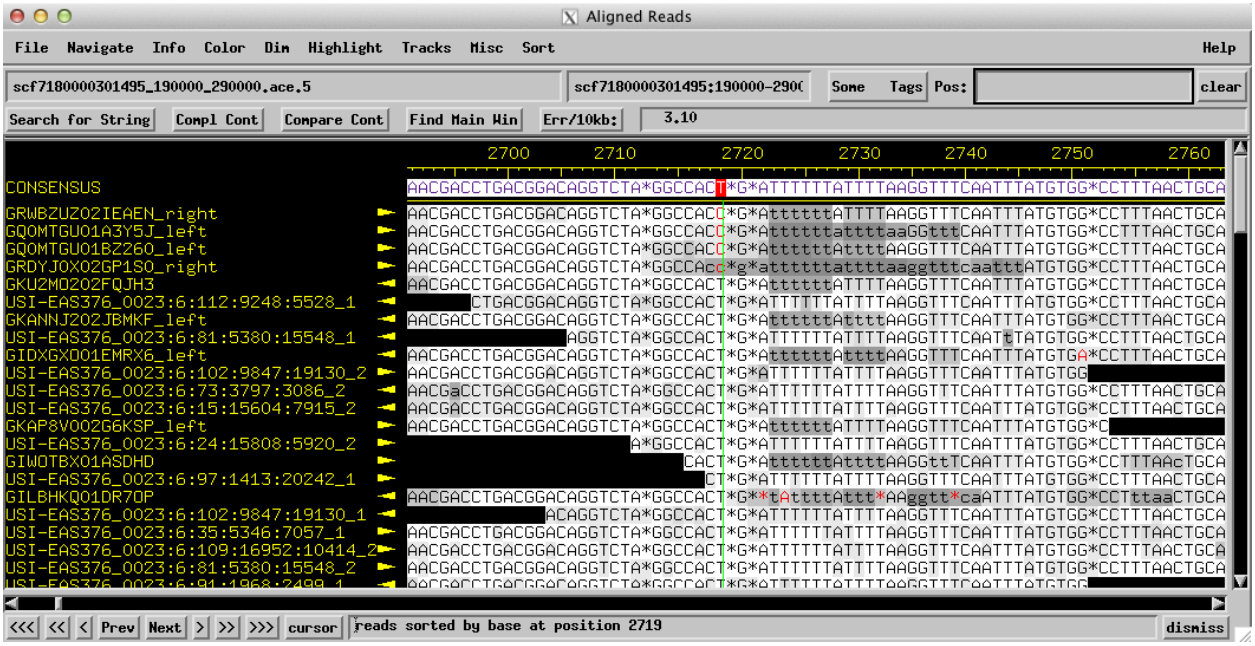


The resulting list will have many regions where the consensus is correct even though there are 3 high quality reads that disagree with the consensus. In many cases, the discrepancies can be attributed to errors in the 454 reads that show different number of bases compared to both the Illumina reads and the consensus. (You can use the read name to distinguish 454 reads from Illumina reads; Illumina reads have the prefix "USI-", 454 reads will start with a "G") Some of the other discrepant regions on the list are caused by reads that have been incorrectly placed in the assembly (e.g. because of large transposons or other repetitive regions in the genome) by the mapping program.

The basic strategy is to navigate through each item in the Highly Discrepant Regions list and look for discrepant regions that are associated with a MNR. When you find a region associated with a MNR (within 5 bases), you should inspect the region carefully and either confirm or edit the consensus based on the available evidence. Note that because of the known weaknesses of the 454 sequencing technology in resolving the correct number of bases in long MNR's, finishers should rely on the Illumina data when determining length.

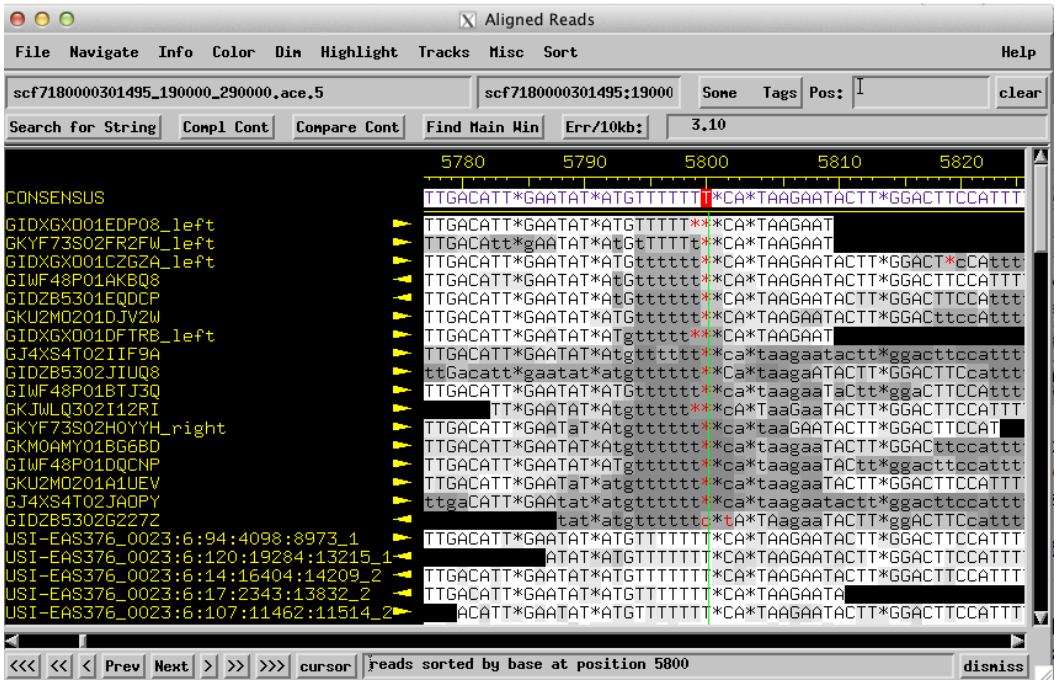


Since adjacent GEP sequence improvement projects overlap with each other, the finishers can ignore the discrepancies that are found within 2500 bases of the ends of the project (because these problem areas will be resolved by the finishers working on the adjacent projects). Consequently, we will ignore the initial highly discrepant regions and examine the region at base position 2719. Inspection of that region in the Aligned Reads window shows that the discrepant position (i.e. 2719) is not near a MNR. Because we are sorting "by base", the four reads with a discrepant C (3 of which are high quality) are listed above all the reads with the T.



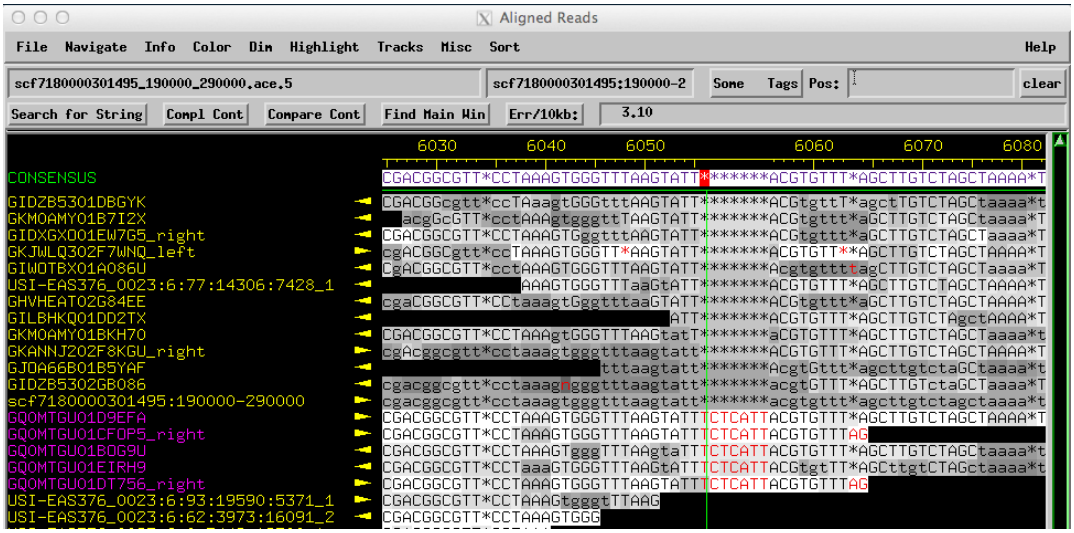
This discrepancy may be due to a base calling error, mis-mapping of the read or it is possible that this site is polymorphic. Regardless of why this discrepancy exists, there is insufficient evidence to support the hypothesis that the consensus is incorrect: there are 70 reads here (with quality score 30 or above) that agree with the consensus and show a T and only 3 high quality reads that show a C.

In fact, given that the discrepant position is **NOT** part of the nearby MNR, we could simply move on to the next region with 3+ HQD's. However, if you examine these reads carefully, you will find additional discrepancies further downstream (e.g. at 2799 of the consensus). This makes it very likely that these reads actually belong somewhere else in the genome and it strengthens the argument that the consensus should remain a T. Given the high frequency of mis-mapped reads, discrepancies of this type will not be examined carefully. Instead finishers should focus their efforts exclusively on discrepant regions associated with MNR's. Click on the "Next" button on the Aligned Reads window to navigate to the next highly discrepant region. Continue to click next until you navigate to a discrepant region that is associated with a MNR. The first discrepant position within a MNR is at 5800.



The sorting "by base" option has placed all the reads with the pad (*) at the top of the Aligned Reads window, below these are all the reads with a T at this position. There are quite a few 454 reads in this region that show fewer number of T's than are in consensus (i.e. 16 of the 454 reads show a pad). Use the scroll bar on the right to scroll down and examine the Illumina reads. All the high quality Illumina reads that aligned to this region agree with the consensus and shows 7 T's (Illumina reads start with the prefix **USI-**; 454 read names are shorter and start with **G**.) Because all the Illumina reads agree with the consensus, there is insufficient evidence to change the consensus and we will keep the consensus at 7 T's and move on.

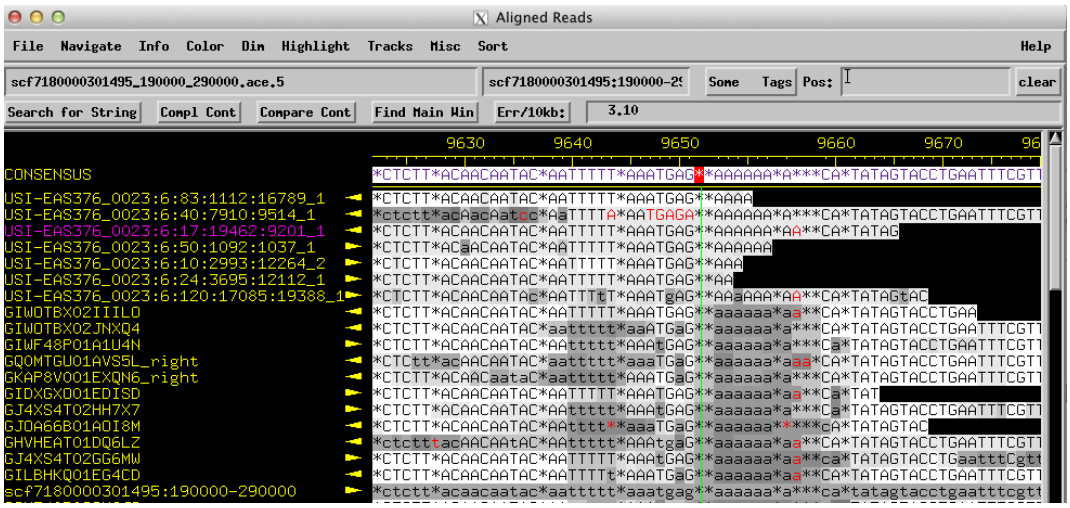
Click “next” in the Aligned Reads window to navigate to the next discrepant region. The next location on the HQD list that shows an interesting discrepancy is located at 6055. Examination of the region shows five 454 reads that have a 7 bp insertion (TCTCATT) compared to the consensus (if you can only find 3 discrepant reads, make sure “Dim nothing” is selected and look again). Again the preponderance of evidence suggests that the consensus is correct (i.e. there are ~90 reads covering this region, 85 of which are high quality that agree with the consensus while only a few disagree with the consensus). The skew in the distribution of reads that disagree with the consensus (3%) compared to the percentage of reads that agree with the consensus (97%) makes it extremely unlikely for the discrepancy to be caused by a polymorphic site.



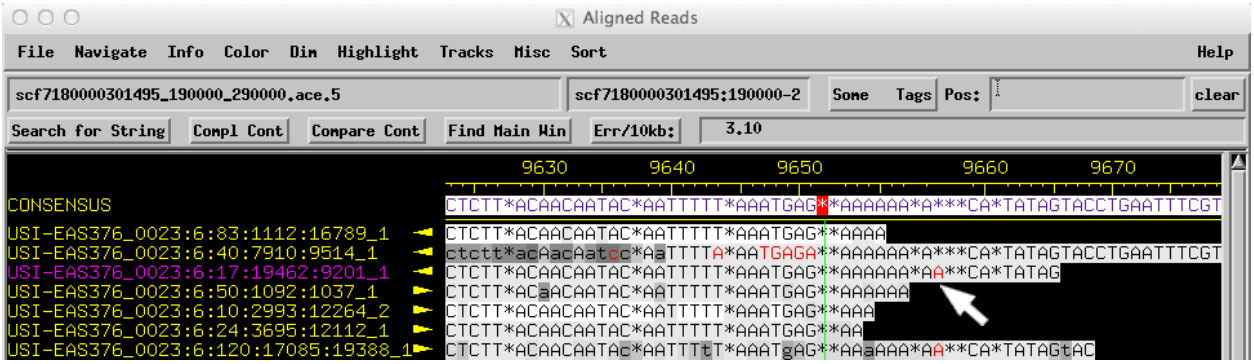
The next discrepant MNR site with 3 or more HQD’s is at 7557. It is associated with a MNR of 4 T’s. However there are 52 (high quality) reads with a T at this position and only 3 (high quality) reads with a C, so again there is insufficient evidence to change the consensus.

Click on the “Next” button in the Aligned Reads region to navigate to the next discrepant region. There are several other regions that are associated with MNR’s downstream, but in each of these cases the consensus is fine, no changes are justified by discrepant reads.

The next interesting region is located at 9651. This region is an example of a problematic alignment which necessitates careful examination. Here the consensus sequence (ignoring pads) is AAATGAGAAAAAAACATAT. When you scroll down in the Aligned Reads Window, you will find many base discrepancies in this region. However no one position is discrepant across all the high quality Illumina reads. Note how some of the high quality Illumina reads show a discrepant A which differs from the pad just to the right of base 9651. Other high quality Illumina reads have a (consistent) pad at that position (e.g. USI-EAS376_0023:6:17:19462:9201_1.

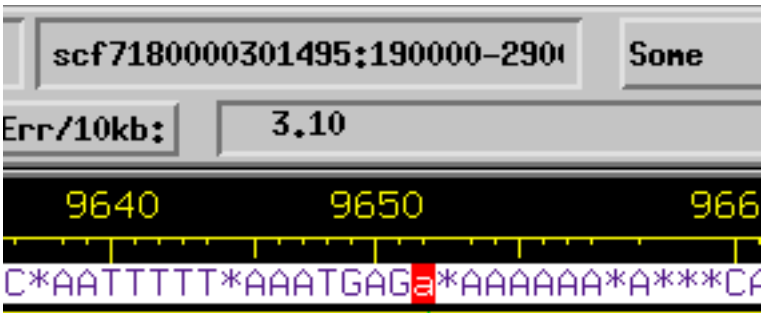


In this region, careful examination reveals that all of the properly mapped high quality Illumina reads have 8 A's irrespective of how they were aligned to the consensus. Hence the available evidence suggests we should change the consensus from 7 A's to 8 A's.



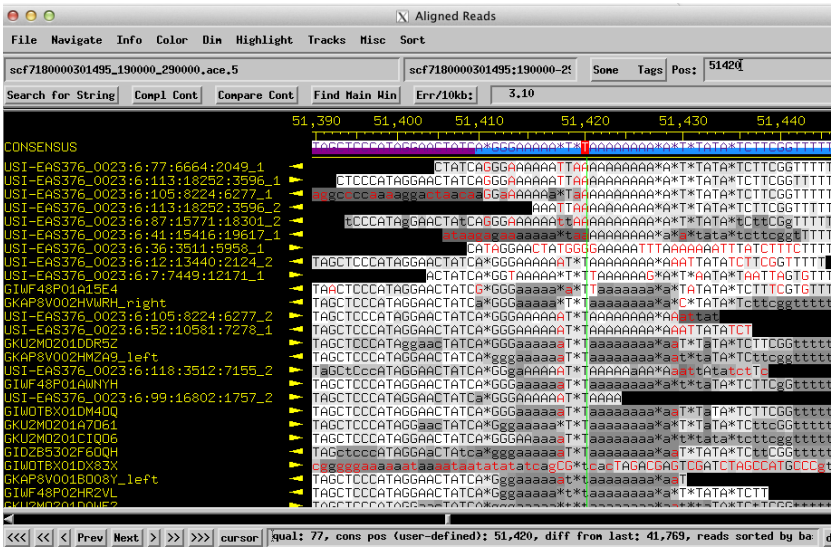
Inconsistent alignments of this type are a common issue with *Consed*. This is particularly true for regions with long MNR's or many smaller MNR's. For this reason it is very important to inspect the discrepant regions carefully, counting the number of bases in each read if necessary, to determine the consensus sequence.

If they prefer, finishers can use the standard technique of opening a trace window for one of the Illumina reads and using change consensus to make the proper edit. However, unlike previous versions of *Consed*, *Consed* 25 allow users to edit the consensus sequence directly. To add an extra A to the consensus using this technique, click on one of the pad (*) adjacent to the 7 A's in the consensus. Hit the A key to change the consensus from a pad to an **a**. Note that all edits to the consensus are kept as lowercase (i.e. low quality edit) by Consed.



After editing the consensus save the assembly (`scf7180000301495_190000_290000.ace.6`).

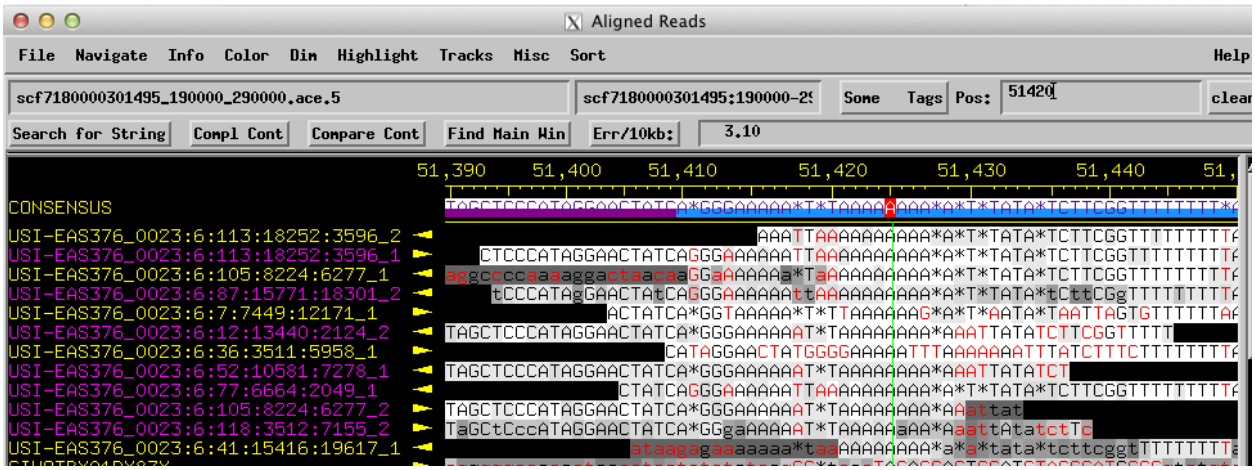
For practice, you can continue to assess regions on the HQD list that are associated with MNR's and correct the consensus sequence if it were supported by the Illumina data. There are several additional regions in this project that requires corrections to the consensus sequence but they will not be discussed in this walkthrough. The next region discussed will be the region around the long A MNR around 51,425. This regions is a good example of a very complex region where the reads must be analyzed one by one in order to resolve the consensus.



This region has 2 MNR of A's that are separated by 2 T's. Regions with multiple MNR's are particularly error prone for 454 sequencing and are also more likely to have been misaligned by *Consed*. If some of the reads have a completely black background at the end of the read be sure to use the Dim menu to select "Dim Nothing".

A careful examination of this region shows that many of the red discrepant bases in this region can be attributed to misalignments and are not genuine discrepancies in the lengths of the MNR's. The consensus hasa sequence of GGGAAAAATTAAAAAAAAAAATTAT, with the critical issue of how many A's are in the two MNR's. Click on the middle A at position 51,425, this should bring all the Illumina reads to the top. Irrespective of where they are

sorted or how they were aligned careful examination shows that 7 of the Illumina reads (highlighted in the figure below) have the sequence 3 G's followed by 6 A's followed by 2 T's then 11 A's followed by TTAT (i.e. GGGAAAAAATTAAAAAAAAAAATTAT).



Of the reads that do not match, the read name that ends with :3596_2 (top read in the figure above) does not cover the entire region being discussed and is therefore uninformative. The read name that ends with :6277_1 only disagrees with this sequence at positions that have very low quality. In contrast, the read names that end with :12171_1 and :5958_1 contains many HQD's compared to the consensus throughout their entire length. Given that this region has a blue repeat tag, it is very likely that these reads been improperly mapped to this region and the differences indicated by these reads are unreliable. The read name that ends with :9617_1 is very low quality for most of this region and the evidence it provides would not be sufficient to overrule the 7 high quality reads that are all consistent with each other.

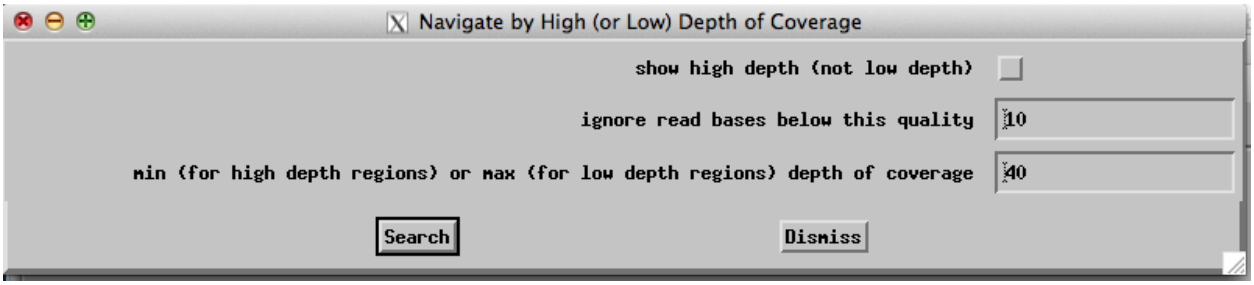
Collectively, our analysis of the high quality Illumina reads that aligned to this region suggests that the correct consensus sequence is GGG AAAAA**A** TT AAAAAAAAAA**A** TTAT. Edit the consensus adding an addition A to each mononucleotide run and save the assembly. (scf7180000301495_190000_290000.ace.7)

There are other regions further downstream in this contig with at least 3 HQD's and where most of the Illumina reads disagree with the consensus. Analysis of these regions is left as an exercise for the reader.

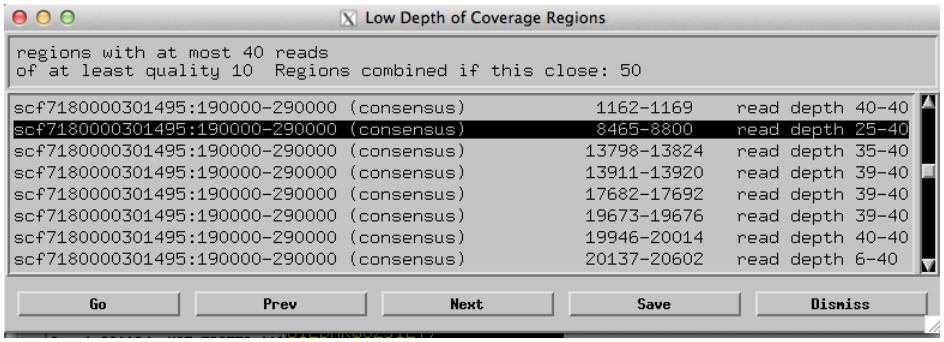
Checking areas with fewer than 40 reads

The protocol for resolving MNR's described above assumes that the discrepant region contains many reads. Setting the minimum of at least 3 HQD's allows us to filter out many locations with mis-mapped reads. However, we will need to use an alternate approach to check for errors in MNR's found in regions with low coverage. A finisher must navigate to these regions and carefully inspect them for any potential errors in the consensus sequence.

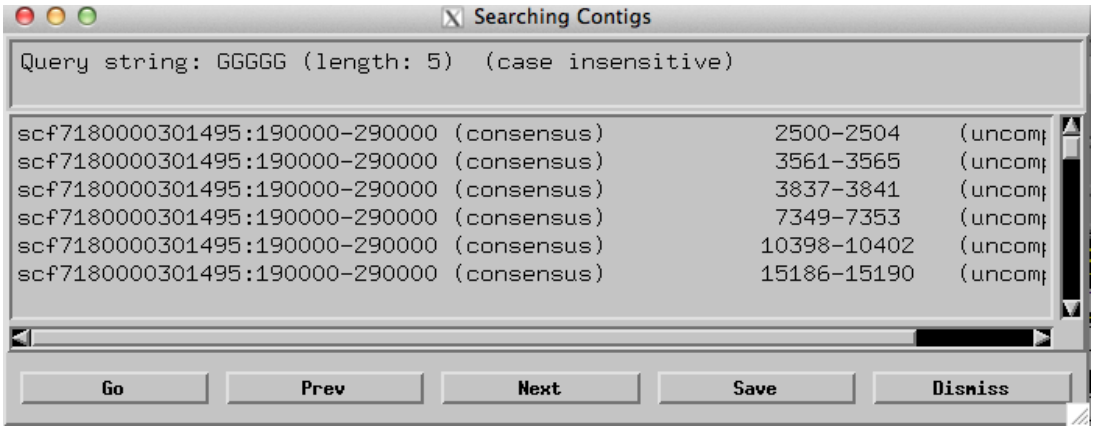
To search for regions with low read coverage, use the “Main Window -> Navigate -> Search for High (low) Depth of Coverage” menu. In the “Navigate by High (or Low) Depth of Coverage” window, uncheck the “show high depth (not low depth)” field to search for regions with low read coverage. Set the “ignore read bases below this quality” field to 10 and then set the “max (for low depth regions) depth of coverage” field to 40. Click Search.

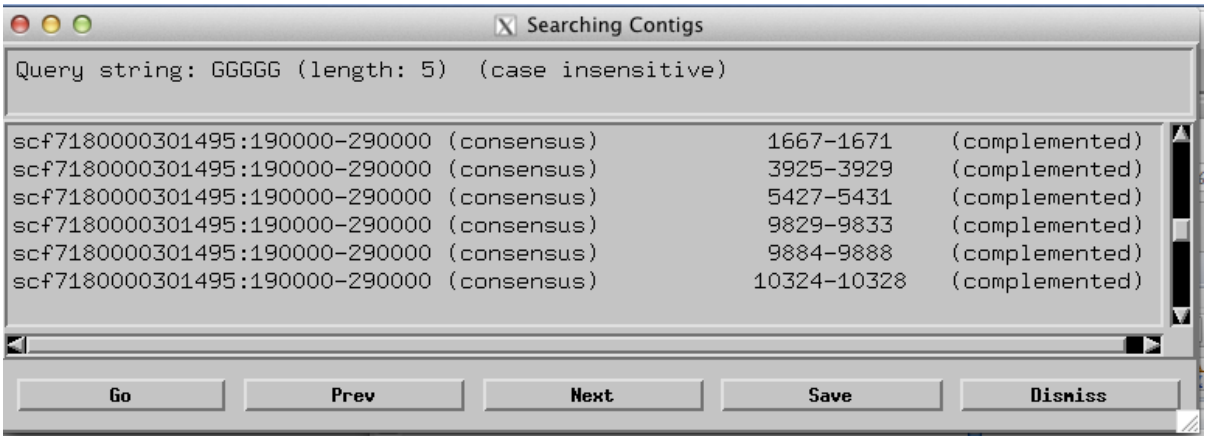


The “Low Depth of Coverage Regions” navigator window will appear and you can use this to quickly navigate to the areas with low read coverage. The entries at the beginning of the list correspond to the single reads that we have previously removed from the main contig. Scroll down until you reach the entries that correspond to your main contig (i.e. scf7180000301495:190000-290000), remember that the first and the last 2.5 kb do not need to be finished.

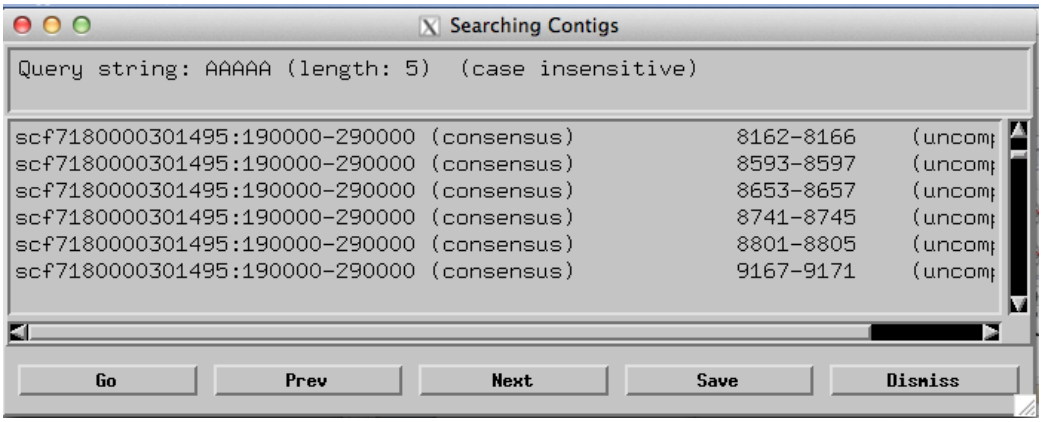


The first low coverage region we need to inspect spans from 8465 to 8800. Search for any sequence within this region with a MNR of 5 or more. This can be accomplished by either manual inspection (easily done if the region is small) or by using “search for string”. Be sure to search with both GGGGG (which will find both GGGGG and CCCCC) and AAAAA (which will find both AAAAA and TTTTT). Searching with GGGGG shows no MNR of 5 or longer within this low coverage region. Also, remember that when searching for string regions that match the complement of the query sequence (i.e. locations with CCCCC) will be listed **AFTER** all the locations that matched the uncomplemented sequence (i.e. GGGGG). Hence you will need to scroll down to ascertain if any CCCCC MNR’s are found between 8465-8800. (In this case, there are no MNR’s of CCCCC in this low coverage region.) To avoid this scrolling issue finishers may wish to run 4 different MNR searches and cross-reference all 4 lists when looking for overlapping regions.





Searching with AAAAA shows 3 regions between 8465-8800 that should be carefully inspected, 8593-8597, 8653-8657, and 8741-8745. Examination of the three regions shows no discrepancies or inconsistencies that would require changes to the consensus sequence. Scroll to the "complemented" section of the "Searching Contigs" list to check for any TTTTT MNR's in this region (there are no TTTTT MNR's that overlap with the low coverage region at 8465-8800).

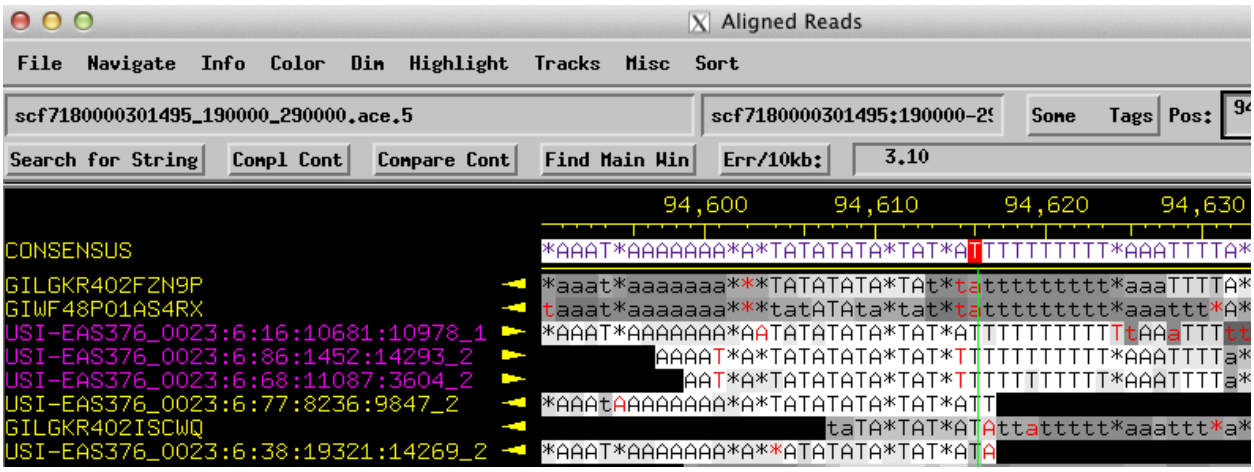


According to the GEP sequence improvement standard for the hybrid assemblies, each consensus position must be supported by a minimum of two reads that agree with each other and are each of sufficient quality. These reads can be either 454 or Illumina reads (or a combination of both).

"Sufficient" quality in this context means that the read position has a phred (quality) score of at least 20 in each read and that the finisher is confident that the read has been mapped to the correct region. To have high confidence that the read is properly mapped, it should have no more that one HQD anywhere in the read compared to the final consensus. If there is only a single high quality read that supports the consensus, then the finisher will need to add a "data needed" tag to the region and highlight the presence of an unresolved low quality region in the final finishing report form. If you are implementing a wet lab PCR/Sanger pipeline you may be asked to cover the region to resolve the issue, check with your mentor. If you are not implementing your own PCR/Sanger experiments, you should NOT design primers for these low consensus quality "data needed" regions.

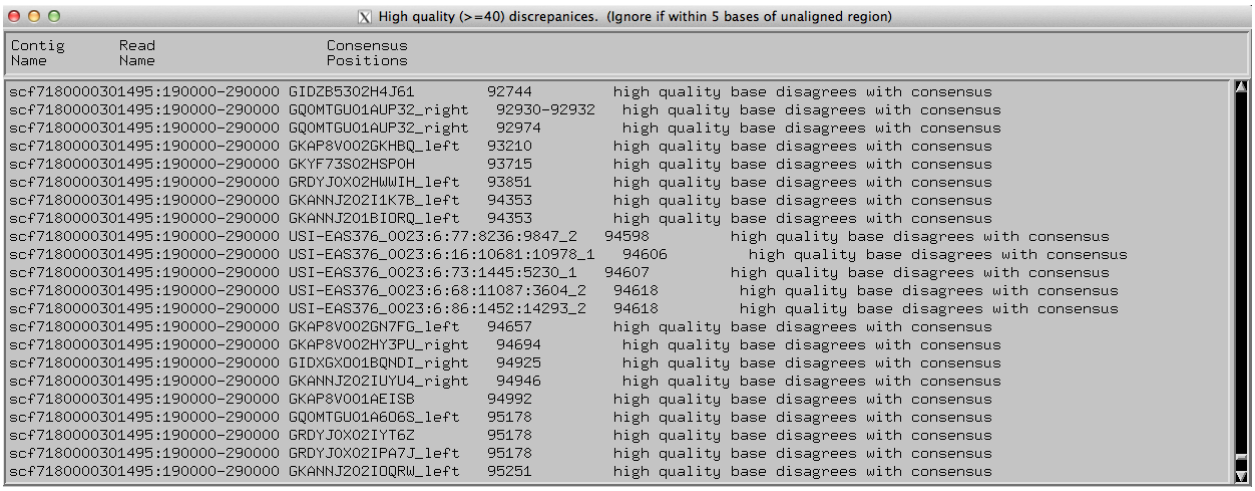
For practice you may continue to work through the regions in the low coverage list by manually inspecting these low coverage regions and cross-reference with the MNR search for string lists. Examine each location to confirm the consensus in these regions.

The next low coverage region discussed here is located at 94415-94757. Within this region is a MNR of 10 T's in the consensus starting at base 94619 Yet again there are alignment problems that make this area difficult to analyze.



While the consensus has 10 T's, careful inspection of the 3 high quality Illumina reads (highlighted in the figure above) all show 11 T's. Note how, because of the vagaries of the alignment algorithm used by *Consed*, no consensus position has 3 HQD's. Two of the reads are discrepant with the consensus A at 94618. The other read is discrepant with the pad just after consensus position at 94628. We will need to verify that these 3 reads contain no more than one HQD outside of this MNR to be confident that they have been mapped correctly to this region. Given that Illumina reads are all relatively short this can be done simply by manual inspection. To check this computationally, pull up the list of all HQD's

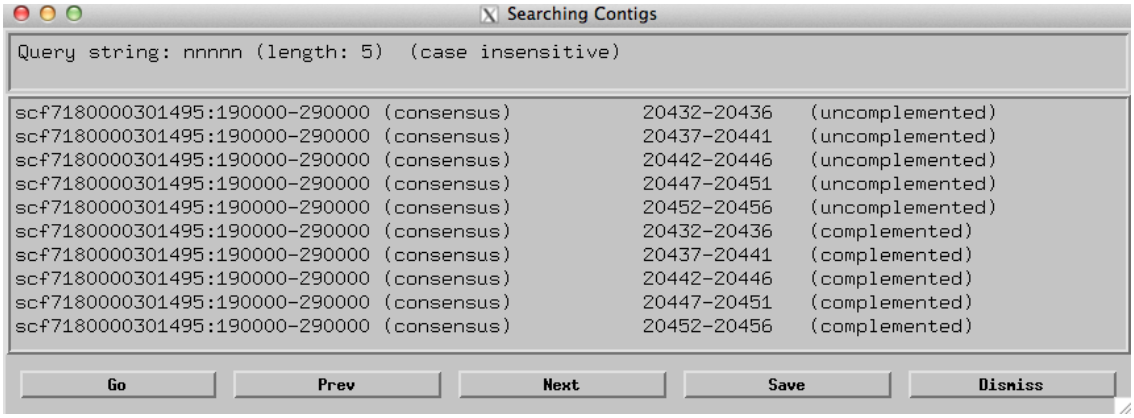
(Navigate -> High quality (>=30) discrepancies, >5 bp from unaligned region). Scroll down to the region of interest at ~94,000 and note which of the reads contain HQD's.



The HQD list shows that each of these three reads (that have an extra T compared to the consensus) only has a single HQD. In each case, the only HQD listed for each of these reads is the HQD that is associated with the incorrect length of the MNR in the consensus. Consequently, we can be confident that these three discrepant reads have been mapped to the correct region and we can correct the consensus using these reads. Correct the consensus (11 T's) and save the assembly. (`scf7180000301495_190000_290000.ace.8`)

Resolving gaps and low consensus quality regions

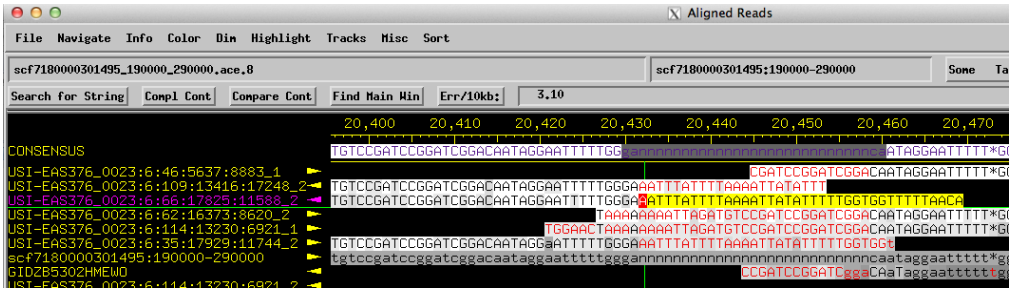
Because resolving gaps and low consensus quality regions sometimes require additional data from PCR/Sanger sequencing reactions, we recommend that finishers doing the optional PCR/Sanger pipeline begin their sequence improvement project by resolving gaps. The primary goal of resolving base errors in MNR's can be worked on while waiting for the PCR/Sanger results. To determine if your project has any gaps, simply use "Search for String" and search for "nnnnn" in the consensus.



These results show n's in the region around 20445 plus or minus about 13 bases. Double click on the first match (at 20432-20436) to navigate to this region. Notice that the gap region actually spans from 20432-20458.

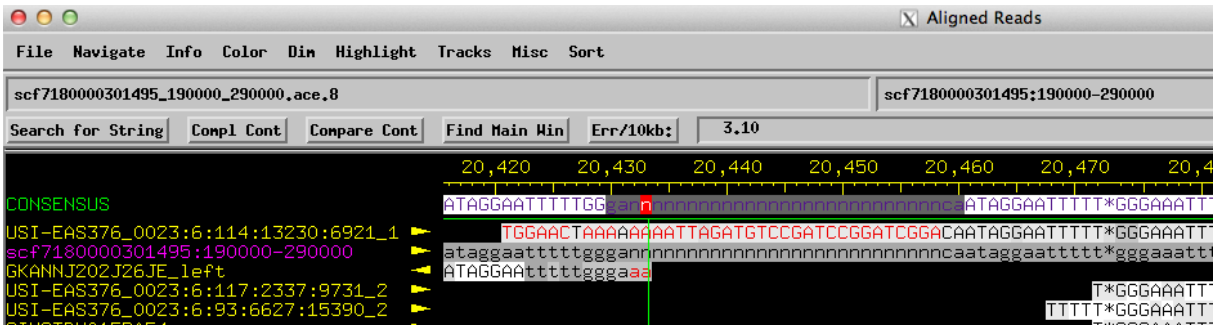
Because of how the projects are constructed (i.e. mapping 454 and Illumina reads against the published consensus sequence), many of the projects contain small gaps that can be resolved without additional data. In this example, there are multiple high-quality Illumina reads that span the entire gap. In addition, note how the bases that are aligning to the n's in the consensus actually match the sequence adjacent to the gap. This suggests that there is no missing data. Finishers may be able to detect the overlap by visual inspection or they can use the "Search for String" functionality in *Consed* to search for overlapping regions.

In this example, we will use the read USI-EAS376_0023:6:66:17825:11588_2 (which extends into and beyond the gap) to help resolve this region. The last bases in the consensus on the left side of the gap are TTTTTGGga, select the bases in read USI-EAS376_0023:6:66:17825:11588_2 immediately following this sequence to the end of the read and perform a Search for String.

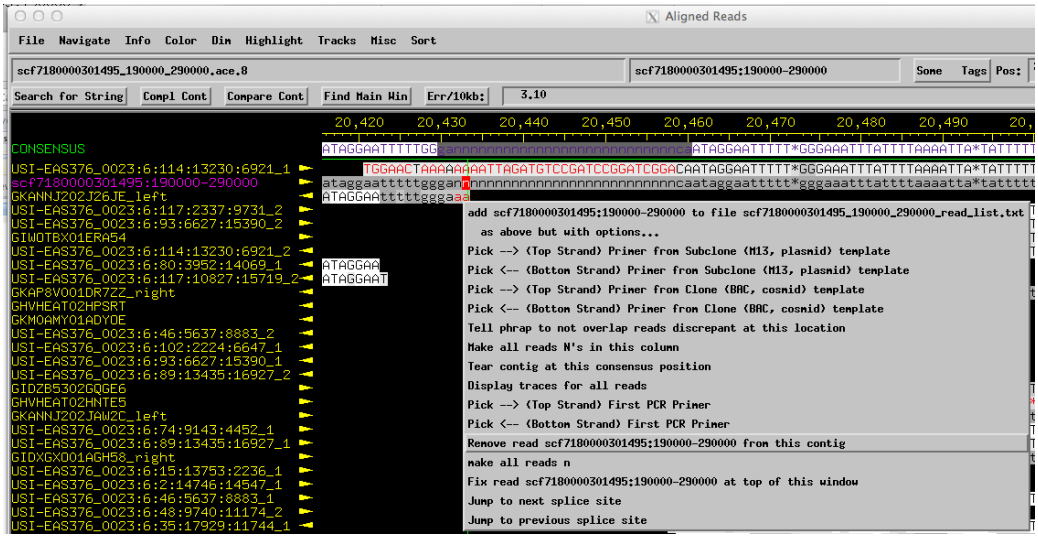


The results of "Search for String" revealed an exact match that is located just on the right side of the gap (at position 20477-20514). Visual inspection of this region reveals an overlap of a few bases on each side of the gap: ataggaatttttggga is actually repeated on each side of the gap. There are many Illumina reads that support the hypothesis that there is only a single instance of this sequence. Hence we might be able to resolve this gap by performing a force join to collapse the overlapping reads into a single region.

If our hypothesis were correct, then the assembly piece (scf7180000301495:190000-290000) for this project is wrong and contains a misassembly. If the assembly piece read remains in the contig, Consed will prohibit us from closing the gap. Thus, we must first pull out the assembly piece that was used to construct the initial assembly before attempting to reassemble this region. This fake read has the same name as the project: scf7180000301495:190000-290000.

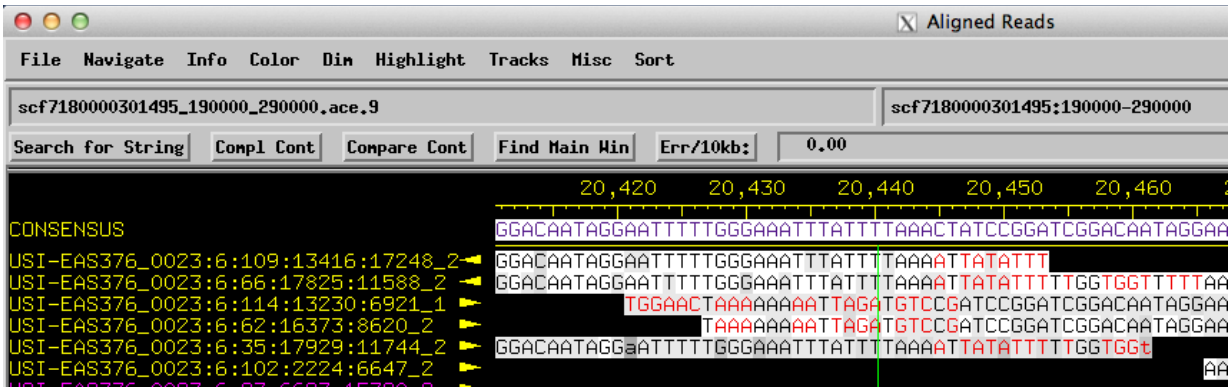


Remove the assembly piece read by right clicking on the read and select “Remove read scf7180000301495:190000-290000 from this contig”. In the “Remove Reads” dialog box that appears make sure that only the “scf7180000301495:190000-290000” read is listed in the “Reads to be removed” box on the right. The default settings are appropriate and do not need to be changed, click “Do it”. A new “Reads Removed” dialog will appear with a note that indicates we must save the assembly. Go back to the Consed Main Window and save the assembly `(scf7180000301495_190000_290000.ace.9)`.



Depending on your project, removing the assembly piece from the main contig could cause the contig to break into multiple smaller contigs. However, in this case, the assembly remained in a single contig. When the assembly piece was removed Consed attempted to calculate a new consensus, replacing the N’s with bases from the reads below.

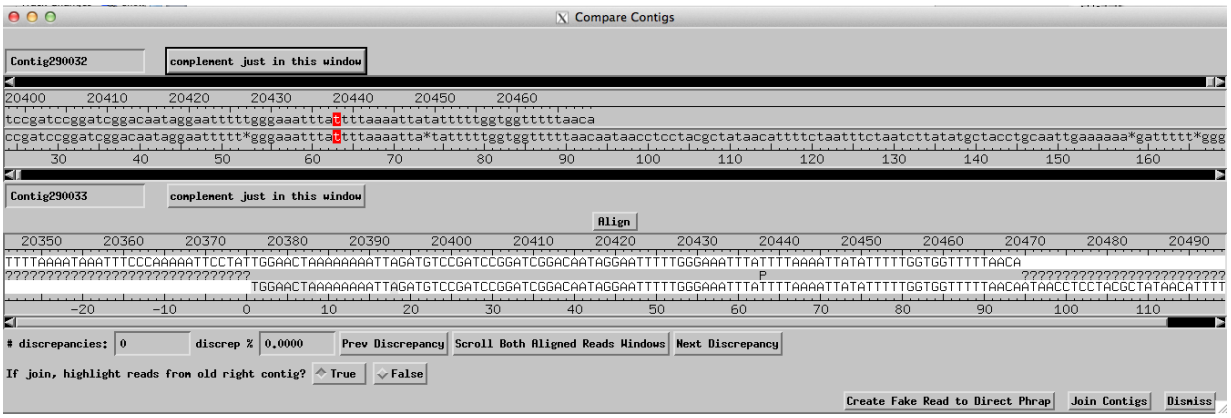
Perform a Search for String using the overlapping sequence we have identified previously (i.e. ataggaatttttggga) to navigate to the gap region that we are trying to resolve. Because the reads below the consensus are not properly aligned, there are many discrepant red bases in region.



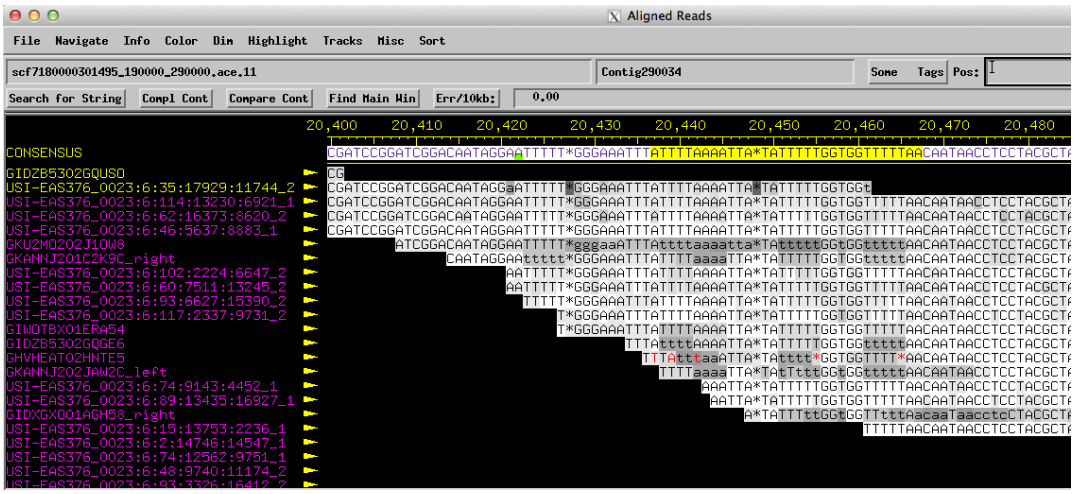
However, careful visual inspection of the region surrounding the location of the previous n's (bases 20432-20456) shows that the region remains the same: there are still two copies of the sequence ataggaatttttggga. To resolve the gap and create an assembly with no HQD's, we will need to tear and re-join the overlapping regions with the correct overlap. Right-click at base 20440 of the consensus and select "Tear contig at this consensus position". Because our hypothesis is that the two repeated regions should be collapsed into a single copy, it does not matter which reads ends up in the left (highlighted) contig versus the right (unhighlighted) contig as long as some reads go in each direction. Hence we will accept the default selection from Consed. Click on "Do Tear" to create two contigs, a left hand one (~20 kb) and a right hand one (~80 kb).

Save the assembly (`scf7180000301495_190000_290000.ace.10`). Use the standard "Compare Contig" technique to join the two contigs together (This is described in Consed exercise "A Complex Drosophila Fosmid"). A brief outline of the procedure is as follows:

Navigate to the far right end of the 20kb contig, use the sequence found there to perform a "Search for String" to identify the matching site in the 80kb contig. Using the "Searching Contigs" results window, navigate to the match at the end of the 20kb contig and click on "Compare Cont". Return to the "Searching Contigs" results window and navigate to the match at the beginning of the 80kb contig and click on "Compare Cont". Click "Align" on the Compare Contigs window and examine the alignment. if there are no high quality discrepancies (as is the case here) Click on "Join Contigs".

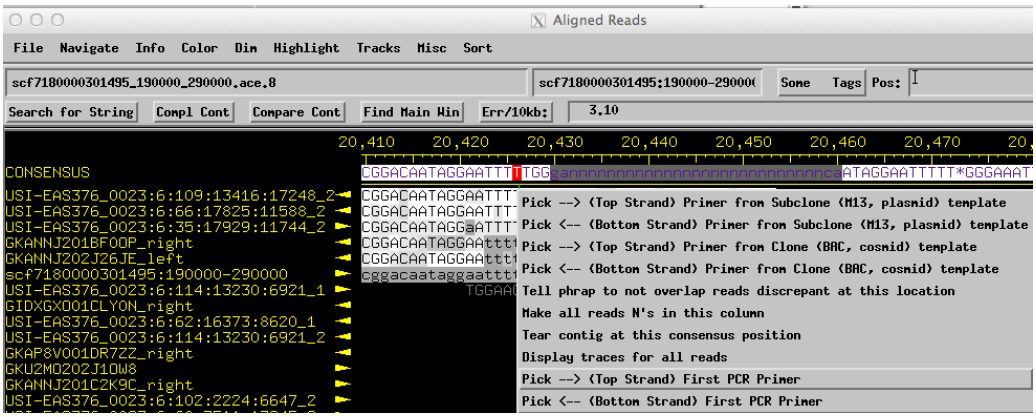


Save the assembly (`scf7180000301495_190000_290000.ace.11`). Inspection of the region after the above tear and join shows that the gap originally at 20432-20456 is no longer present. The resulting assembly is much better with no HQD's.

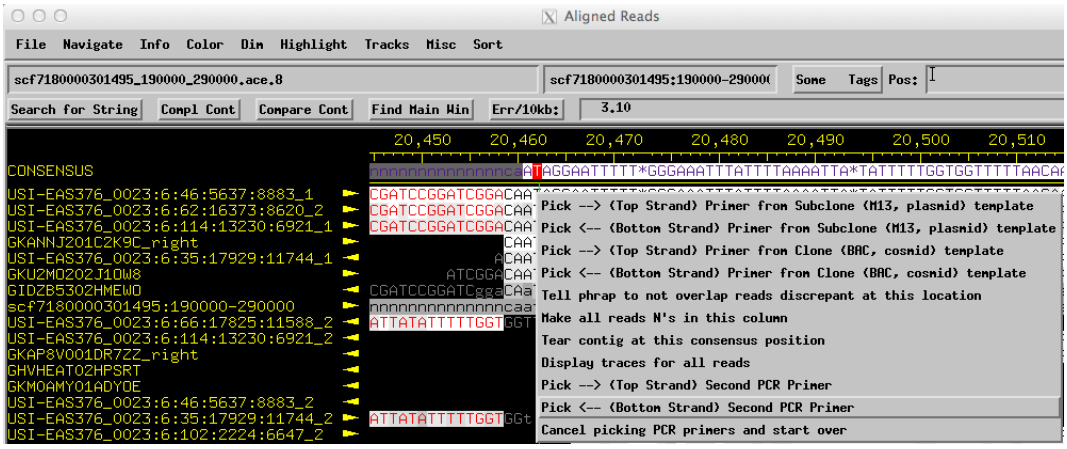


The PCR/Sanger pipeline

Some of the gaps in the *D. biarmipes* projects are genuine and requires additional sequencing data. For gaps of this second type new sequence data must be generated to fill in the missing bases. Because there is no template available for sequencing, new data will need to be generated by first generating template with PCR prior to sequencing with Sanger. To practice this technique, quit and then restart consed and open the ace file scf7180000301495_190000_290000.ace.8 (which still has the gap in the consensus sequence). While finishers should not design primers to cover low consensus quality regions, they should design PCR primers that span any unresolvable gaps. These primers can be used in an attempt to generate the necessary data to close the gap. Consed has a PCR primer picking protocol that will generate a list of possible PCR primer pairs: go to the gap (consensus position 20432), right click on the consensus position just to the **left** of the gap (at position 20426) and select "Pick --> (Top Strand) First PCR Primer"; dismiss the information dialog box.



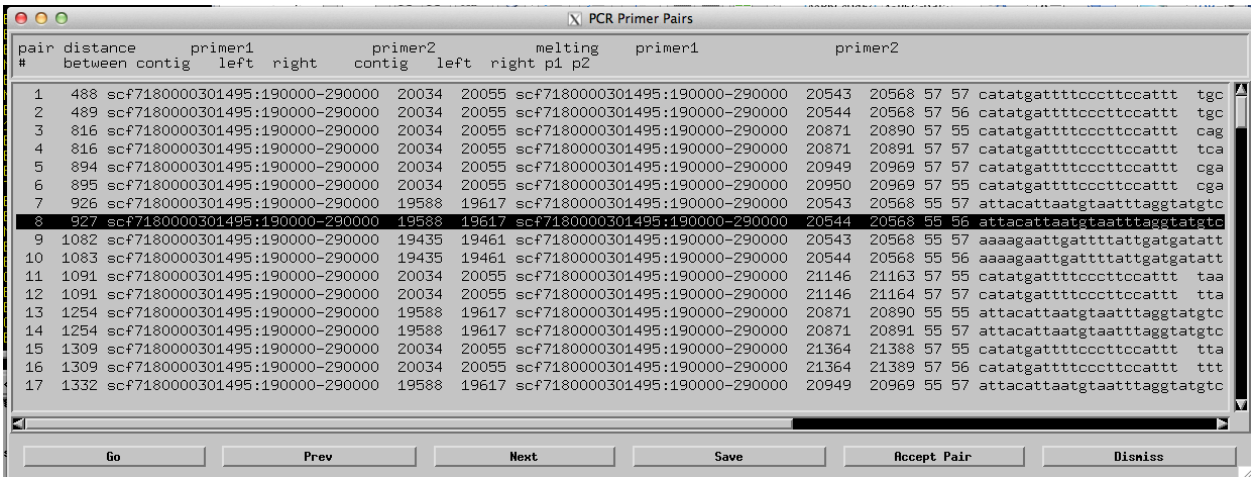
Right-click on the consensus just to the **right** of the gap (at position 20462) and select "Pick <-- (Bottom Strand) Second PCR Primer"



The criteria for picking among the suggested PCR primers are discussed in detail in the "PCR Primer Selection Guide" document. Finishers working slowly and carefully should select a single pair to attempt PCR/Sanger. Finishers who are constrained for time should attempt to pick two different sets of PCR primers for each problem area such that both forward primers are compatible with both reverse primers. In general, the size of the PCR amplicons should be kept less than 1000 bases if possible, and should be kept as small as practical in all cases. If all the possible sizes are larger than 1250 bases, then at least one of the two primers **MUST BE** within 350 bp of the region where new data is needed. If you cannot find primer pairs that satisfy either the "1000 bp size limit" or the "one end within 350 bases of end" rule then you will also need to create a third sequencing primer that is close enough to the problem area to use to prime the Sanger sequencing reaction.

In this practice case, we will carefully examine the list of PCR primers suggested by *Consed* and attempt to find 4 primers (2 left and 2 right) in which all 4 combinations are on the list. In addition, we will attempt to have at least two of the primer combinations have an estimated product size of less than 1000 bases. Finishers may find it helpful to draw the region and the relative positions of the suggested primers to assist in picking the best possible sets of primer pairs. If you cannot find two sets of primer pairs that will produce a product size of less than 1000 bases, then continue to screen the primers on the list and attempt to keep the PCR products as small as possible.

In this instance there are only 8 pairs in which the distance between the primers is less than 1000 bases. Furthermore, the 8 primer pairs consist of only two unique forward primers: 20034-20055 and 19588-19617.



Similarly, there are three unique reverse primers among the first 8 primer pairs suggested by *Consed*: 20543-20568, 20871-20890 and 20950-20969. Note that primers that have substantial overlap with each other (e.g. 20871-20890 versus 20871-20891) are not considered to be "unique" in this context.

Because there are only two unique left hand primers and we want to design two sets of primer pairs, we will pick one pair with the left hand primer at 20034-20055 (i.e. pairs 1-6) and the other pair with the left hand primer at 19588-19617 (i.e. pairs 7-8). Inspection of the primer coordinates shows that pairs 7 and 8 are only trivially different on the right primer; to choose between pairs 7 and 8 we will pick the pair with the smallest difference in melting temperature (Tm) between the left and right primers (i.e. pair 8).

Based on our search criteria, the second left hand PCR primer must be one of the first 6 primer pairs suggested by Consed. We can eliminate the first two primer pairs because they have the "same" right hand primer as primer pair 8. This leaves pairs 3-6 which have 2 possible right hand primers at 20871-20890 and 20950-20969, respectively.

To determine which right hand primer we should use, we can check to see if either of these two primers is listed in combination with our left hand primer in pair 8. Note that pair 13 and 14 has the left hand primer at 19588-19617 and the right hand primer at 20871-20890; this would indicate that primer pairs 3 and 4 are better than primer pairs 5 and 6. Because the right hand primer for the primer pairs 3 and 4 are essentially the same, we will again pick the primer pair based on the closest Tm, which would lead us to pick the primer pair 4 over primer pair 3.

Collectively, we have selected two left hand primers (20034-20055 and 19588-19617) and two right hand primers (20871-20891 and 20544-20568) in which all four pairwise combinations are found on the list (pairs 1, 4, 8 and 13). In addition, 3 of the 4 combinations produce PCR products that are less than 1000 bp in size.

Select one of the primer pairs and click "Accept Pair". Repeat the primer design procedure to regenerate the primer pair list and select the second primer pair. After accepting both primer pairs, save the assembly (`scf7180000301495_190000_290000.ace.12`).

During the primer selection process, if you find multiple primers that you cannot decide between, you can perform a BLASTN search against the *D. biarmipes* whole genome assembly to screen for off-target priming. Using those BLASTN results, you can select the primer that minimizes the probability of off-target priming. See the "PCR Primer Selection Guide" for detailed description of the search protocol.

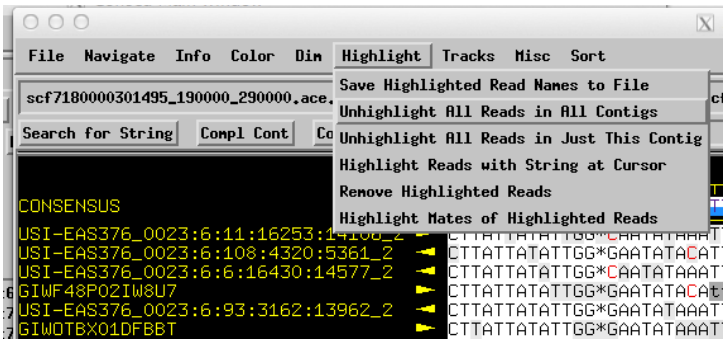
PCR/Sanger for Low Consensus Quality

For the *D. biarmipes* projects, almost all the low consensus quality (LCQ) regions will be associated with either the ends of the project or located adjacent to a gap. As described above, the finisher can ignore LCQ regions within the first and last 2.5kb of the project. In addition, improving a gap will also improve any surrounding LCQ regions. Because of the high read coverage from 454 and Illumina reads, we expect that genuine LCQ regions will be extremely rare. Furthermore, many of the LCQ regions may be of acceptable quality and could be resolved by manual inspection and editing. If you find any true LCQ regions, you should seek assistance from GEP staff members to determine if PCR/Sanger is necessary. If the region is approved for PCR/Sanger, you can apply the same rapid protocol as above for ordering primers for gaps (i.e. order 4 primers such that both forward primers are compatible with both reverse primers).

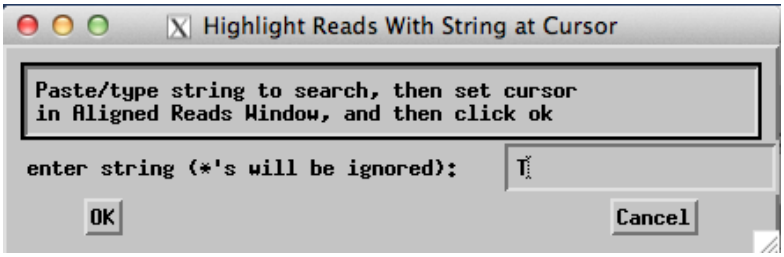
(Optional analysis) Identifying putative polymorphisms

An optional objective for these projects is to identify and tag regions with putative polymorphisms. Preliminary analysis suggests that, at least for *D. biarmipes*, the frequency of polymorphisms is extremely low. In general, we expect the two polymorphic sequences to be represented in at least 40% of all the reads. You can determine the frequency of each allele using the precentages shown in the Highly Discrepant Positions navigator or by using the Misc -> Depth of coverage at Cursor menu item.

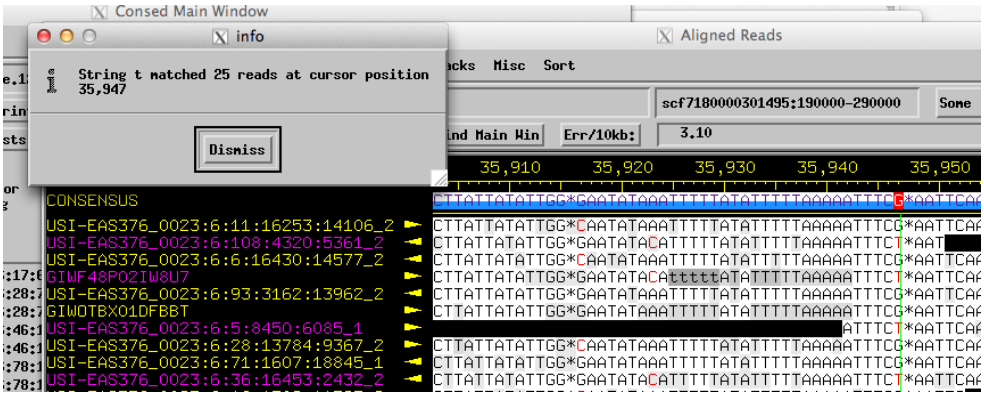
There is an alternate technique to count reads that may be more useful when looking for reads with sequences longer than a single base. We will use this technique to consider the reads that align to the consensus position 35,947. This position consists of a mix of G's and T's. To determine the number of reads that have a T at this position requires a two-step procedure. First, unhighlight all the reads in the project by selecting the "Highlight -> Unhighlight All Reads in All Contigs" option.



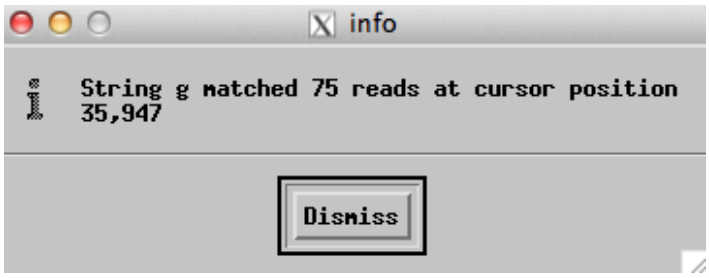
Once all highlights have been removed, select Highlight -> Highlight Reads with String at Cursor. This will open a dialog box where we can enter the allele we would like to search. In this example, enter a T in the search box and click "Ok". Sequence of any length can be entered into this box. Any read that matches the sequence starting at the position of the cursor will be highlighted.



In this case all the read names with a T at that position will be highlighted and you will get an info box stating the number of reads (in this case 25) that matched the T.



Repeat the two-step technique described above to count the number of G's at this position.



Our analysis shows that, of the 100 reads at this position, 25 reads (25%) has a T while 75 reads (75%) has a G. Given this result, this location does not satisfy the minimum 40% and we would not add a “polymorphism” tag to this location.