Finishing *Drosophila Grimshawi* Fosmid Clone DGA19A15

Matthew Kwong
Bio4342
Professor Elgin
February 23, 2010

Abstract

The primary aim of the Bio4342 course, Research Explorations and Genomics, is to finish and annotate the genomes of the selected regions of the various *Drosophila* species. After having successfully finished and annotated key regions of *D. virilis* and *D. mojavensis*, many Bio4342 students have moved to work on the dot chromosome (chromosome 4) of *D. grimshawi*. The dot chromosome is of particular interest because it exhibits both heterochromatic and euchromatic properties. Comparison of these finished and annotated species with *D. melanogaster* has shown that the fourth chromosome can contain genes found in euchromatic sites on other chromosomes in other species. This paper will discuss the finishing work on fosmid clone DGA19A15 of *D. grimshawi*.

Workflow

Initial Assembly:

Figure 1 shows the initial assembly view of project DGA19A15. There are three contigs: 4, 3c, and 2. Contigs 4 and 3c seem to be spanned by multiple forward and reverse pairs and appear to be portions of a single larger contig. After running crossmatch, it appears that there is a repititious region found between Contig 3c and 2. Contig 3c is inverted with respect to the rest of the assembly and I chose to investigate this abnormality first.

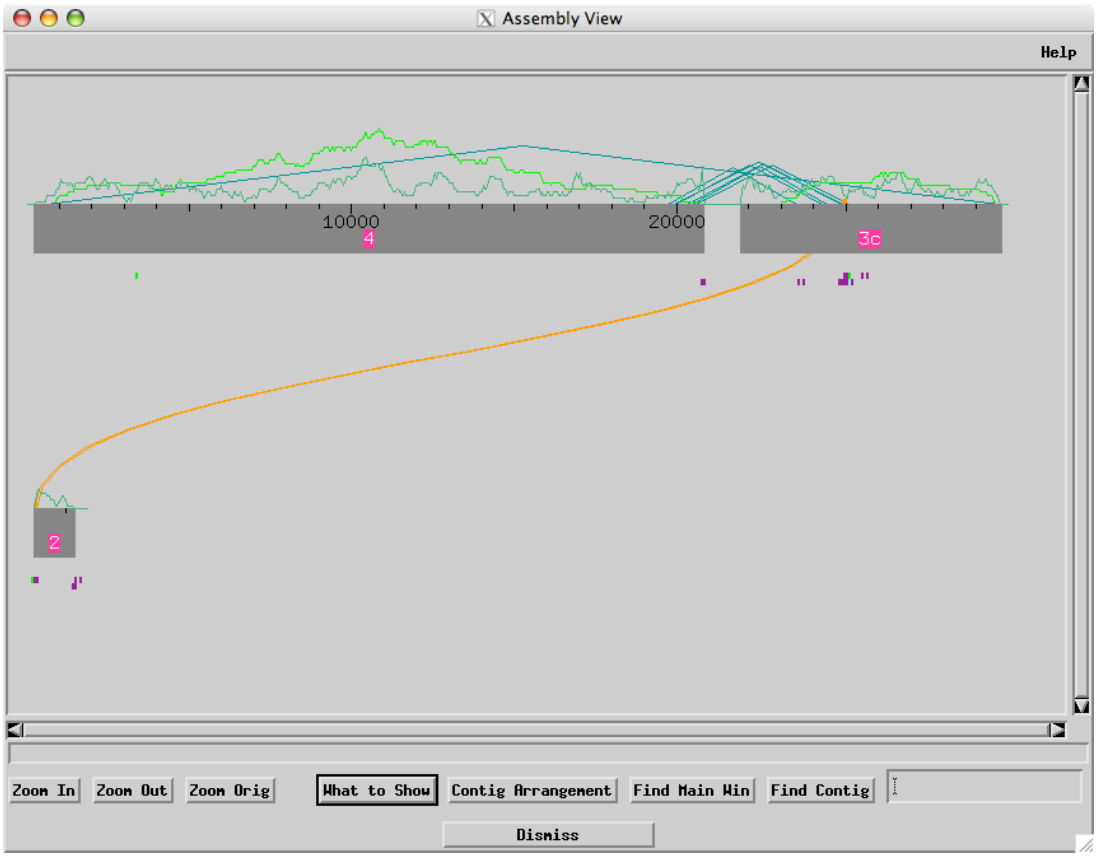

Fig 1: Initial assembly view of project DGA19A15.

Contig 3c and tagging ends:

To decide whether reorienting contig 3c would be the proper course of action, I searched for the assembly ends. Knowing that the assembly ends had reads with my project name, DGAA-A19A15a.?1, I searched for the reads with this criterion. I found that the left ends of both contig 3c and contig 4 contained this read name and concluded both must be the assembly ends (see Figures 2 and Figure 3). In addition, finding that the assembly ends were on the left ends of both contigs was evidence that Contig 3c was oriented incorrectly and should be inverted to attain proper orientation with the rest of the assembly. Figure 4 shows the properly oriented contig 3 in the assembly view. Cloning ends were identified and tagged (Left: 39 bp, Right: 8402 bp [28228 bp after join]).

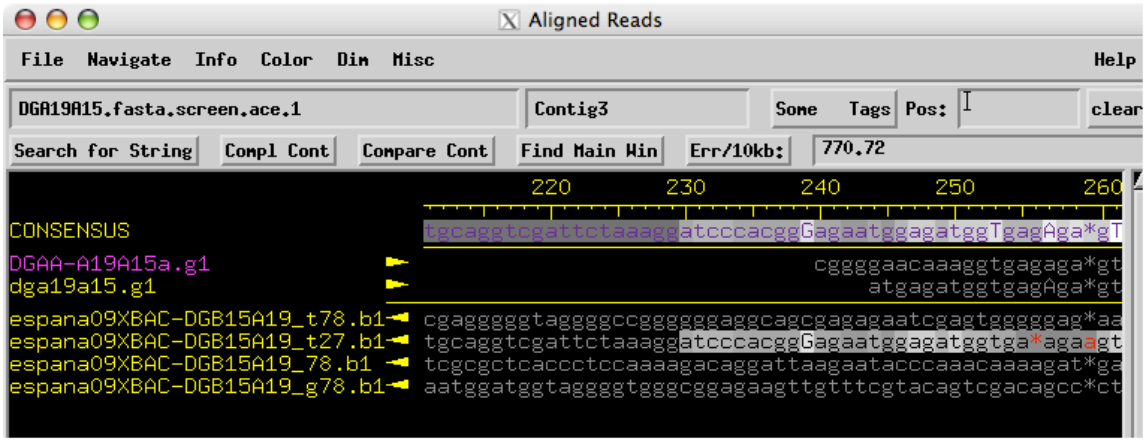

Fig 2: The left end of contig 3c containing read DGAA-A19A15a.g1, which is indicative of an assembly end.

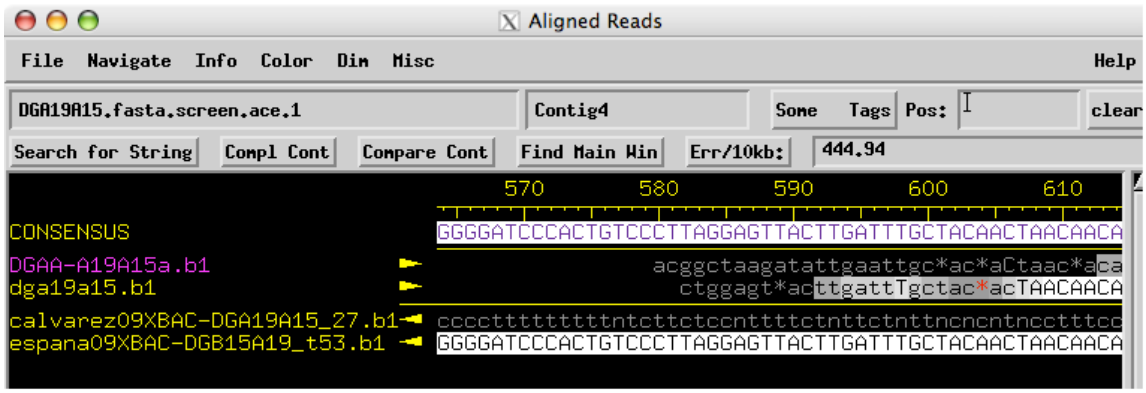

Fig 3: The left end of contig 4 containing read DGAA-A19A15a.b1, which is indicative of an assembly end.

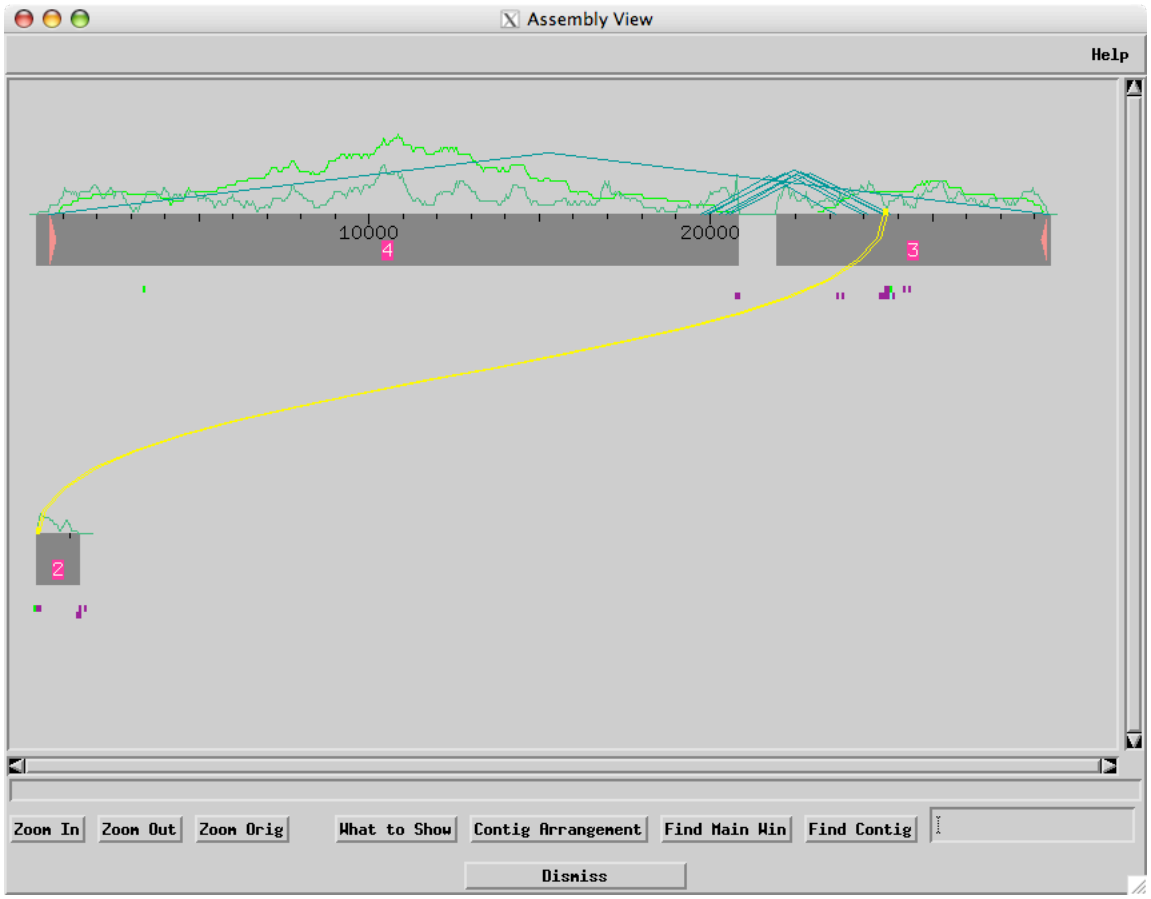

Fig 4: Shows contig 3c oriented in line with rest of assembly and renamed contig 3. Cloning ends are identified and tagged.

Contig 2:

Contig 2 does not appear to have a relationship with contigs 3 and 4 except for a single identical region that is similar between contigs 2 and 3. Thus, I had to decide whether contig 2 should be a part of contig 3 or not part of the assembly at all. To investigate the situation, I first compared contigs 2 and 3 to observe the contig similarities. Using "Compare Contigs" and aligning contigs 2 and 3, I could see that the contigs were almost entirely discrepant outside this similar region (Figure 5). Furthermore, Consed had tagged the similar region as a "repeat region" (named specifically "IS5#Artefact") on contig 2. This indicates that the similar region between contigs 2 and 3 is poor evidence for arguing that contig 2 belongs in contig 3. Both these facts supported the idea that contig 2 is not part of contig 3.

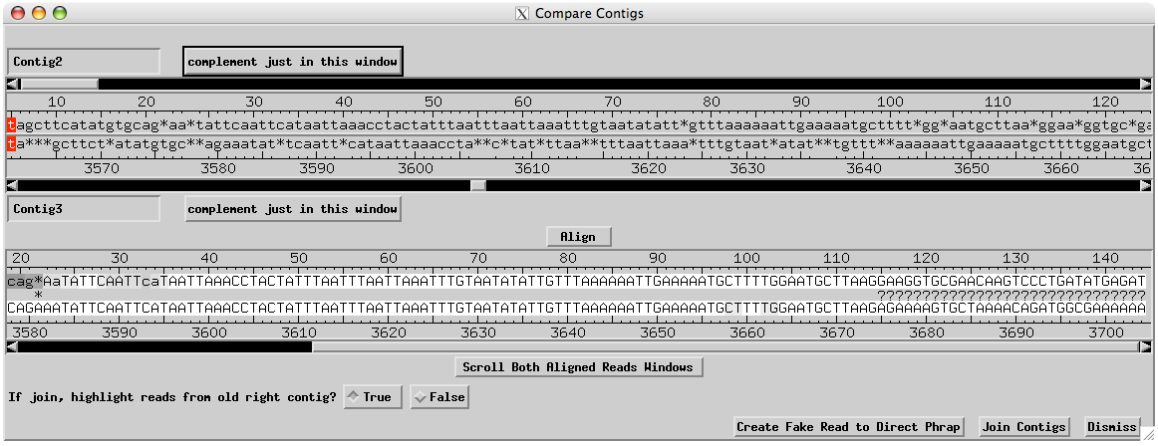

Fig 5: Compare Contigs show that there is only one region of similar bases between contigs 2 and 3.

After finding evidence that contig 2 was not a part of contig 3, I decided to look at the restriction digests to determine whether contig 2 belonged in the assembly and whether contigs 3 and 4 should be joined. By comparing real with *in silico* digests, I obtained useful data from the EcoRI, EcoRV, and HindIII restriction digests. Acting on my previous evidence that contig 2 was not a part of my assembly, I set the parameters for the *in silico* digests to include only contigs 3 and 4. Figure 6 shows the EcoRI digest, which illustrates a similar situation in the EcoRV and Hind III digests. The EcoRI digest shown is further evidence that contig 2 does not belong in the assembly. If contig 2 were a part of the assembly, the real restriction digest would show a 1.5 kb (the size of contig 2) fragment greater than the *in silico* digest. However, not only is this 1.5 kb value absent, the *in silico* digest shows a 1.5kb fragment that is unaccounted for in the real fragment digest. One possible explanation for the increased *in silico* fragment size is an overlap region between contigs (likely 3 and 4) that has been laid out as separate fragments. This is evidence that contigs 3 and 4 should be joined together. Further evidence that contigs 3 and 4 should be joined exists in the 9238 bp *in silico* fragment that consists of portions of both contig 3 and 4. Thus, the EcoRI digest demonstrates not only that contig 2 does not belong in the assembly, but also serves as evidence that contigs 3 and 4 should be joined.

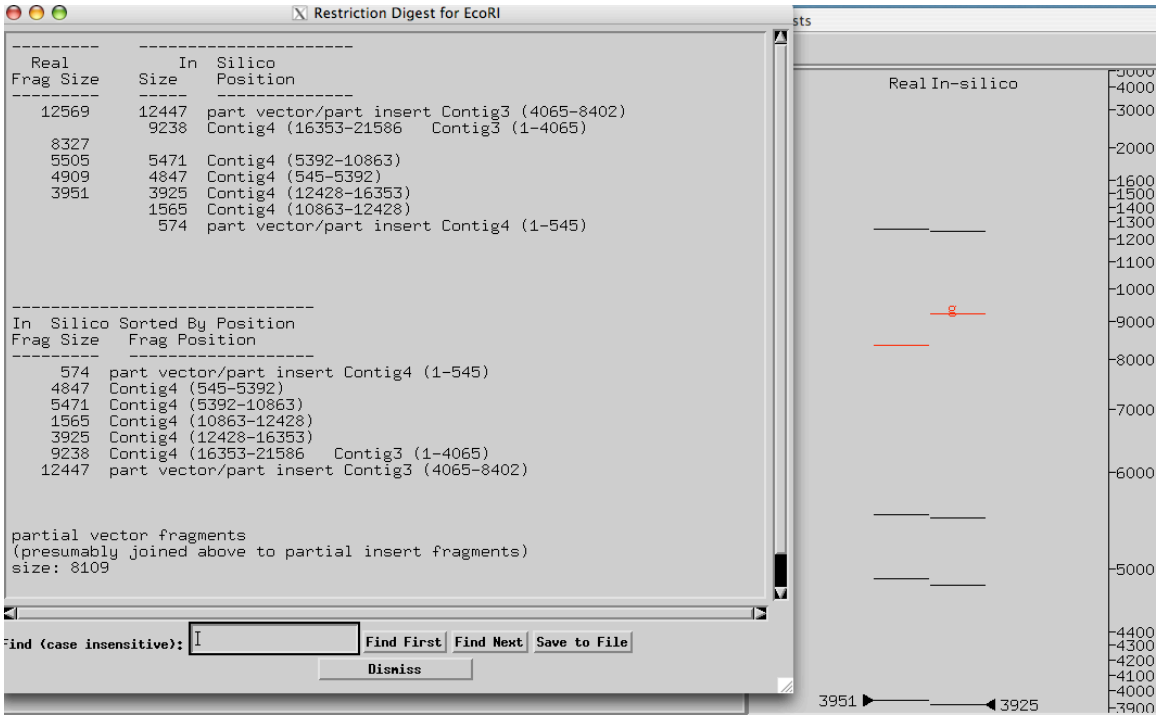

Fig 6: EcoRI digest that includes contigs 3 and 4 in the in silico digest.

Contigs 3 and 4:

From the restriction digests, I have strong evidence that contig 2 does not belong in the assembly and that contigs 3 and 4 should be joined. I decided against immediately forcing a join between contigs 3 and 4 because of the high inconsistency of reads and low Phred values for the consensus sequence that existed on the left end of contig 3 and right end of contig 4 (Figure 7).

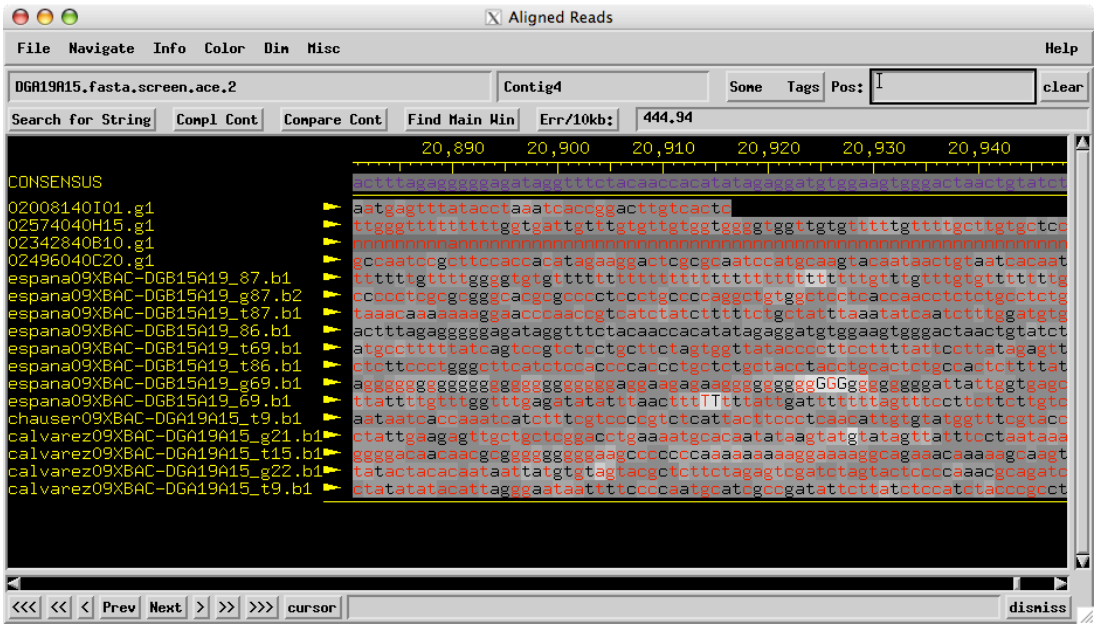

Fig 7: Large amounts of low quality reads at both the left end of contig 3 and right end of contig 4 indicates that more information may be necessary before joining the two contigs. It also may be the source of the 1.5 Kb excess of *in silico* bases over real digest bases.

Calling Reactions using Oligo Primers:

In order to attain better quality information about the contig ends and span the gap between contigs 3 and 4, I decided to call oligo reactions between these two contigs. If the reactions ran successfully, I would have the information necessary to create the join. Looking at the ends of contigs 3 and 4 (see Figure 7), I could see that a large number of reactions had previously been called in an attempt to close this gap. Thus, to ensure success, I decided to call 3 oligos and use all three reaction chemistries available (Big Dye, dGTP, 4:1) for each. I chose one oligo on contig 4 pointing in the rightward direction towards the gap and two oligos on contig 3 in the leftward direction towards the gap. I called only one reaction on contig 4 because it appears that multiple reactions have been previously called that were unsuccessful at spanning the gap. Thus, I believed it would be more efficient to attack the gap with more oligos from contig 3 and chose to call two oligos on this contig. I did not call reactions for contig 2 because sufficient evidence pointed to it not belonging to my assembly. Furthermore, I would run a BLAST search on it to verify this conjecture. Since the labeled ends of contigs 3 and 4 connected with another project that likely had a higher quality consensus, I did not order reactions off of these ends. Figure 8 shows the reaction name, reaction chemistry, and oligo sequence of the oligos called.

#	Reaction Name	Rx Chemistry	Oligo Name	Oligo Sequence	Created Date
1	XBAC-DGA19A15_5.b1	BigDye	DGA19A15.5	cacattttcctgcaacaatc	2010-02-03
2	XBAC-DGA19A15_g5.b1	dGTP	DGA19A15.5	cacattttcctgcaacaatc	2010-02-03
3	XBAC-DGA19A15_t5.b1	4:1	DGA19A15.5	cacattttcctgcaacaatc	2010-02-03
4	XBAC-DGA19A15_4.b1	BigDye	DGA19A15.4	cacgaggtcgaatgctg	2010-02-03
5	XBAC-DGA19A15_g4.b1	dGTP	DGA19A15.4	cacgaggtcgaatgctg	2010-02-03
6	XBAC-DGA19A15_t4.b1	4:1	DGA19A15.4	cacgaggtcgaatgctg	2010-02-03
7	XBAC-DGA19A15_6.b1	BigDye	DGA19A15.6	gaaaggccctcaataggac	2010-02-03
8	XBAC-DGA19A15_g6.b1	dGTP	DGA19A15.6	gaaaggccctcaataggac	2010-02-03
9	XBAC-DGA19A15_t6.b1	4:1	DGA19A15.6	gaaaggccctcaataggac	2010-02-03

Fig 8: Oligo reactions called. Numbers 1-6 correspond to contig 3 and numbers 7-9 correspond to contig 4 (pp11-14 DGA19A15 Manual).

Comparing Called Reactions to Autofinish:

Autofinish called reads in all three contigs:

Contig 2- I concluded Contig 2 had a high probability of not belonging in my main assembly based on restriction digest data and crossmatch calling. Thus, I did not call oligos for contig 2. Autofinish called 3 oligos for contig 2 likely to ensure it does not belong to this assembly. Autofinish called 2 oligos to acquire higher base quality for the left end and 1 oligo to acquire higher base quality for the right end.

Contig 3- I concluded the left end of this contig should be joined to some degree with contig 4 both from viewing the restriction digest and from the number of matching forward and reverse pairs between the contigs. Thus, I called 2 oligos with Big Dye, dGTP, and 4:1 chemistries for this end. Autofinish called 2 oligos at the left end of the contig near the same regions that I called them.

Contig 4- I concluded the right end of this contig should be joined to some degree with contig 3 (based on the same reasoning as above). I called one oligo with Big Dye, dGTP, and 4:1 chemistry for this end. I called no more than one oligo from this end because the high number of unsuccessful reads called before indicate a low probability of acquiring useful data. Attacking the contig 3 end would likely be more efficient and effective. Autofinish called five oligos for contig 4, all for the right end of the contig. This is the same region where I called my single oligo. Autofinish likely called such a high number of reactions to obtain more high quality reads for the right end of contig 4 and because the region had proven to be problematic in the past.

Figures 9 and 10 compare the oligo reactions that I called to the ones that autofinish called. Figure 9 shows all the oligos that both Autofinish and I called, while Figure 10 compares the oligo locations and sequencing direction on the assembly view.

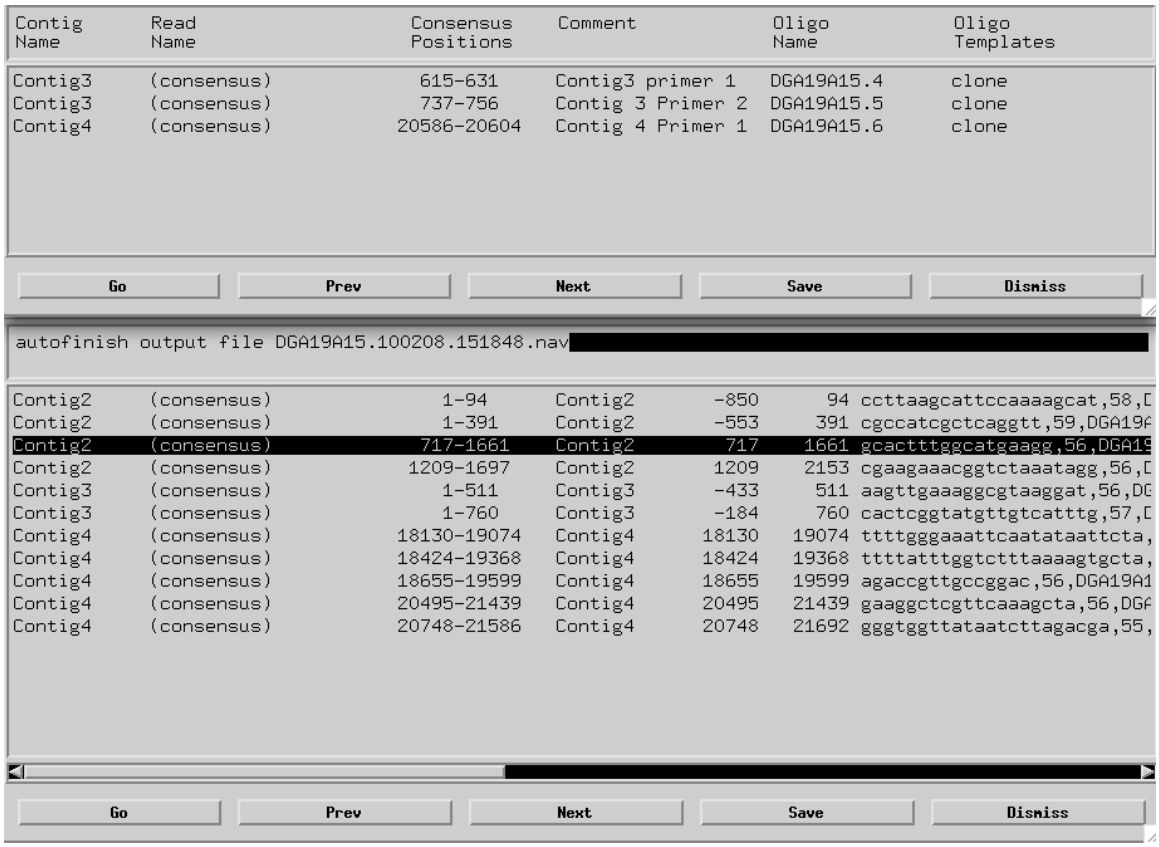

Fig 9: My calls (above) versus autofinish calls (below).

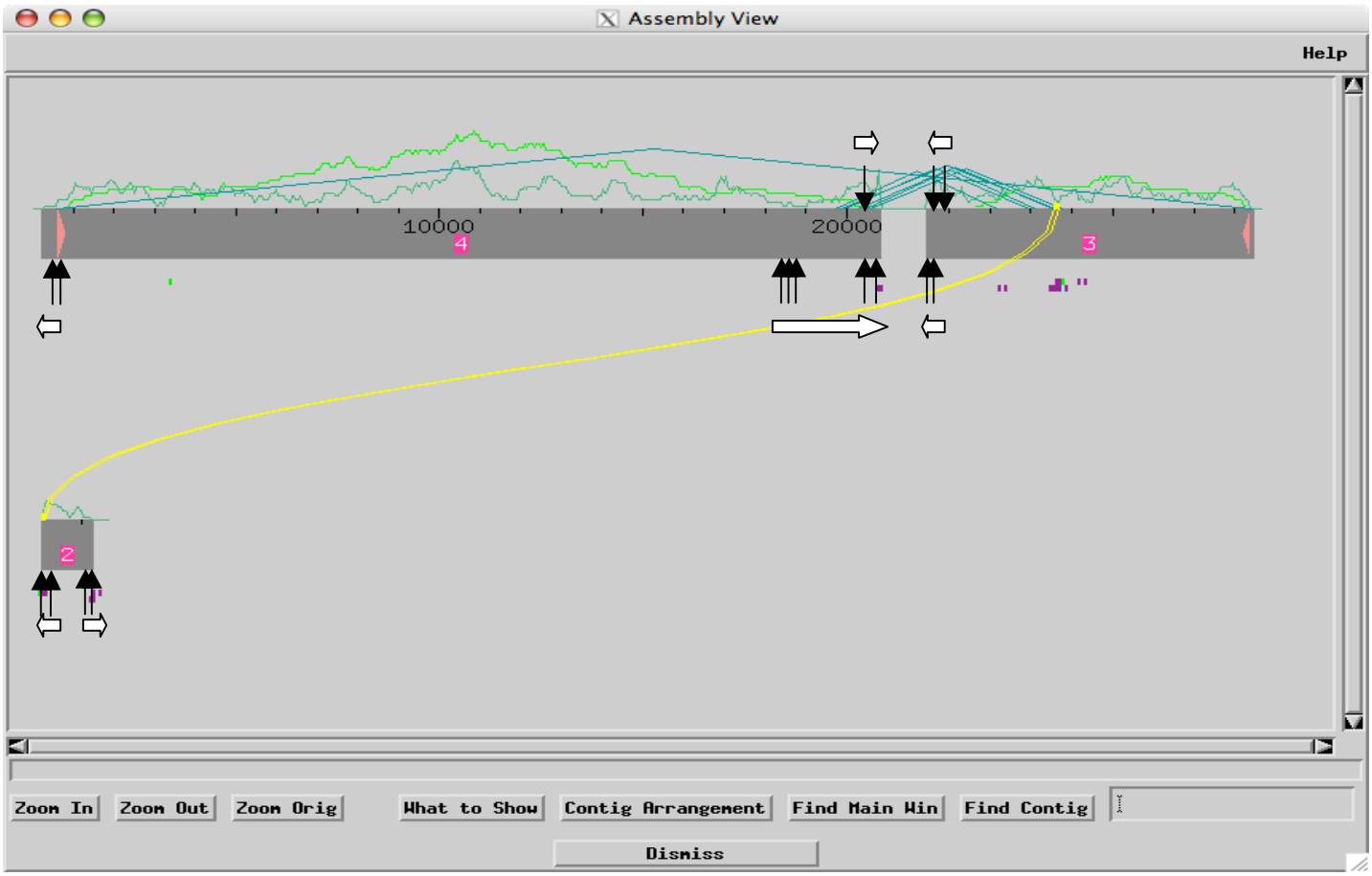

Fig 10: Assembly view consensus position comparison of autofinish called oligos (black arrows on bottom of contigs) to oligos that I called (black arrows on top of contigs). The white arrows indicate the direction each oligo or group of oligos would run and sequence.

Adding New Reads and Closing the Gap:

After receiving the reaction results, I acquired multiple new reads to add to my assembly as shown in Figure 11.

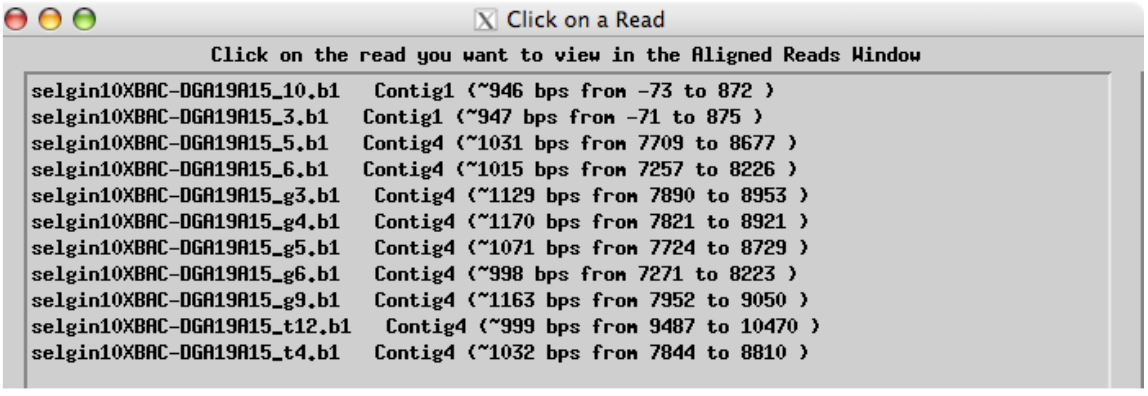

Fig 11: Reads obtained from running oligo reactions (contig 1 was not considered because it does not add useful data to the assembly).

After using Phred/Phrap to incorporate these new reads into the assembly, I found that the additional reads provided enough useful data for consed to acknowledge a closed gap (Figure 12).

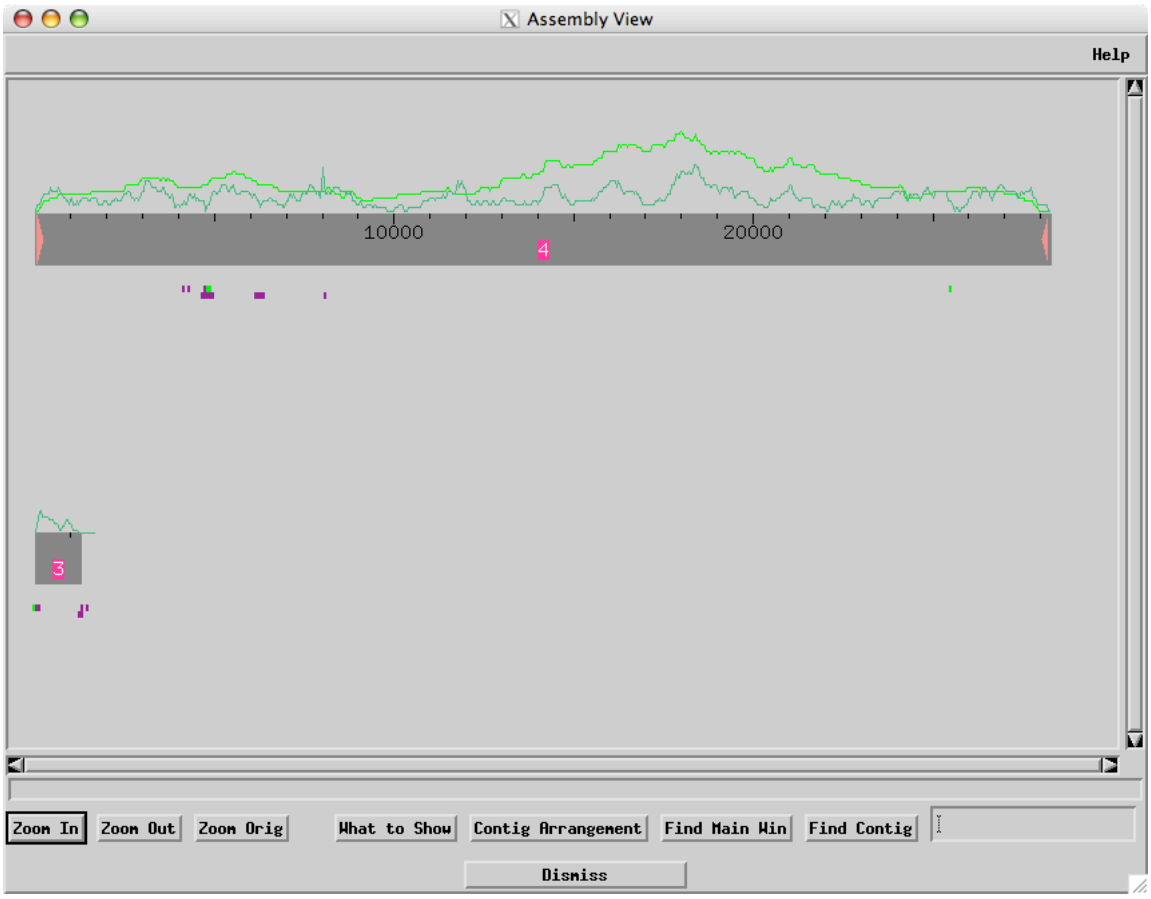

Fig 12: Assembly view after gap is closed.

Note that the previously named contigs 3 and 4 have merged to become contig 4 and the previously named contig 2 is now named contig 3. The high quality discrepancy and unaligned high quality region sections refer to the contigs before the merging and addition of new reads. Only one round of reactions was necessary for this project.

Only High Quality Discrepancies, Low Quality Discrepancies, and Unaligned High Quality Regions inside 2 kb of contig ends in contigs 3 and 4 (before merger) and contig 4 (after merger) were considered. This is because regions near the contig ends are generally lower quality in my project and likely to be higher quality after assembling adjacent projects. Thus, those assembling adjacent projects are better equipped to handle issues pertaining to the assembly ends. Contig 2 (contig 3 after merger and read addition) was not considered because there is ample evidence that it is contamination to my project and not part of my main assembly.

High Quality Discrepancies:

Most High Quality Discrepancies in this project were phred misreads that substituted a "pad" (*) for a base. A second type of High Quality Discrepancy includes a poor reaction that leads to a phred misread. These conclusions were reached after examining the traces of all High Quality Discrepancies. Furthermore, the absence of clusters of High Quality Discrepancies is evidence that these High Quality Discrepancies are not polymorphisms.

I solved the High Quality Discrepancies (HQ) in a similar manner at every instance and will show two examples indicative of the methods and reasoning used:

1. Contig 4, Consensus position 2491--Made HQ discrepancy Low Quality (LQ):
   I compared the discrepant read position 2491 that, according to the aligned reads window (Figure 13), contained a "pad" to 2 HQ reads that agreed with the consensus (Figure 14). The discrepant read trace appears to have an overly large signal for 2 T nucleotides compared to reads that agree with the consensus sequence. Comparing this signal with a previous TTT signal earlier in the read shows that the *TT discrepancy has a similar width signal size corresponding with the TTT signal (Figure 15). This is sufficient evidence to label read position 2491 as a LQ base t (Figure 16).

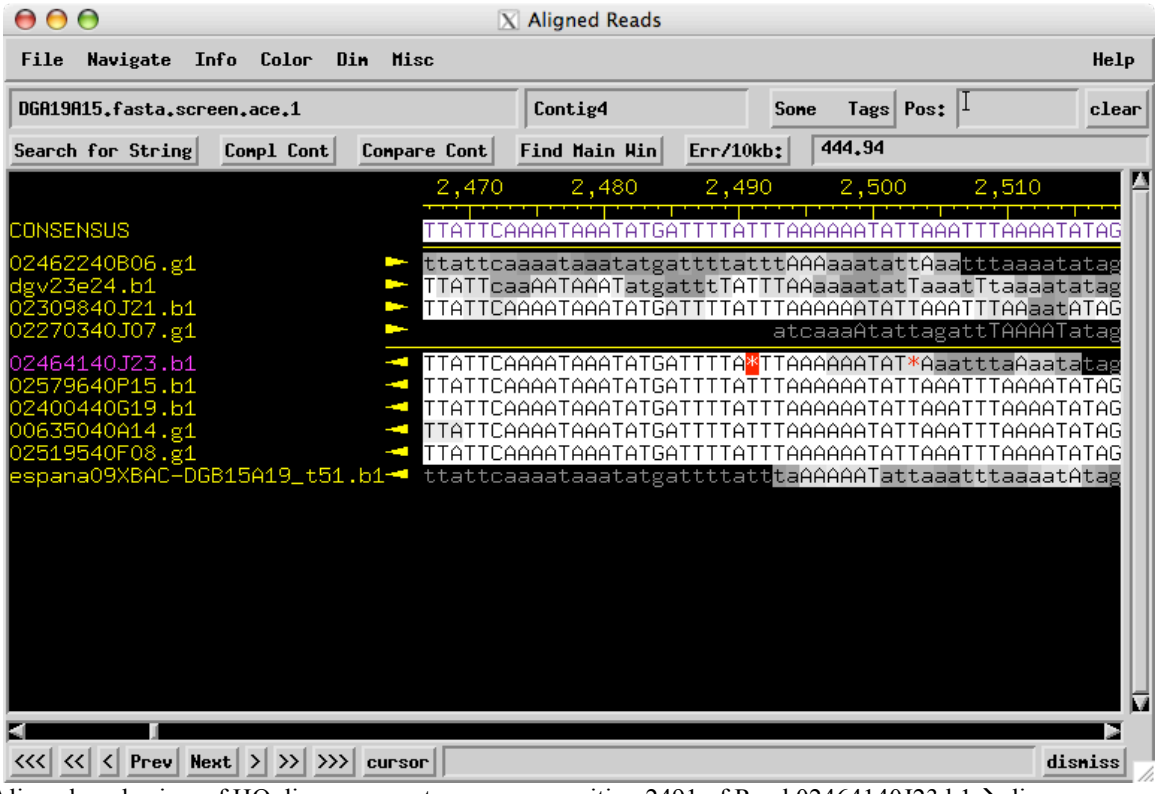

Fig 13: Aligned reads view of HQ discrepancy at consensus position 2491 of Read 02464140J23.b1→ discrepancy in form of "pad" (*).

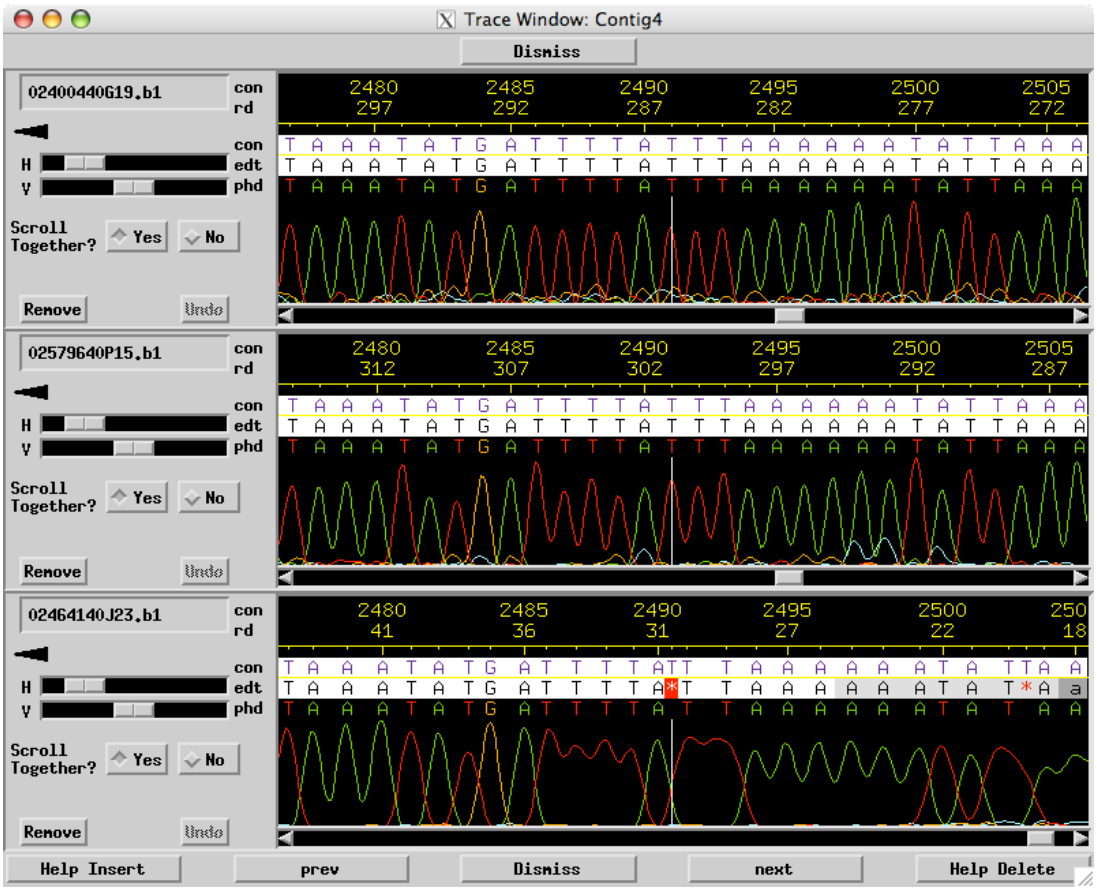

Fig 14: Bottom read has a discrepant base at position 2491 due to a phred miscall. Other reads that match the consensus have strong T traces at 2491 consensus location.

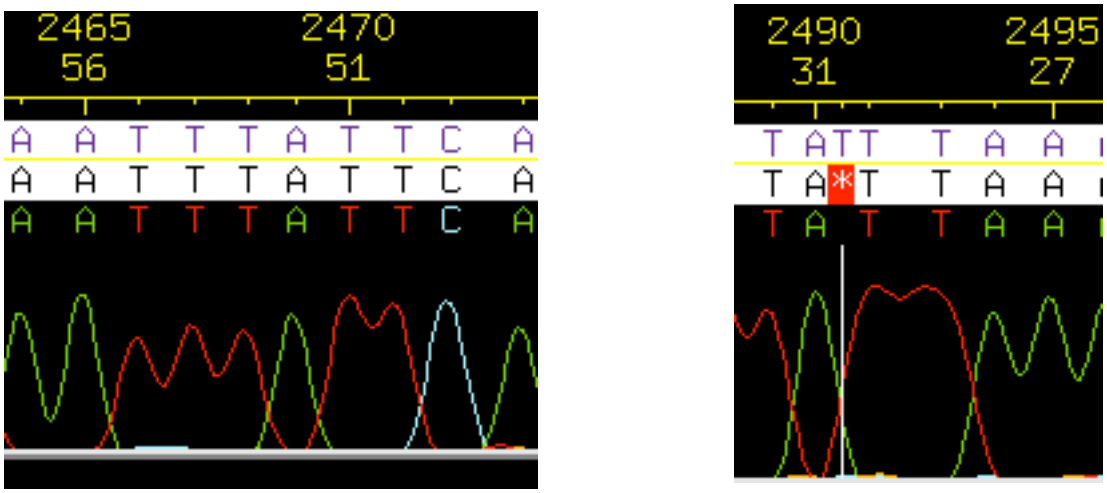

Fig 15: Comparing the signal of the discrepant *TT base (pictured right) with a TTT signal (pictured left) downstream of the discrepancy (on the same read) shows that both signals show a similar area. This supports the user call that the discrepant base position be marked as a low quality t base.

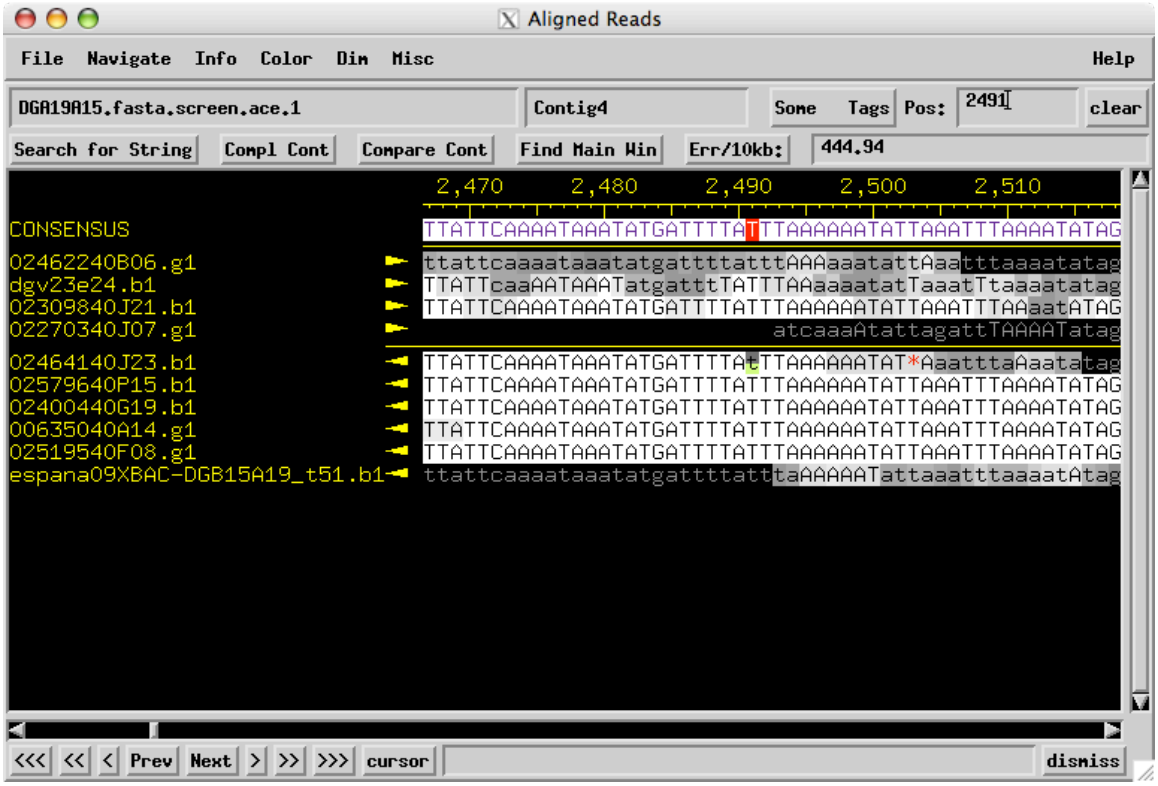

Fig 16: Aligned reads view of HQ discrepancy→LQ base t at consensus position 2491 of Read 02464140J23.b1. Consensus sequence still has sufficient information to have a phred score of 90.

Additional High Quality Discrepancies made low quality using this method and reasoning:
Contig 4-- Consensus positions: **6492, 7731, 20778**
Contig 4 (after merge+new read addition)—Consensus positions: **8067, 21115, 22355, 26354**

2. Contig 4, Consensus position 16778—Made HQ discrepancy LQ:
After looking at the aligned reads window (Figure 17), I could see that the HQ C base called by Phred was likely a miscall since it disagrees with the consensus and is surrounded by low quality bases. This conjecture was reinforced after looking at the trace window (Figure 18). The discrepant read showed an unusually large C signal for multiple bases that largely disagreed with the consensus. It is likely that the sequencing chemistry used to acquire this read malfunctioned at this section. Thus, it is logical to label this position on the read as low quality.

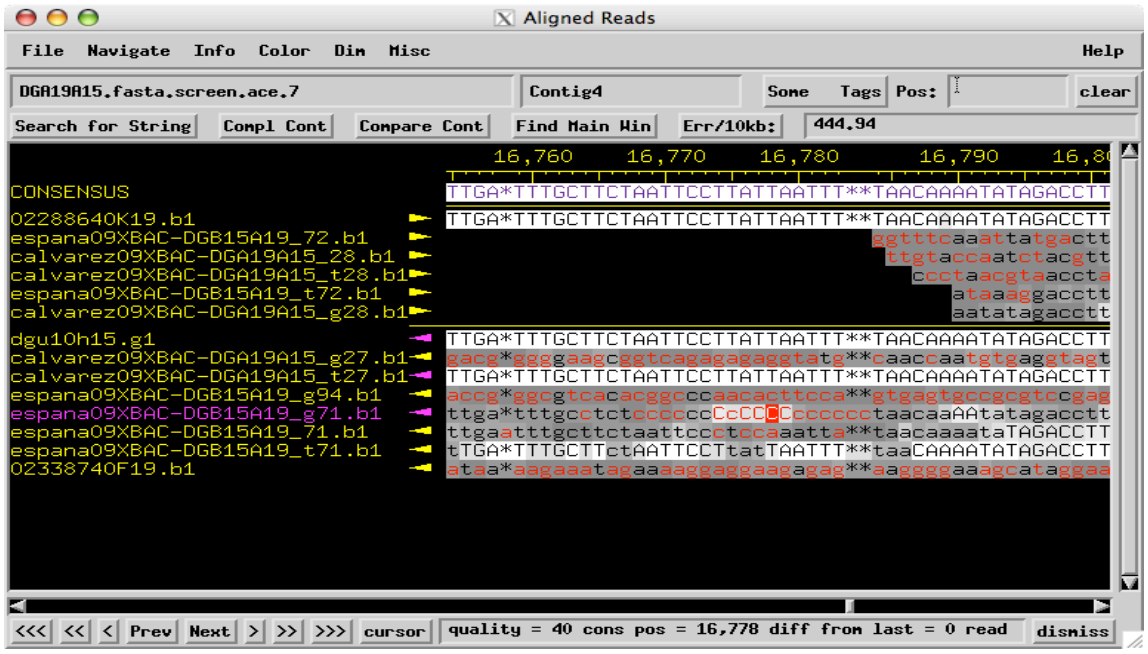

Fig 17: Aligned Reads Window shows discrepant HQ read with consensus at position 16778. Made low quality after viewing trace window. Although there are other discrepant LQ reads in this portion, additional reads do not need to be called because the consensus has a phred score of 90 supporting the T nucleotide.

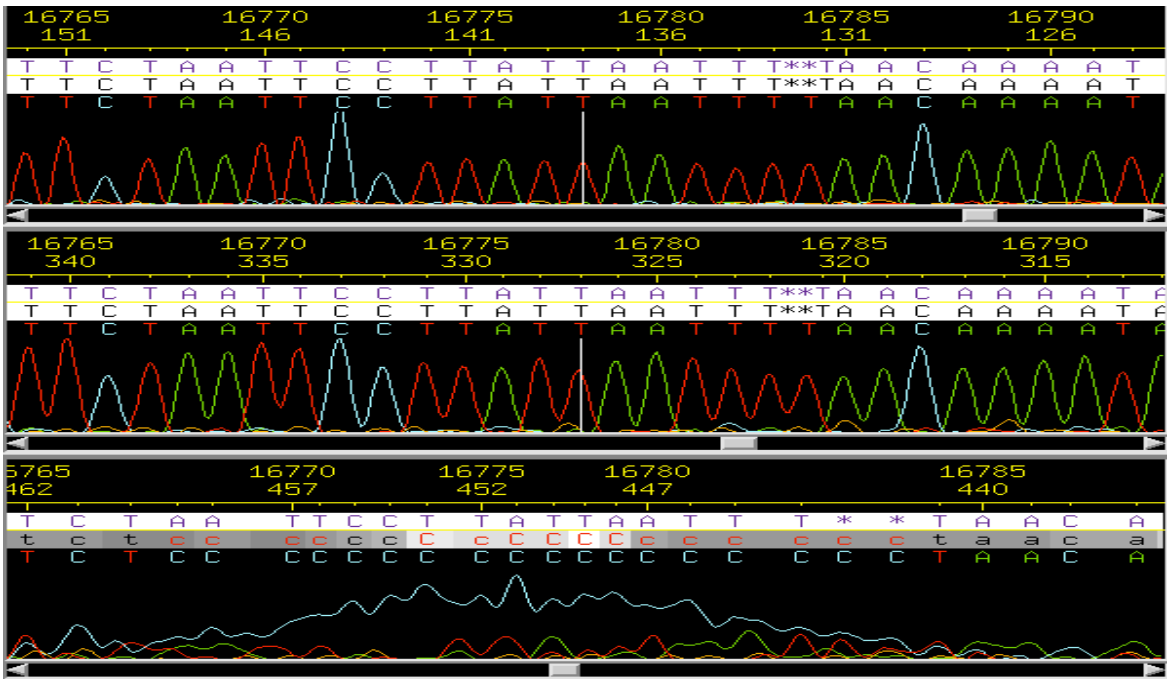

Fig 18: Discrepant read shows an unusually large C signal for multiple bases that seems to be overwhelming the other signals. Looking at these covered signals one can see that many agree with the consensus sequence. Thus, it is likely that phred miscalled the discrepant C base since there appears to be sequencing errors in this region. When comparing this signal to the two high quality traces above there is no question that the discrepant C is a miscall.

Unaligned High Quality Regions:

Since my project had ample coverage over most locations, Unaligned High Quality Regions did not pose an issue because reads that did not add useful data to the assembly could be removed into their own contigs. Thus, all unaligned high quality regions were solved the same way: looking to see whether the read provided useful data to the assembly (none did), and removing the read into a separate contig if it did not provide useful data (this occurred in all six cases).

Process used to solve Unaligned High Quality regions:

Contig 4 Consensus position 805-846: Looking at the aligned reads window (Figure 19), one can see that the read calvarez09XBAC-DGA19A15_27.B1 has unaligned high quality regions compared to the consensus. However, looking at the read as a whole, there is little similarity between the read and any portion of the consensus. Furthermore, no portion of this read provides necessary data to the consensus. Looking at the trace window (Figure 20), I noticed that the unaligned high quality region had poor signals compared to reads that corresponded to the consensus. This was sufficient evidence for me to put the read into a separate contig (contig 5).

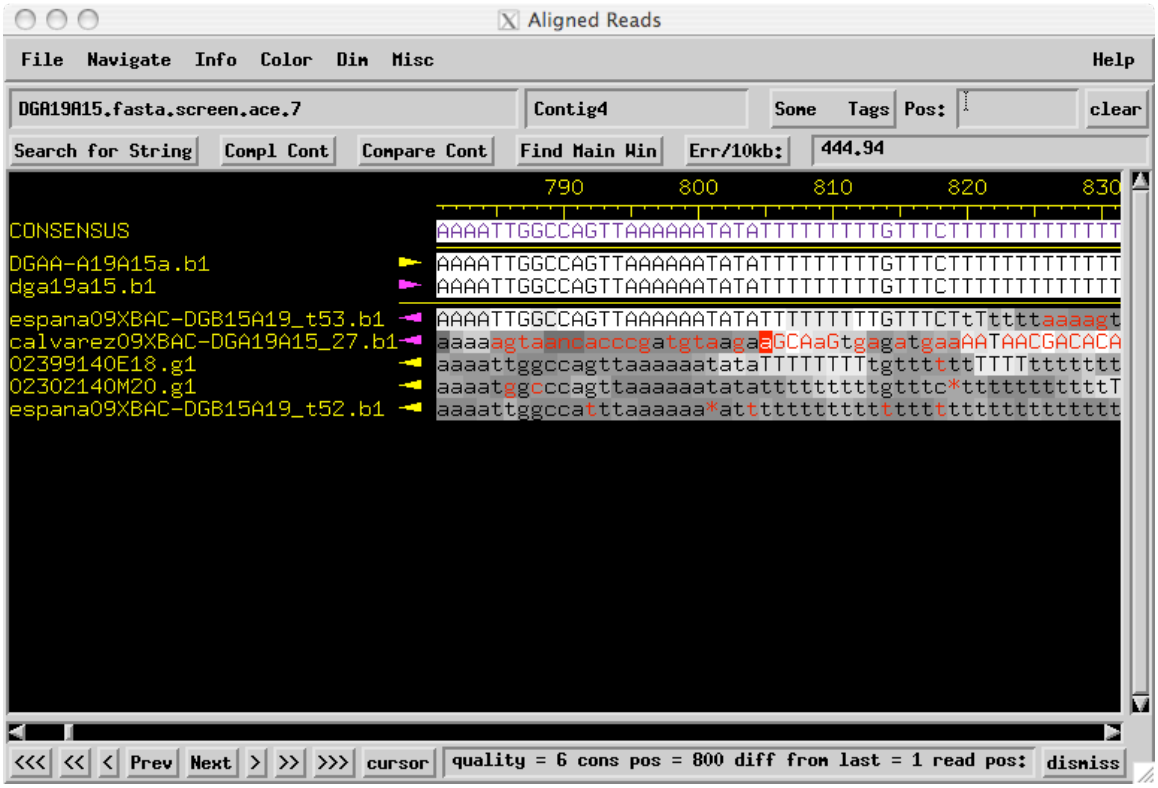

Fig 19: Aligned reads window shows unaligned high quality region found in contig 4, consensus position 805-846.

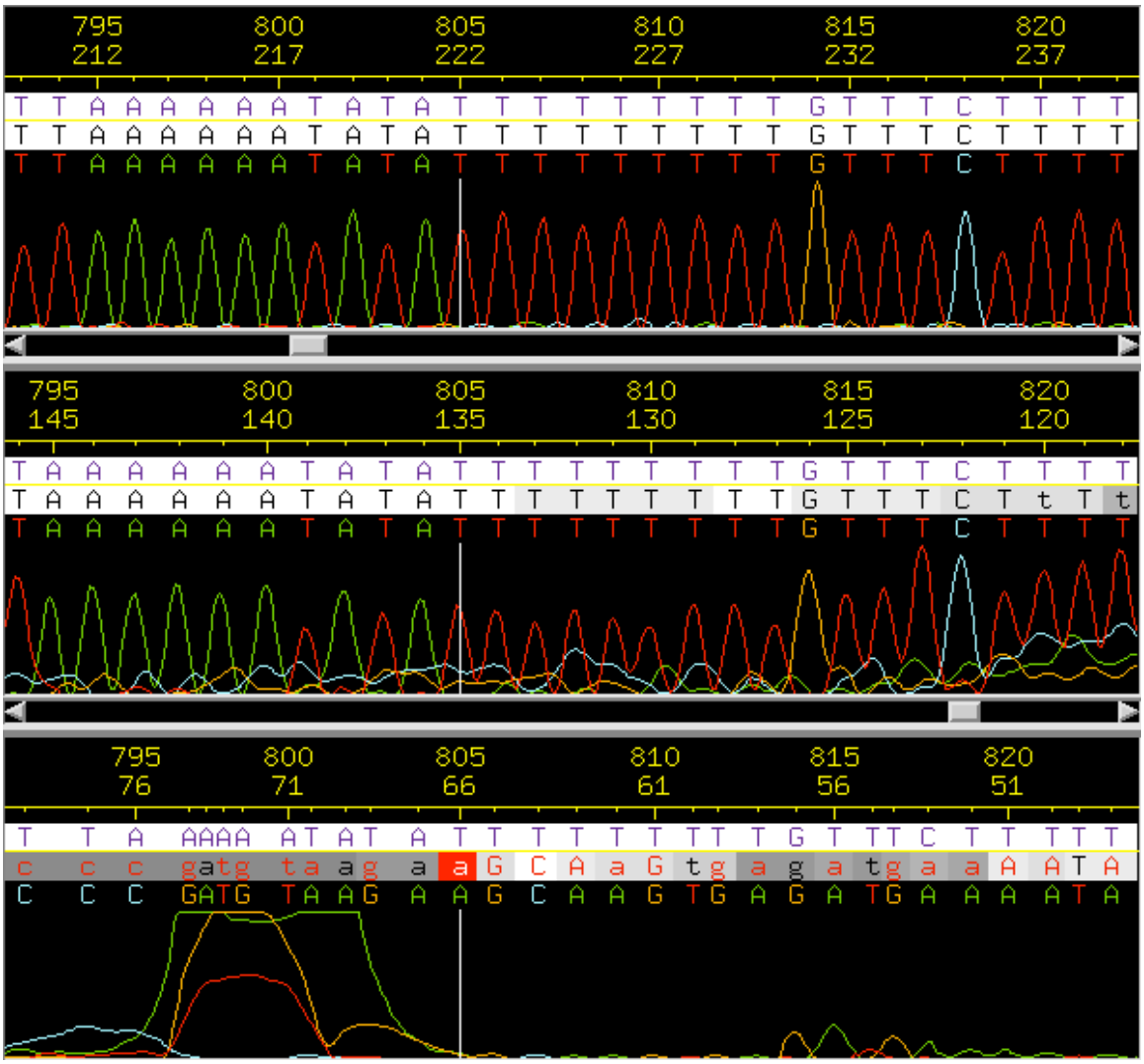

Fig 20: Trace window of contig 4 consensus positions 805-846 that has an unaligned high quality region. The read in question has a large number of bases discrepant with the consensus. Looking at the traces, one can see that the discrepant read has few signals coinciding with the consensus. Furthermore, the signals themselves lack the sharp signals in this region that lead to reliable data. The sharp signals that do lead to reliable data are shown for comparison in the two high quality traces above. Thus, there is evidence that the read in question is of low quality due to a malfunction during the sequencing reaction.

Additional Unaligned High Quality Regions that led to read removal into a separate contig using this method and reasoning:
Contig 3—Regions: 3382-3686 (read chauser09XBAC-DGA19A15_g12.b1 put into contig 6), 3749-3940 (read espana09XBAC-DGB15A19_105.b1 put into contig 7), 4700-4769 (read espana09XBAC-DGB15A19_t62.b1 put into contig 8)
Contig 4 (after merge+new read addition) —Regions: 3672-3741 (read espana09XBAC-DGBA19_t62.b1 put into contig 5), 4501-4692 (read espana09XBAC-DGB15A19_105.b1 put into contig 6),

Note: Many changes in contig names after merging of contigs 3 and 4 into contig 4 and new read additions.

Low Quality Consensus:

As shown in Figure 21, there was only one low quality consensus in the main assembly more than 2 kb from the contig ends. This occurred at 7999-8000 bp after contigs 3 and 4 joined into contig 4.

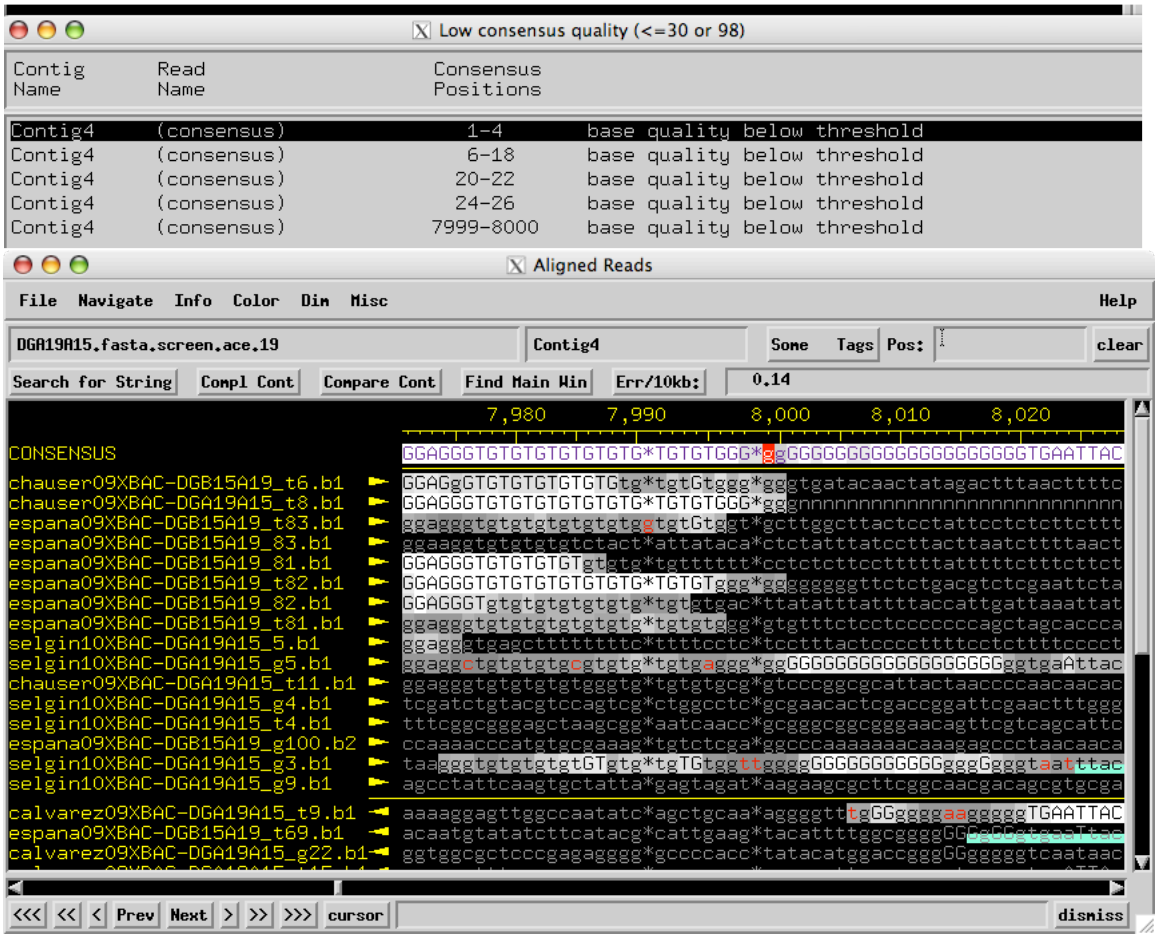

Fig. 21: Shows low consensus quality positions 7999-8000 in joined contig 4

Looking at the comparatively higher quality bases from reads in the 7999-8000 bp LQ region, I concluded that both low consensus quality g bases could be made high consensus quality. The average signal strength for G in both low quality bases is high and apparent in both reads shown in Figure 22. The second read was miscalled by phred likely because it is masked by the abnormally high T signal next to it. This T signal is likely an error from read generation because the other high quality reads do not show as strong a signal for this base. Thus, there is ample evidence for the low quality g in the 7999-8000 bp regions to be changed into a high quality G. After making the change, I tagged the consensus noting that this is a user-mediated change (Figure 23).

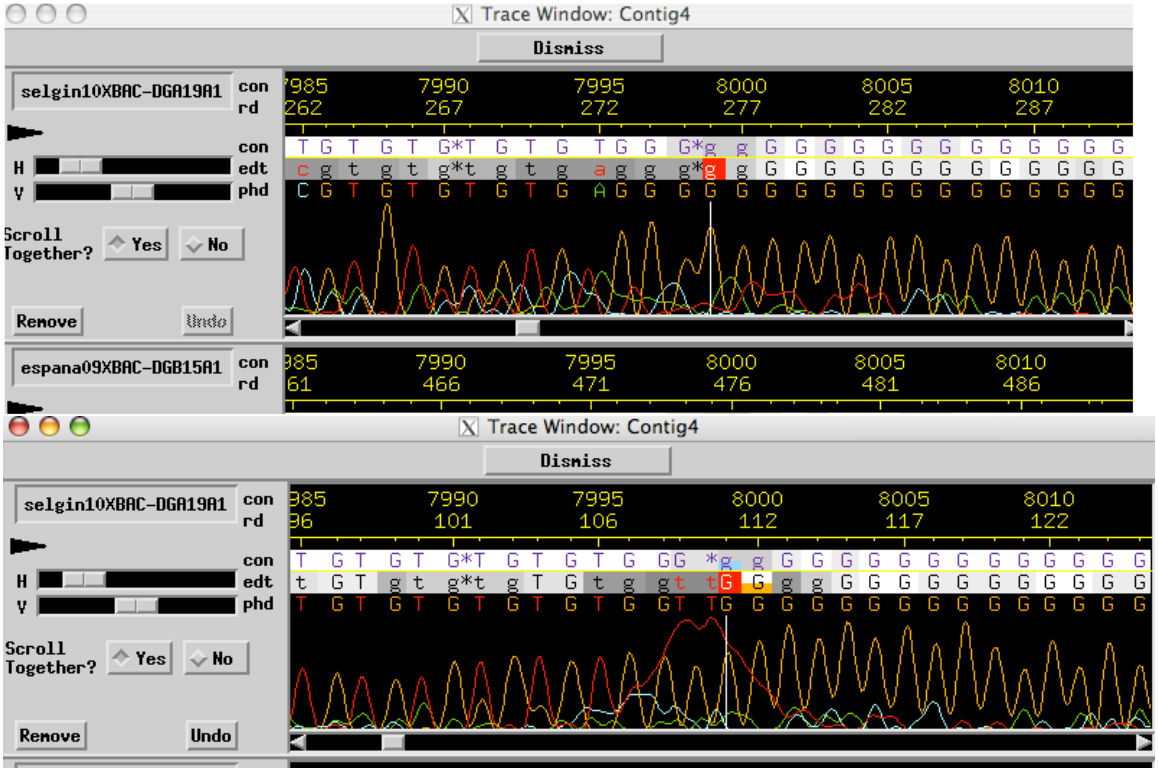

Fig 22: Traces that demonstrate strong peak signal and an active signal to the right and left of the 7999-8000 bp bases. Evidence to change consensus from LQ to HQ.

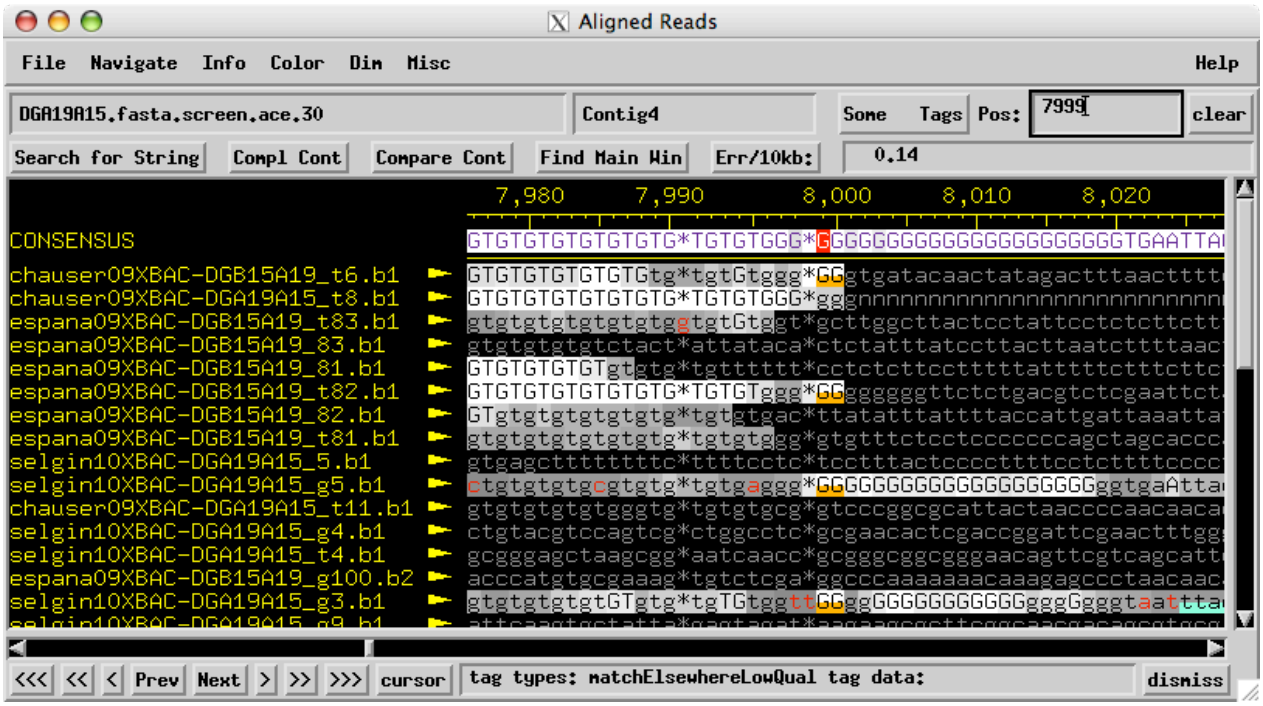

Fig 23: Aligned Reads view of 7999-8000 bp after the user-mediated LQ→HQ decision.

Single Strand/chem.

Sequences with single strand/chemistry issues inside 2 kb of the contig ends are above threshold quality so that additional reactions do not need to be considered. I placed comment tags at the beginning of the single strand/chemistry regions if they were 2 kb inside the assembly ends. Consed tagged the following regions for single strand/chem 2 kb inside assembly ends:

4000-4010 bp for single strand/chem→ 3× coverage: see below
7720-7973 bp for single strand/chem→ 9× coverage
9403-9536 bp for single strand/chem→2× coverage

Looking at the aligned reads window (Figure 24), the 4000-4010 bp region is most questionable regarding whether additional reads are necessary to improve the consensus confidence due to read quality. This region is further discussed below. On the other hand, the 7720-7973 bp region has 9× coverage, which is sufficient in number and quality to have a high confidence regarding consensus accuracy. Although the 9403-9536 bp region has only 2× coverage, the quality of the two reads is sufficient to generate a phred score of 90. Thus, while one could run oligos to generate reverse pairs, it is likely not necessary since confidence in consensus accuracy is high.

The 4000-4010 bp trace window (Figure 25) reveals that the low quality reads have relatively high quality signals. The reads all have visible signals that match with the consensus and with each other. Thus, I have confidence that the consensus is accurate and no additional oligo reactions need to be called for this region.

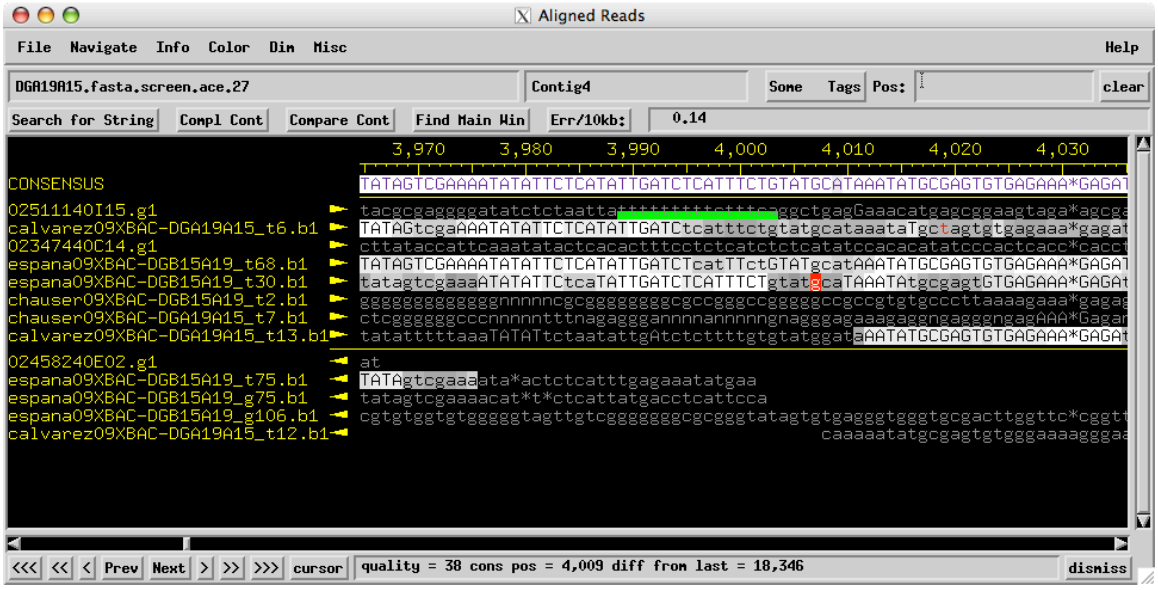

Fig 24: The aligned reads window of region 4000-4010 shows a lower quality of reads than usual (all are still above threshold quality phred 30 for single stranded regions) supporting the consensus.

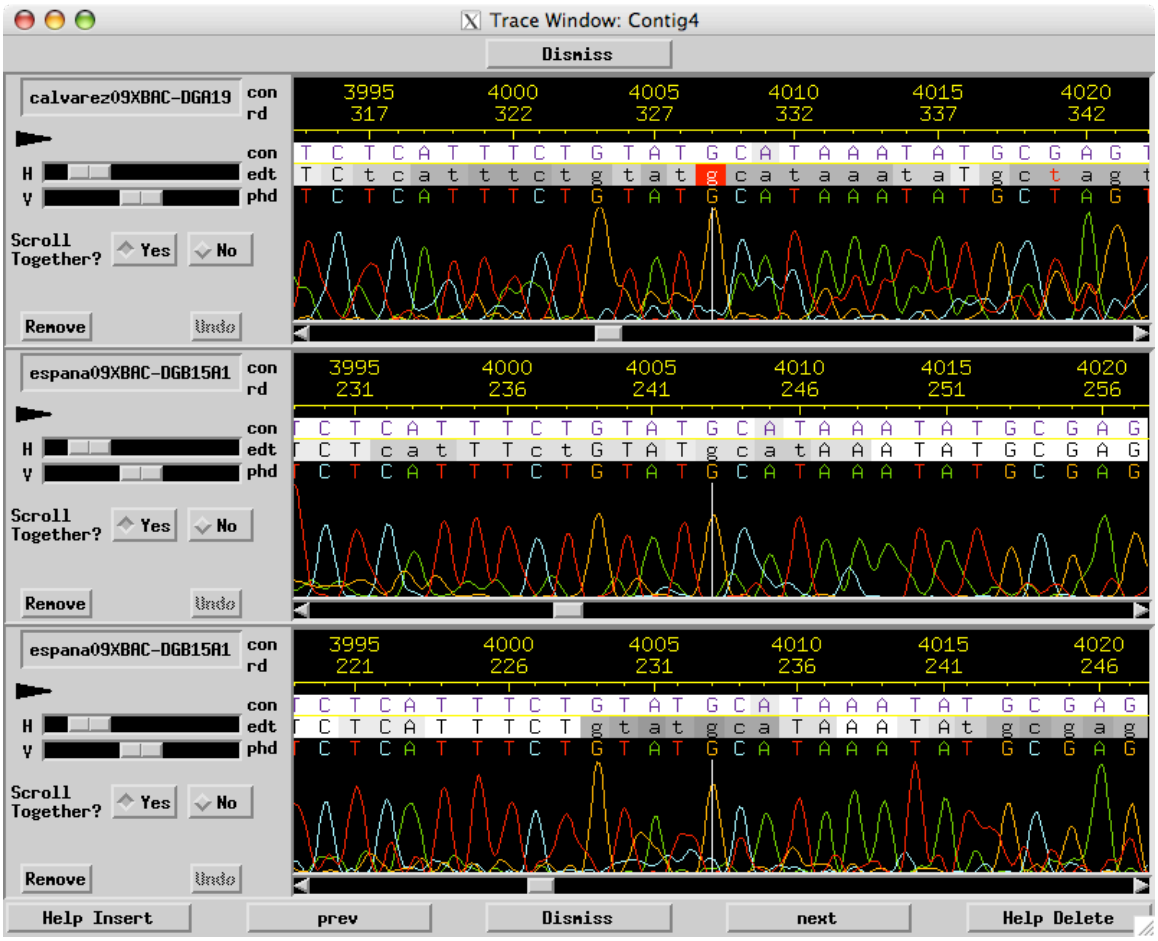

Fig 25: Trace window of region 4000-4010 that has single strand/chemistry issues. The traces all visibly support the consensus, thus no additional reactions are necessary.

BLAST searches and Comparison:

Two BLAST searches were run using the National Center for Biotechnology Information (NCBI) website maintained by the National Institutes of Health (NIH) (www.ncbi.nim.nih.gov/blast). I ran the first BLAST search on contig 2 (named contig 3 after additional reads were called) to ask whether it was a contamination in my assembly and thus not part of my project. The BLAST result indicated contig 2 had a high probability match with various fragments of the *E. coli* genome (Figures 26 and 27). Thus, I could conclude with confidence that contig 2 is contamination and not a part of my main assembly. I ran the second BLAST search to determine whether there were impurities in my main assembly contig 4 (named contigs 3 and 4 separately before additional reads were called). The BLAST result indicated that there was a significant probability that there was a small amount of *Wolbachia* DNA (a type of *Drosophila* parasite) in my main assembly (Figures 28 and 29). However, I concluded that this was not contamination (see discussion below) and that my main assembly could be considered high quality data.

Blast Search for Contig 2:

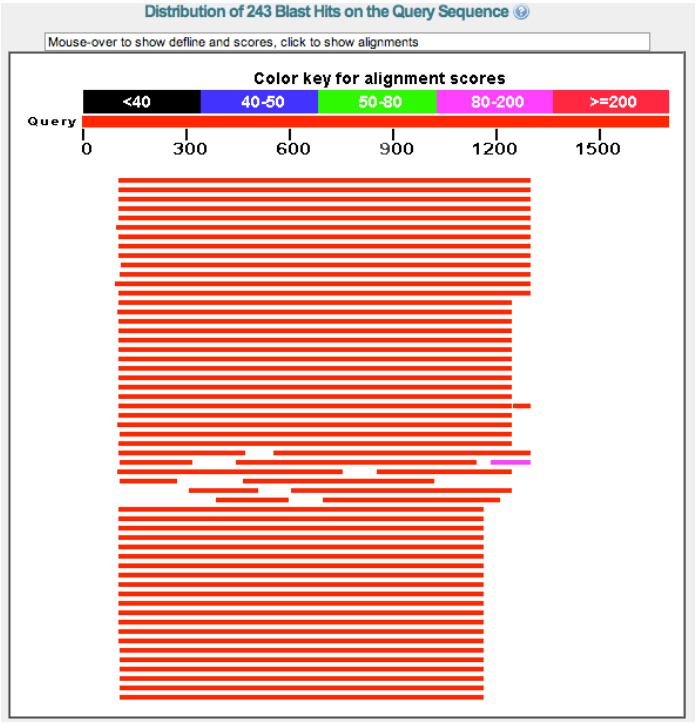

Fig 26: BLAST results indicate that contig 2 has very high alignment scores with a large number of microbes. This is further evidence that contig 2 is contamination and does not belong in my assembly.

Sequences producing significant alignments:		Score (Bits)	E Value
ref|NC_012947.1|	Escherichia coli BL21(DE3), complete genome	2122	0.0
ref|NC_012759.1|	Escherichia coli BW2952, complete genome	2122	0.0
ref|NC_010473.1|	Escherichia coli str. K-12 substr. DH10B, co...	2122	0.0
ref|AC_000091.1|	Escherichia coli str. K-12 substr. W3110, co...	2122	0.0
ref|NC_000913.2|	Escherichia coli str. K-12 substr. MG1655 ch...	2122	0.0
ref|NC_007795.1|	Staphylococcus aureus subsp. aureus NCTC 832...	2118	0.0
ref|NC_010468.1|	Escherichia coli ATCC 8739, complete genome	2114	0.0
ref|NC_011751.1|	Escherichia coli UMN026, complete genome	2102	0.0
ref|NC_011749.1|	Escherichia coli UMN026 plasmid p1ESCUM, com...	2102	0.0
qb|ACID01000030.1|	Escherichia sp. 1_1_43 cont1.30, whole gen...	2096	0.0
ref|NC_011745.1|	Escherichia coli ED1a, complete genome	2091	0.0
ref|NC_011083.1|	Salmonella enterica subsp. enterica serovar ...	1824	0.0
ref|NC_011415.1|	Escherichia coli SE11, complete genome	1820	0.0
ref|NZ_AAJV02000017.1|	Escherichia coli E22 gcontig_111249565...	1698	0.0
ref|NZ_AAJV02000020.1|	Escherichia coli E22 gcontig_111249565...	1698	0.0
ref|NZ_AAJV02000004.1|	Escherichia coli E22 gcontig_111249566...	1696	0.0
ref|NZ_AAJV02000011.1|	Escherichia coli E22 gcontig_111249565...	1696	0.0
ref|NZ_AAJV02000015.1|	Escherichia coli E22 gcontig_111249565...	1696	0.0
ref|NZ_AAJV02000019.1|	Escherichia coli E22 gcontig_111249564...	1696	0.0
ref|NZ_AAJV02000021.1|	Escherichia coli E22 gcontig_111249565...	1696	0.0
ref|NZ_AAJV02000039.1|	Escherichia coli E22 gcontig_111249565...	1696	0.0
ref|NZ_AAJV02000026.1|	Escherichia coli E22 gcontig_111249566...	1694	0.0
ref|NZ_AAJV02000028.1|	Escherichia coli E22 gcontig_111249564...	1694	0.0
ref|NZ_AAJV02000029.1|	Escherichia coli E22 gcontig_111249565...	1692	0.0

Fig 27: BLAST results indicate that contig 2 has a very low probability of being discrepant with a large number of *E. coli* genome fragments. Thus, the contig 2 contamination is likely from *E. Coli*.

Blast Search for Contig 4:

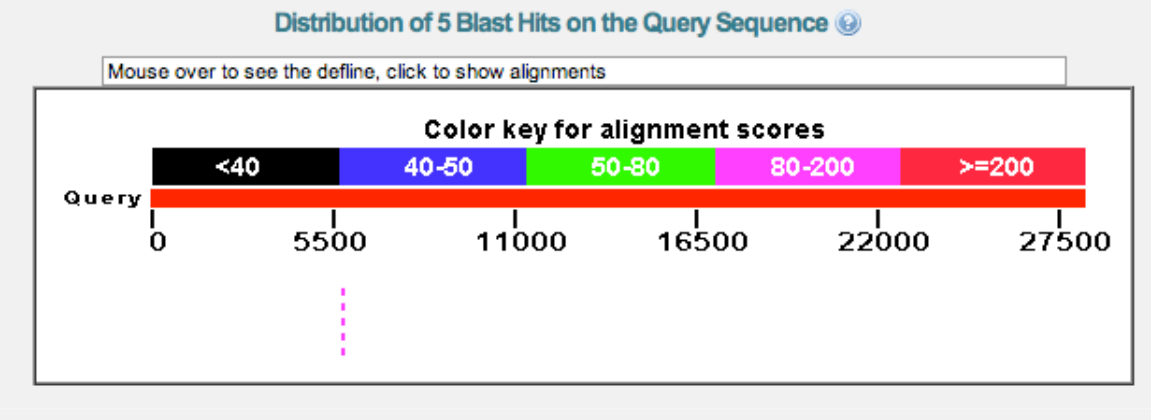

Fig 28: The BLAST search distribution for contig 4 indicates that there are no very high and large alignment scores with microbes that would suggest contamination in my assembly. However, a microbe with a small score was identified near the 5500 bp mark.

Sequences producing significant alignments:		Score (Bits)	E Value
ref|NZ_AAGB01000007.1|	Wolbachia endosymbiont of Drosophila a...	87.9	1e-12
ref|NZ_AAGB01000209.1|	Wolbachia endosymbiont of Drosophila a...	87.9	1e-12
ref|NZ_AAGB01000337.1|	Wolbachia endosymbiont of Drosophila a...	87.9	1e-12
ref|NZ_AAGB01000377.1|	Wolbachia endosymbiont of Drosophila a...	87.9	1e-12
ref|NZ_AAGB01000133.1|	Wolbachia endosymbiont of Drosophila a...	82.4	4e-11

Fig 29: BLAST search found that *Wolbachia* sequences had a significant probability of matching with sequences in my assembly.

While it was likely that there was a sequence in my assembly corresponding to the *Drosophila* parasite *Wolbachia*, the matching portion was short compared to the entire length of the genome. Furthermore, BLAST showed that gaps between matching base pairs were common (Figure 30). This evidence allowed me to conclude that there was weak homology between the *Wolbachia* DNA and my assembly *Drosophila* DNA. Furthermore, *Wolbachia* is a type of *Drosophila* parasite whose genome has been previously documented to appear in *D. melanogaster*[1]. Thus, citing the short lengths of sequence similarity and the gaps between similar bases, it is likely that *Wolbachia* genome pieces had been incorporated into this portion of the *Drosophila* genome for an extended period of time. This would support the idea that the BLAST result did not imply contamination in my assembly. In addition, there was no evidence of additional DNA lengths in my real restriction digest over my *in silico,* and the assembly length was not abnormal.

```
>ref|NZ_AAGB01000007.1|  Wolbachia endosymbiont of Drosophila ananassae gdan_401, whole
genome shotgun sequence
 gb|AAGB01000007.1|  Wolbachia endosymbiont of Drosophila ananassae gdan_401, whole
genome shotgun sequence
Length=23298

 Score = 87.9 bits (47),  Expect = 1e-12
 Identities = 103/128 (80%), Gaps = 11/128 (8%)
 Strand=Plus/Plus

Query  5838   ACGGACGGACGGACATGACT-ATATCGGTTC-GGCTGTTATCCTGATCAAGAATATATAT  5895
              |||||| |||||||||| || ||||||  || ||  || ||||||||||||||||||||
Sbjct  22642  ACGGACAGACGGACATG-CTCATATCGACTCAGGAGGTGATCCTGATCAAGAATATATAT  22700

Query  5896   ACTTTAT-GGGATCGGAGATG-CTTCTTT-CTGCCTGTTACATACATTTGCAC-AAAACC  5951
              ||||||| ||| |||||||||| || |||  ||||  ||| || || |||| || ||||
Sbjct  22701  ACTTTATAGGG-TCGGAGATGTCTCCTTCACTGC--GTTGCACACTTTTG-ACCAAAATT  22756

Query  5952   ATTATACC  5959
              || |||||
Sbjct  22757  ATAATACC  22764


>ref|NZ_AAGB01000209.1|  Wolbachia endosymbiont of Drosophila ananassae gdan_89, whole
genome shotgun sequence
 gb|AAGB01000209.1|  Wolbachia endosymbiont of Drosophila ananassae gdan_89, whole
genome shotgun sequence
Length=1460

 Score = 87.9 bits (47),  Expect = 1e-12
 Identities = 103/128 (80%), Gaps = 11/128 (8%)
 Strand=Plus/Minus

Query  5838  ACGGACGGACGGACATGACT-ATATCGGTTC-GGCTGTTATCCTGATCAAGAATATATAT  5895
              |||||| |||||||||| || ||||||  || ||  || ||||||||||||||||||||
Sbjct  311    ACGGACAGACGGACATG-CTCATATCGACTCAGGAGGTGATCCTGATCAAGAATATATAT  253

Query  5896  ACTTTAT-GGGATCGGAGATG-CTTCTTT-CTGCCTGTTACATACATTTGCAC-AAAACC  5951
              ||||||| ||| |||||||||| || |||  ||||  ||| || || |||| || ||||
Sbjct  252    ACTTTATAGGG-TCGGAGATGTCTCCTTCACTGC--GTTGCACACTTTTG-ACCAAAATT  197

Query  5952  ATTATACC  5959
              || |||||
Sbjct  196    ATAATACC  189
```

Fig 30: The BLAST search indicates which portions of the *Wolbachia* genome correspond with my contig 4. Eight percent of the matching sequences are gaps and twenty percent are mismatches. Furthermore, the matching sequences are short. Thus, the *Wolbachi*a sequence present in my *Drosophila* assembly is likely not a contamination.

Final Restriction Digest Comparison:

The final restriction digest comparison with the digest run, "fragSizes.txt" showed (an) extra fragment(s) in the *in silico* relative to the real digest when using the EcoRI (Figure 31), HindIII, and SacI restriction enzymes. This discrepancy is most easily seen when looking at the 1.5 kb unmatched fragment in the EcoRI digest (Figure 32). Since my assembly is of high quality and few discrepancies remain, I concluded that the digest must have run poorly. Thus, I chose an alternate digest to compare with my project. The digest run, "fragSizesc.txt," had the closest agreement to my *in silico* digest using the restriction enzymes EcoRI (Figure 33), EcoRIV, HindIII (Figure 34), and SacI. Thus, the digest supports my final consensus.

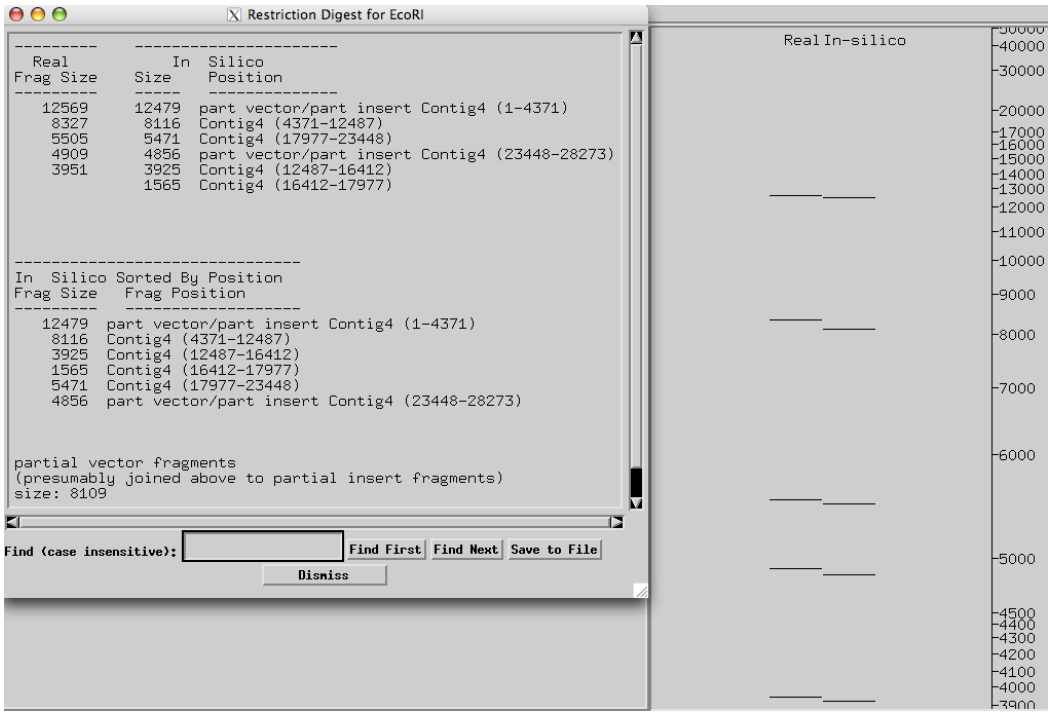

Fig 31: Restriction Digest for EcoRI using run, "fragsizes.txt."

Fig 32: The EcoRI "fragsizes.txt" *in silico* digest has an unmatched 1.5Kb fragment with the real digest data.

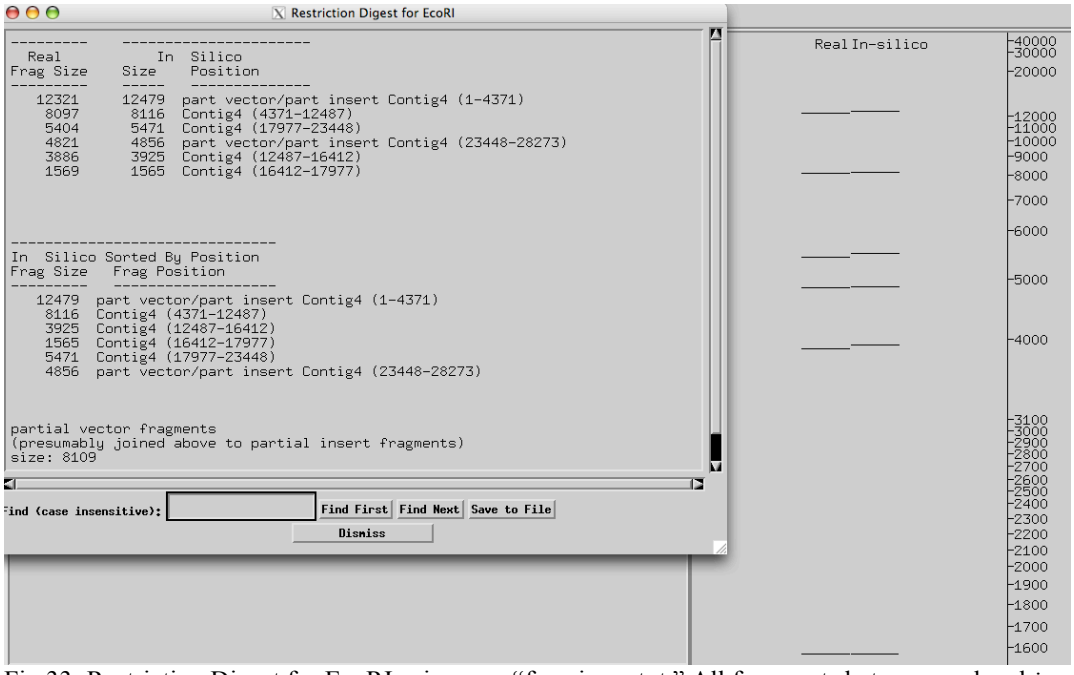

Fig 33: Restriction Digest for EcoRI using run, "fragsizesc.txt." All fragments between real and *in silico* restriction digests match.

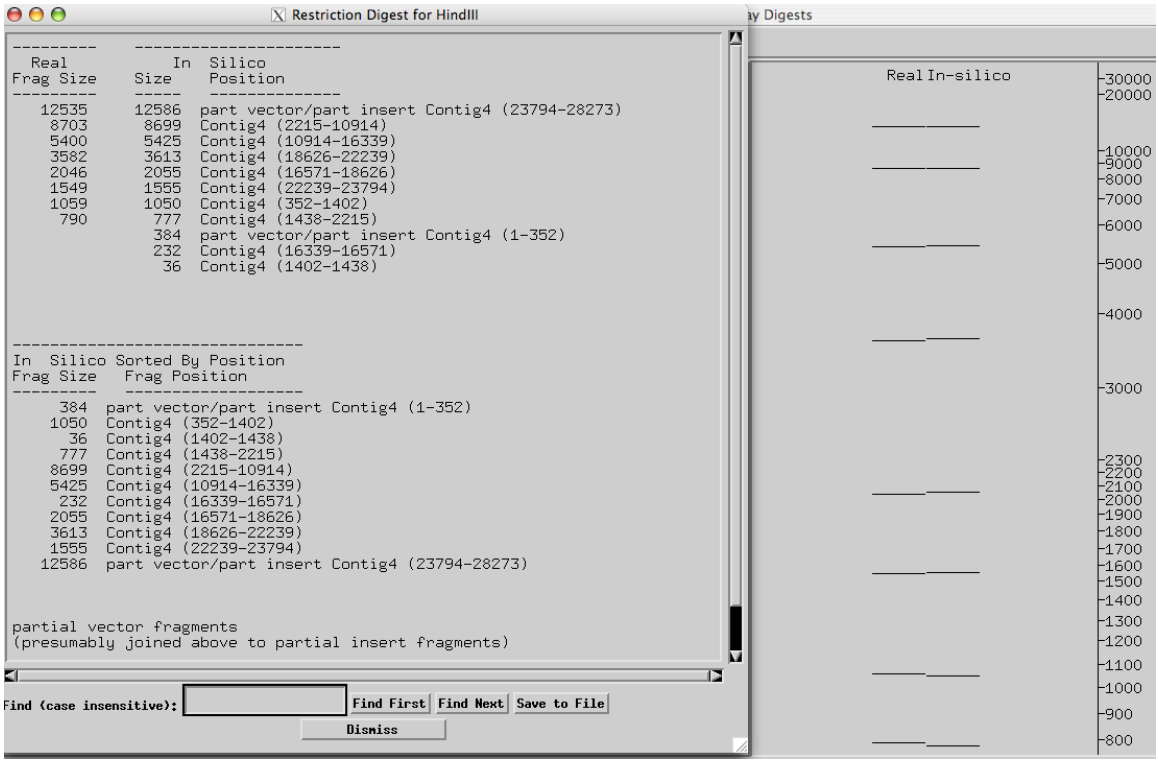

Fig 34: Restriction Digest for HindIII using run, "fragsizesc.txt." All fragments greater than 0.7 kb between real and *in silico* restriction digests match.

Conclusion

This project can be considered completely finished because it is in a single contig (as shown by final assembly view in Figure 35), cloning ends have been identified and tagged, a BLAST search has been run and no contamination was found on the main assembly (contig 4), all bases meet finishing standards, and the real restriction digests correspond to my *in silico* digests. The single stranded/chemistry regions 2 kb inside of the ends has been addressed (page 20). I have concluded that contig 2 (named contig 3 after the addition of new reads) is an *E. coli* contamination and should not be in my assembly. While crossmatch finds the region 4878-4767 in contig 4 to match with a sequence in contig 2, I have concluded that this is merely a shared repeat. Through this reasoning, the project: *Drosophila grimshawi* fosmid clone DGA19A15 is complete.

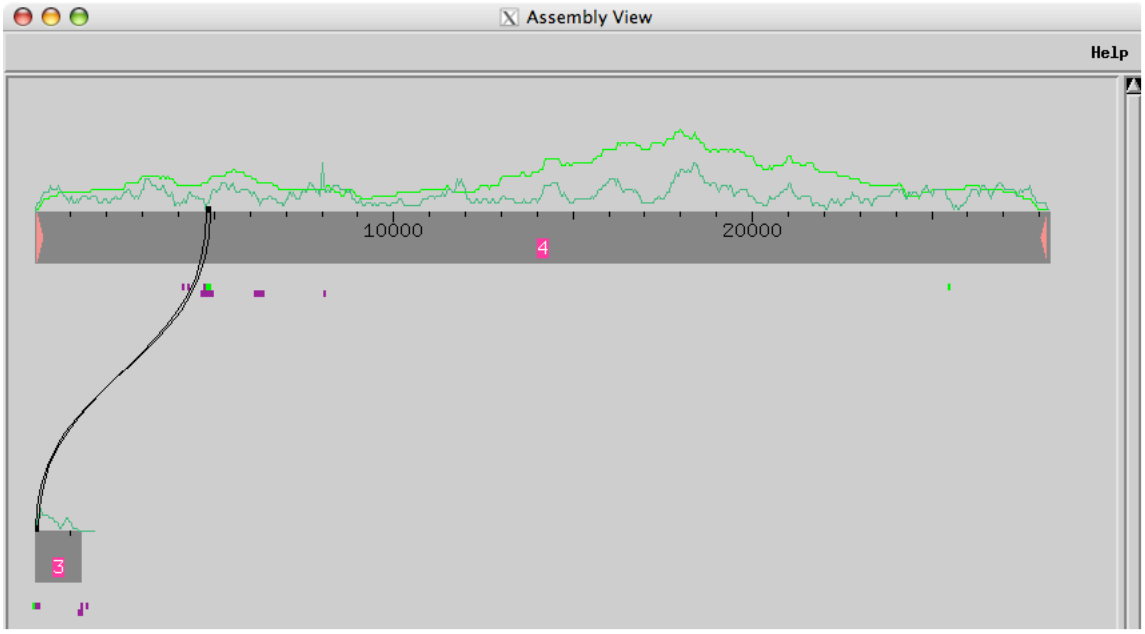

Fig 35: Final assembly view with crossmatch run.

Final Discrepancies:

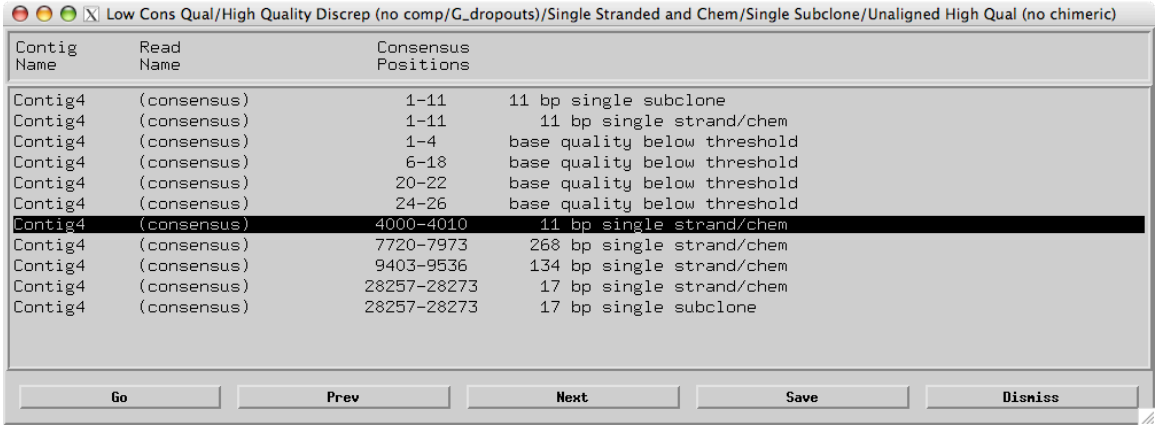

Fig 36: Problematic regions within 2 kb of the assembly ends not considered for reasons described on page 11. The single strand/chem regions within 2 kb of the assembly have been tagged.

Checklist Notes:

-Did not use any Assembly pieces (fake reads)
-Did not have any “PCR only” regions

Special Thanks

Special thanks to teaching assistant Wilson Leung and GSC finishers Neha Shah and Lee Trani for assisting me in this project at multiple levels. Their knowledge and expertise enriched my learning experience immensely. In addition, I would like to thank Professor Elgin and Professor Shaffer for making this class and, thus this project, possible.

Citation:

1. Wu M, Sun LV, Vamathevan J, *et al.* (2004). "Phylogenomics of the reproductive parasite Wolbachia pipientis wMel: a streamlined genome overrun by mobile genetic elements". *PLoS Biol.* **2** (3): E69. doi:10.1371/journal.pbio.0020069