Stephen McDaniel
April 18, 2007
Elgin, Bio 434W, Finishing Paper

Finishing Project 465-C16

Abstract:

Bio 4342/W seeks to teach students how to utilize the new tools of the genomics era through active involvement in a research project. This class began by finishing and annotating the fourth chromosome of *D. virilis*, an evolutionarily distant ancestor to *D. melanogaster*. *D. virilis* was chosen because its fourth chromosome displays a different heterochromatin protein 1 binding pattern than the fourth chromosome of *D. melanogaster*. The binding pattern indicated that the *D. virilis* fourth chromosome could be euchromatic, unlike *D. melanogaster*'s, which is heterochromatic. This semester, we have begun a new project, finishing and annotating the fourth chromosome of *D. mojavensis*, which is a close relative of *D. virilis*.

My project was to finish and annotate the 40 kb fosmid 465-C16. This paper details the process of finishing this fosmid.

The Process:

The Initial Assembly:

After sequencing reactions were carried out, the Phred program called bases as they left the sequencer and the Phrap program assembled all of these reads into four contigs. When Consed, the program that we would be using to analyze our sequence and finish it, was first opened I was presented with the assembly seen below in Figure 1.

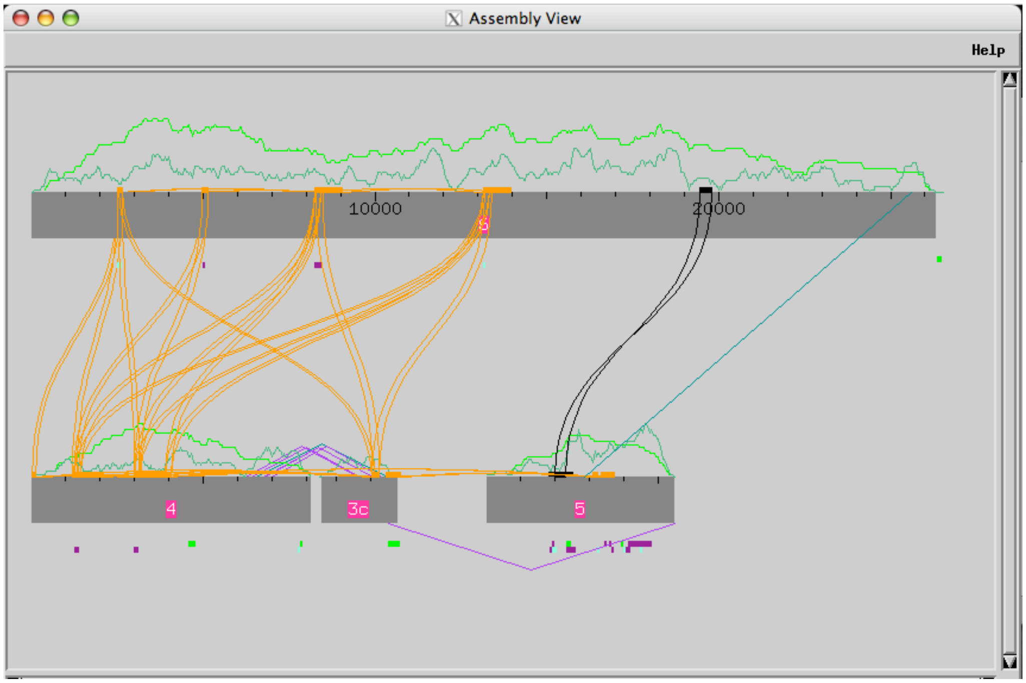


Figure 1: Initial Assembly View.

Crossmatch, a program used to identify same-orientation and inverted repeats was run. The orange lines represent same-orientation repeats in the fosmid sequence, while the black lines represent inverted repeats in the fosmid sequence. The purple arches represent forward and reverse pairs of sequence reads that are separated by a gap in the fosmid. These pairs help to piece together the order of the contigs in the overall fosmid structure.

First, the ends of the fosmid were marked with a comment tag based on sequence reads from the original fosmid. The forward read, oriented left to right, signals the left end of the fosmid, while the reverse read, oriented right to left, signals the right end. The end reads are XBAB-465C16f.b1 (forward) and XBAB-465C16f.g1 (reverse) and are contained in contigs three and five respecively.

High Quality Discrepancies:

Once the ends of the fosmid were marked, other navigation tools were used to look for other problems in the fosmid. Navigating by high quality discrepancies was the first search tool used. Quite a few were found, but most of them were due to pads that were inserted into my sequence erroneously. If the traces in Figure 2 are examined carefully, we can see that all three of the sequences actually agree with one another. The middle sequence reads ACGTGTATAATGCATGC. That same sequence is found in both of the other two sequences. But because the middle sequence is spread out, pads were inserted into it and into the other two normal sequences to make all three of them align properly. Because inspection indicated that there was no problem with these areas, no further action was taken.

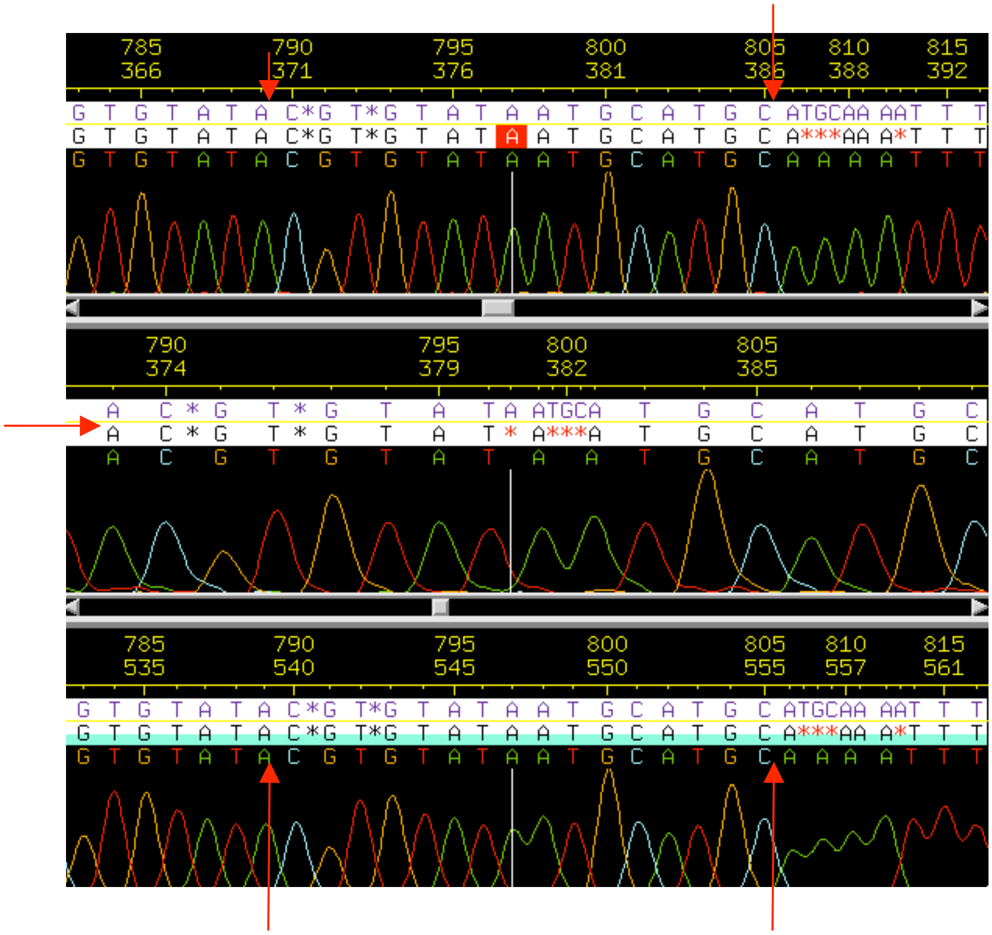

785 790 795 800 805 810 815
366 371 376 381 386 388 392
G T G T A T A C*G T*G T A T A A T G C A T G C ATGCAA AAT T T
G T G T A T A C*G T*G T A T A A T G C A T G C A***AA A*T T T
G T G T A T A C G T G T A T A A T G C A T G C A A A A T T T
790 795 800 805
374 379 382 385
A C * G T * G T A T A ATGCA T G C A T G C
A C * G T * G T A T * A***A T G C A T G C
A C G T G T A T A A T G C A T G C
785 790 795 800 805 810 815
535 540 545 550 555 557 561
G T G T A T A C*G T*G T A T A A T G C A T G C ATGCAA AAT T T
G T G T A T A C*G T*G T A T A A T G C A T G C A***AA A*T T T
G T G T A T A C G T G T A T A A T G C A T G C A A A A T T T

Figure 2: Example of incorrectly placed pads. Arrows are pointing to the correct sequence.

There were, however, two genuine examples of high quality discrepancies in the project (Figure 3). Upon examining the traces, it was evident that there was no question that the bases called were correct in all cases. Because the traces were unambiguous, these discrepancies were either due to polymorphisms in the fly population or a large duplicated region in my fosmid. Because the fly population went through ten generations of inbreeding, it was tentatively concluded that these reads were due to a collapsed duplication.

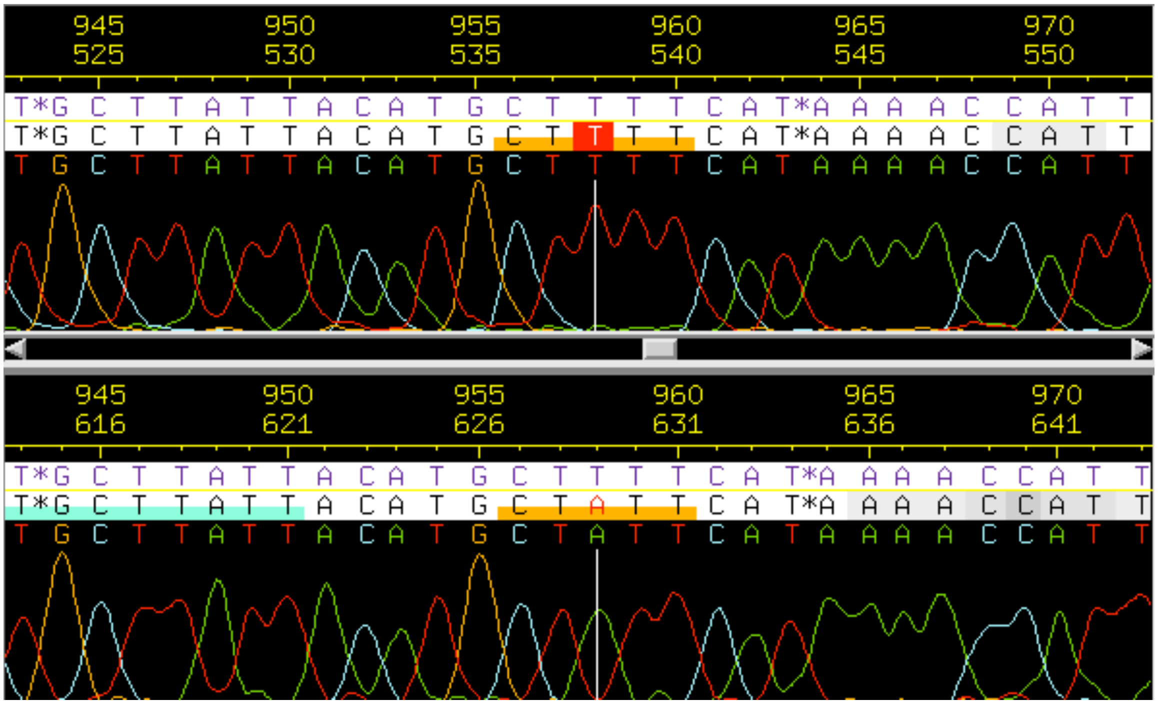

945 950 955 960 965 970
525 530 535 540 545 550
T*G C T T A T T A C A T G C T T T T C A T*A A A A C C A T T
T*G C T T A T T A C A T G C T T T T C A T*A A A A C C A T T
T G C T T A T T A C A T G C T T T T C A T A A A A C C A T T
945 950 955 960 965 970
616 621 626 631 636 641
T*G C T T A T T A C A T G C T T T T C A T*A A A A C C A T T
T*G C T T A T T A C A T G C T A T T C A T*A A A A C C A T T
T G C T T A T T A C A T G C T A T T C A T A A A A C C A T T

Figure 3: Example of high quality discrepant traces.

These two bases were marked as being high quality. When Phred/Phrap was rerun, these discrepant reads would not be aligned again. Because these regions were not overlapped, the assembly went from four contigs to six (Figure 4). The next step was to look for matching reads at the ends of these contigs to try to force join them into a single contig. There was a problem though, because the new contigs were largely identical to each other. They only had a few bases of unique sequence each. Thus the gaps could not be resolved at this point. Each of the attempts to join between the contigs contained high quality discrepancies. More unique information would be needed.

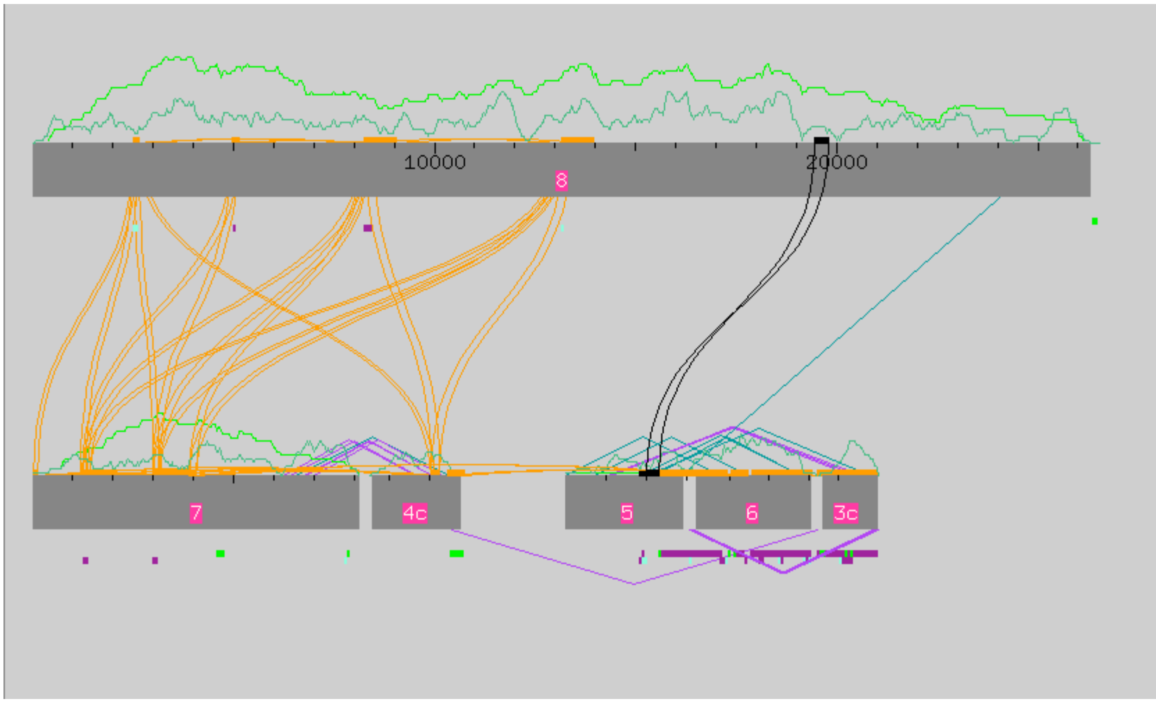


Figure 4: Post Phred/Phrap assembly view.

Because no new sequence data was available, a new approach was attempted. Reads that had forward and reverse pairs that contained both of the discrepancies were pulled out of their contigs. Then the *miniassembly* function was used to create a bridge between the hypothesized duplications in the fosmid. There was one such read, 0389637E08. This read had a forward read with one high quality discrepancy and a reverse read with the other high quality discrepancy (Figure 5). Using this read, similar reads could be pulled out and assembled into the "correct" location. By doing this, force joins were carried out, artificially recreating the duplication in the fosmid (Figure 6).

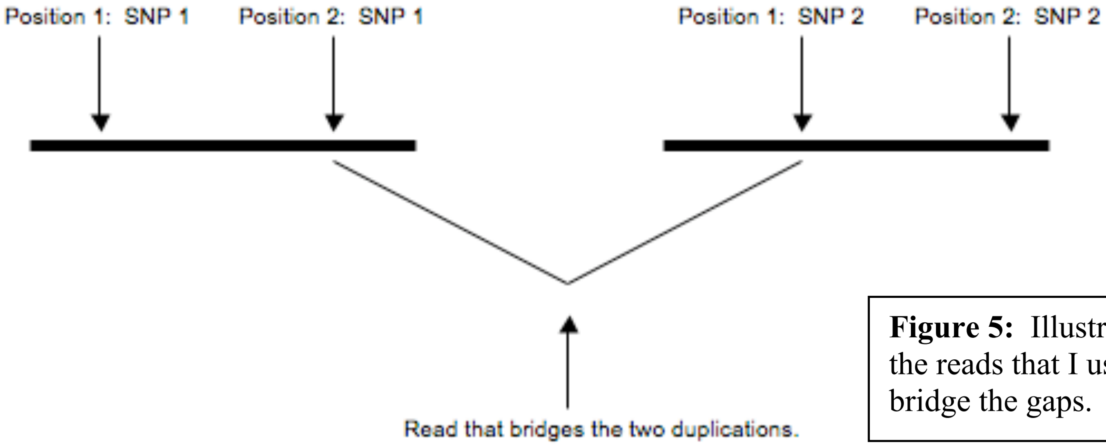
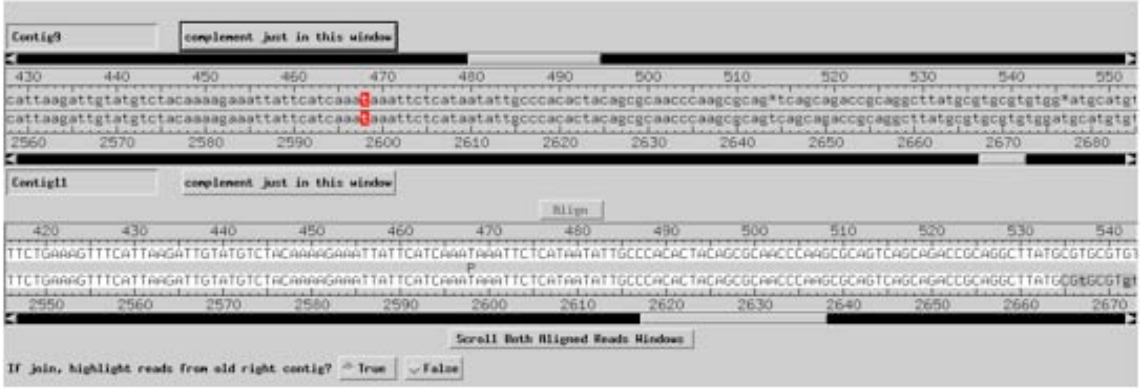


Figure 5: Illustration of the reads that I used to bridge the gaps.



Contig9
complement just in this window
Contig11
complement just in this window
Align
Scroll Both Aligned Reads Windows
If join, highlight reads from old right contig?
True
False

```
 430       440       450       460       470       480       490       500       510       520       530       540       550
cattaagattgtatgtctacaaaagaaattattcatcaaataaattctcataatattgcccacactacagcgcaacccaagcgcag*tcagcagaccgcaggcttatgcgtgcgtgtgg*atgcatgt
cattaagattgtatgtctacaaaagaaattattcatcaaataaattctcataatattgcccacactacagcgcaacccaagcgcagtcagcagaccgcaggcttatgcgtgcgtgtggatgcatgtg
2560      2570      2580      2590      2600      2610      2620      2630      2640      2650      2660      2670      2680

 420       430       440       450       460       470       480       490       500       510       520       530       540
TTCTGAAAGTTTCATTAAGATTGTATGTCTACAAAAGAAATTATTCATCAAATAAATTCTCATAATATTGCCCACACTACAGCGCAACCCAAGCGCAGTCAGCAGACCGCAGGCTTATGCGTGCGTGT
                                                   P
TTCTGAAAGTTTCATTAAGATTGTATGTCTACAAAAGAAATTATTCATCAAATAAATTCTCATAATATTGCCCACACTACAGCGCAACCCAAGCGCAGTCAGCAGACCGCAGGCTTATGCGtGCGTgt
 2550      2560      2570      2580      2590      2600      2610      2620      2630      2640      2650      2660      2670
```

Figure 6: Example of a contig join.

Discrepant Forward/Reverse Pairs:

This duplication introduced a new problem into the assembly. As reads were moved, and contigs joined, discrepant forward/reverse pairs were created. The forward/reverse end pair reads are all supposed to be about 2000 bp apart. If there is a gap in the assembly, Consed estimates its size and then decides whether or not the end read pairs are spaced appropriately. The new duplication of the region containing the two high quality discrepancies in the assembly, forced some of the read pairs too far apart from each other. Consed then marked them as discrepant forward/reverse pairs. These are the red triangles seen in Figure 7. Because these reads are too far apart, they are most likely misassembled and have to be looked at individually. Each of the reads was pulled out into its own contig and reassembled in the correct location. The *miniassembly* function was used when there were several reads from the same location being reassembled. The *search for string* function was used for single reads. At the end of this process several contigs were joined together, though not all of the discrepant forward reverse pairs were resolved.

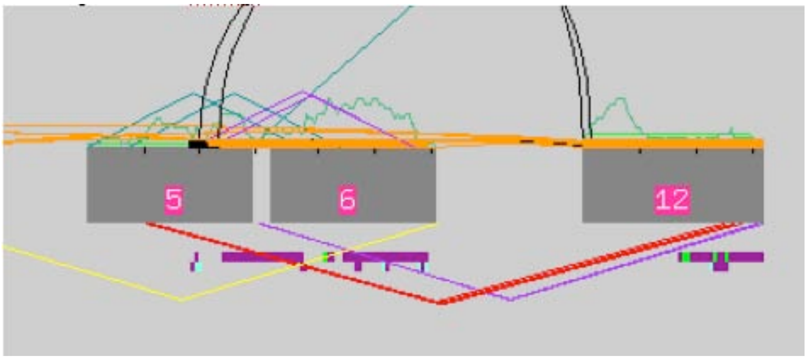
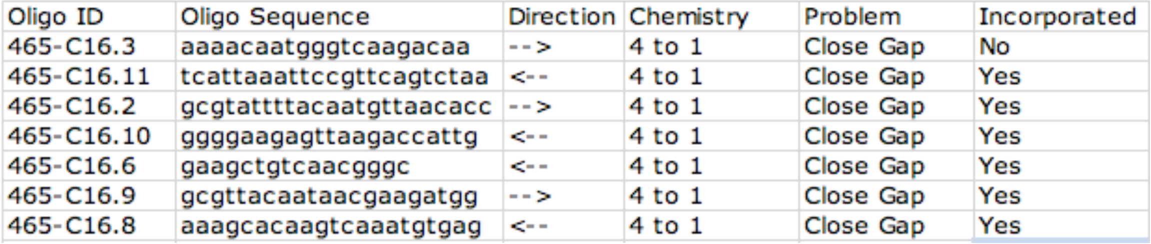


Figure 7: Discrepant forward reverse pairs.

The First Round of Reactions and a Comparison to Autofinish:

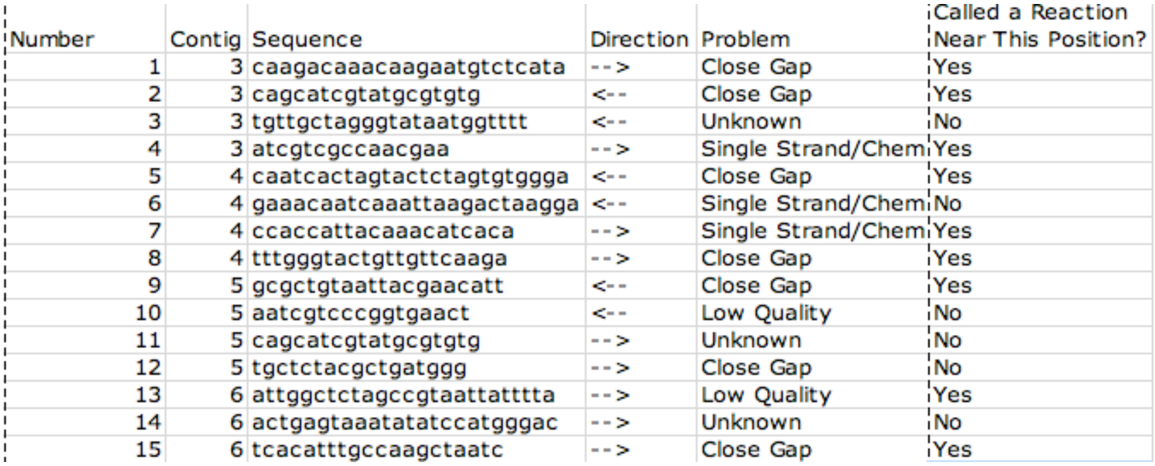
At this point the project was in five contigs. By tearing the discrepant reads out, several of the smaller contigs were joined together. Now the larger contigs needed to be connected. To do this, unique sequence primers near most of the gaps were found. The reaction chemistry used for all of these was 4:1. The previous sequencing reactions were all done with Big Dye chemistry because it is the cheapest. But these regions appear to be difficult to sequence, so a more robust sequencing chemistry was used. Some of the gaps did not have sequencing reactions called for them however. These gaps were surrounded by regions containing both the high quality discrepancies and repetitive sequence. Unique sequence was not found in these regions, making it impossible to create unique sequencing primers. Thus, seven primers were called to close the gaps between the unique contigs (Table 1). These reactions helped to close all of the gaps surrounded by unique sequence in the project, due to a high reaction success rate.



Oligo ID	Oligo Sequence	Direction	Chemistry	Problem	Incorporated
465-C16.3	aaaacaatgggtcaagacaa	-->	4 to 1	Close Gap	No
465-C16.11	tcattaaattccgttcagtctaa	<--	4 to 1	Close Gap	Yes
465-C16.2	gcgtattttacaatgttaacacc	-->	4 to 1	Close Gap	Yes
465-C16.10	ggggaagagttaagaccattg	<--	4 to 1	Close Gap	Yes
465-C16.6	gaagctgtcaacgggc	<--	4 to 1	Close Gap	Yes
465-C16.9	gcgttacaataacgaagatgg	-->	4 to 1	Close Gap	Yes
465-C16.8	aaagcacaagtcaaatgtgag	<--	4 to 1	Close Gap	Yes

Table 1: Round one reactions.

None of the Autofinish primers were chosen (Table 2). By the time the first round of reactions were called, the assembly was quite different from the one in the original ace.1 file. Most of the above primers were near an Autofinish primer, but primers that Consed recommended were chosen instead. Consed recommended longer primers or ones with a higher melting temperature. These qualities aid in insuring locus specificity and successful sequencing reactions. Autofinish also recommended too many reactions. A few of these reactions were not called because the ends of my fosmid were marked, and there was not a gap to fill at these locations. Autofinish did not recognize this and called reads off both ends of the fosmid. Autofinish called other reactions that had no clear purpose, or had multiple annealing sites (number 11, table 2). The multiple annealing sites would render these reactions useless because I would get sequencing product from both annealing locations, and the signals would be equally represented in the trace.



Number	Contig	Sequence	Direction	Problem	Called a Reaction Near This Position?
1	3	caagacaaacaagaatgtctcata	-->	Close Gap	Yes
2	3	cagcatcgtatgcgtgtg	<--	Close Gap	Yes
3	3	tgttgctagggtataatggtttt	<--	Unknown	No
4	3	atcgtcgccaacgaa	-->	Single Strand/Chem	Yes
5	4	caatcactagtactctagtgtggga	<--	Close Gap	Yes
6	4	gaaacaatcaaattaagactaagga	<--	Single Strand/Chem	No
7	4	ccaccattacaaacatcaca	-->	Single Strand/Chem	Yes
8	4	tttgggtactgttgttcaaga	-->	Close Gap	Yes
9	5	gcgctgtaattacgaacatt	<--	Close Gap	Yes
10	5	aatcgtcccggtgaact	<--	Low Quality	No
11	5	cagcatcgtatgcgtgtg	-->	Unknown	No
12	5	tgctctacgctgatggg	-->	Close Gap	No
13	6	attggctctagccgtaattatttta	-->	Low Quality	Yes
14	6	actgagtaaatatatccatgggac	-->	Unknown	No
15	6	tcacatttgccaagctaatc	-->	Close Gap	Yes

Table 2: Autofinish Suggestions

Note: Contig numbers refer to the original Assembly View.

Restriction Digests:

The assembly with the bridged region was about 45 kb, while fosmids are usually tightly limited to about 40 kb in length. An extra five kb would be highly unusual for a fosmid construct. This suggested that the high quality discrepancies were due to polymorphisms in the fly population as opposed to a duplicated region. There is no recombination on the fourth chromosome; therefore if any SNPs are present in the population, there can be no recombination to get rid of them. Four restriction enzyme digests were used to analyze the assembly: HindIII, EcoRI, EcoRV, and SacI. Consed utilizes the real digest information and compares that to an artificial, or in-silico, digest that it performs on the current assembly. This tool is useful to test for any misassembled areas in the project because any large misassemblies will create a discrepancy between the real and in-silico digests. Because this tool was now available, all of the discrepant regions were collapsed down into a single copy with the high quality discrepancies once again present (Figure 8). If the digest showed problems with the single contig assembly, a previous ace file could be loaded and a different approach tried.

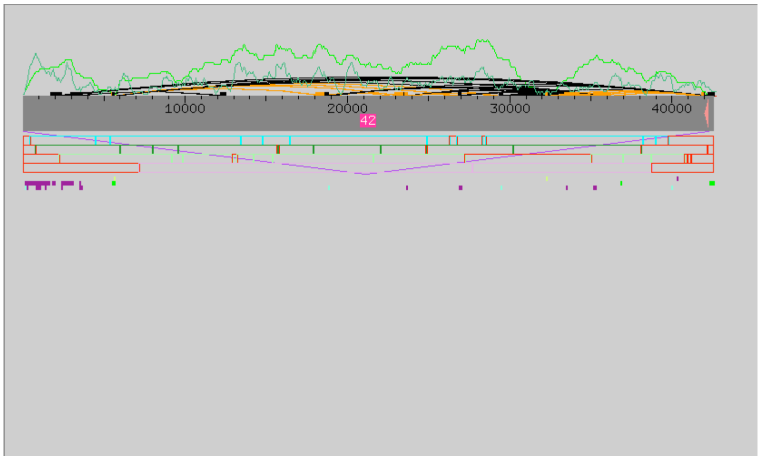
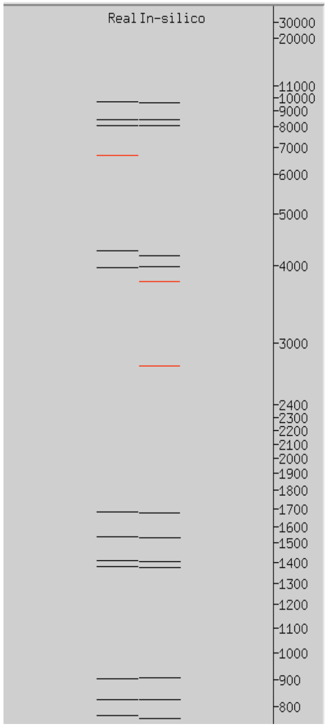
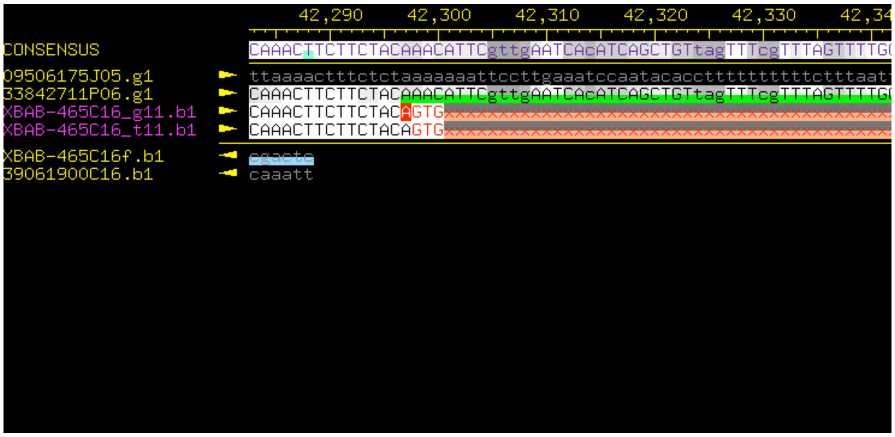


Figure 8: Assembly view after round one reads were added.

9

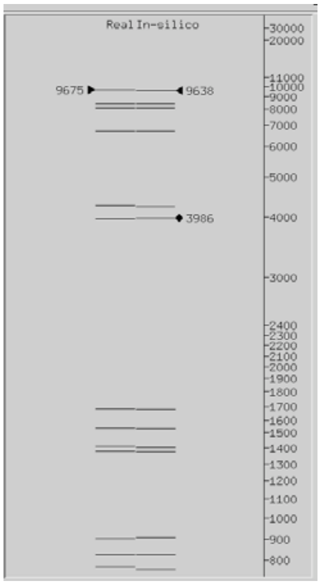


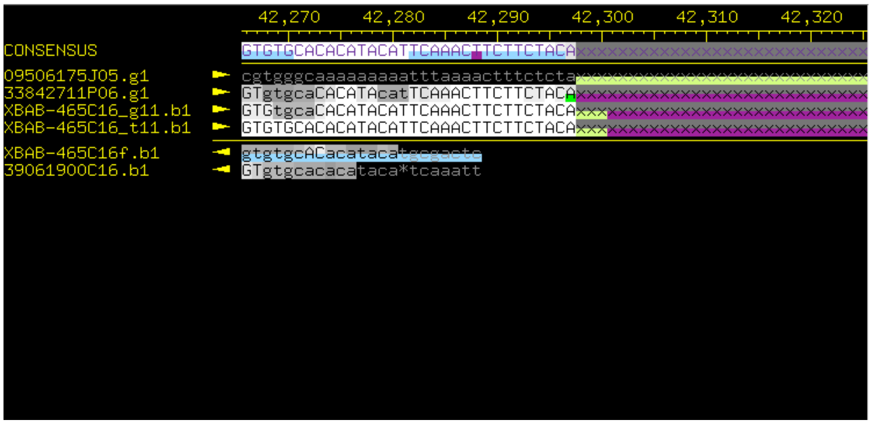
10



```
                    42,290    42,300    42,310    42,320    42,330    42,34
CONSENSUS              CAAACTTCTTCTACAAACATTCgttgAATCAcATCAGCTGTtagTTTcgTTTAGTTTTG
09506175J05.g1         ttaaaactttctctaaaaaaattccttgaaatccaatacacctttttttttctttaatt
33842711P06.g1         CAAACTTCTTCTACAAACATTCgttgAATCAcATCAGCTGTtagTTTcgTTTAGTTTTG
XBAB-465C16_g11.b1     CAAACTTCTTCTACAGTGxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
XBAB-465C16_t11.b1     CAAACTTCTTCTACAGTGxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
XBAB-465C16f.b1        cgactc
39061900C16.b1         caaatt
```

Upon looking at the EcoRV digest, a problem was immediately evident (Figure 9). The comparison showed a major discrepancy between the real and in-silico digests. The real digest had a band between 6000 and 7000 bp. The in-silico digest had two bands at about 2600 bp and 3700 bp. It was first thought that this was due to incorrectly compressing the duplicated region, but this was in fact due to a mistake in marking vector sequence. These reads at the end of the fosmid are made up of both read sequence and vector sequence. One read at the right end of the fosmid had not been marked for vector sequence (Figure 10). When Consed was searching for cut sites for the EcoRV enzyme, it found some in the vector sequence and this caused the discrepancy in the in-silico digest. Once this mislabeled read was marked correctly for vector sequence (Figure 11), the problem in the digest was immediately resolved (Figure 12). With this new evidence, it is highly probably that the single contig assembly is correct and the two high quality discrepancies are due to polymorphisms and not a duplicated region in the fosmid. If the bridged read assembly had been correct, the real and in-silico digests would have still shown several discrepancies in the single contig assembly. Because they do not, it was concluded that polymorphisms in the fly population were the cause of the high quality discrepancies found in the fosmid. The actual bridge read must





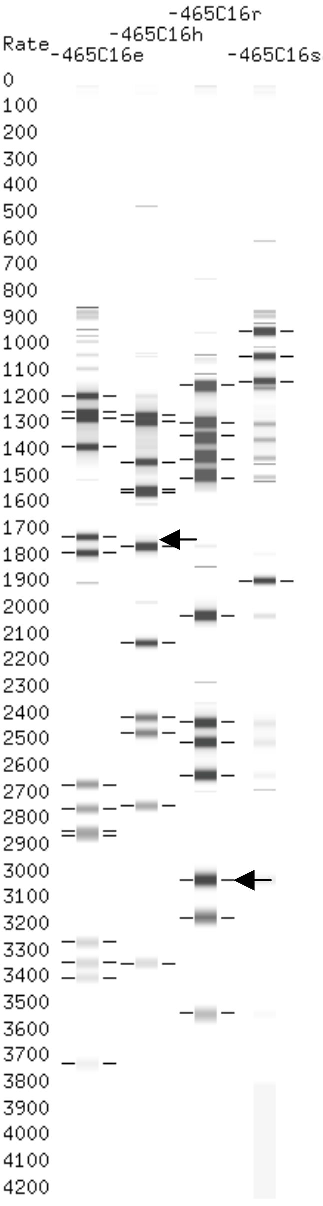
```
                    42,270    42,280    42,290    42,300    42,310    42,320
CONSENSUS              GTGTGCACACATACATTCAAACTTCTTCTACAxxxxxxxxxxxxxxxxxxxxxxxxxxxx
09506175J05.g1         cgtgggcaaaaaaaaatttaaaactttctctaxxxxxxxxxxxxxxxxxxxxxxxxxxxx
33842711P06.g1         GTgtgcaCACATAcatTCAAACTTCTTCTACAxxxxxxxxxxxxxxxxxxxxxxxxxxxx
XBAB-465C16_g11.b1     GTGtgcaCACATACATTCAAACTTCTTCTACAxxxxxxxxxxxxxxxxxxxxxxxxxxxx
XBAB-465C16_t11.b1     GTGTGCACACATACATTCAAACTTCTTCTACAxxxxxxxxxxxxxxxxxxxxxxxxxxxx
XBAB-465C16f.b1        gtgtgcACacatacatgcgactc
39061900C16.b1         GTgtgcacacataca*tcaaatt
```

11

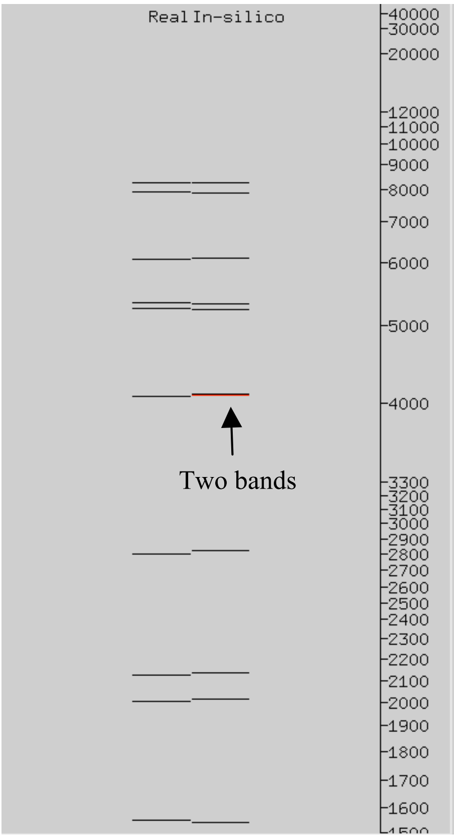
Figures 9, 10, 11, and 12: The incorrect EcoRV gel digest (9). The read that was not marked for vector data (10). The red x's below that read signify vector sequence. The corrected read (11) and the corrected digest (12). Note that all of the reads are now appropriately marked with x's signifying vector data, and there is no longer a problem in the digest.

have come from a fly that had both polymorphisms.

All of the other digests appeared to show a correct assembly as well. There were two small discrepancies in the EcoRI and HindIII digests (Figure 13), but when the picture of the gel showing the four restriction enzyme digests was examined (Figure 14), there was evidence of a doublet that had not been called by the program. The band in question is thicker and darker than the surrounding bands. This gel helped to confirm that the one-contig assembly was correct.



14



13

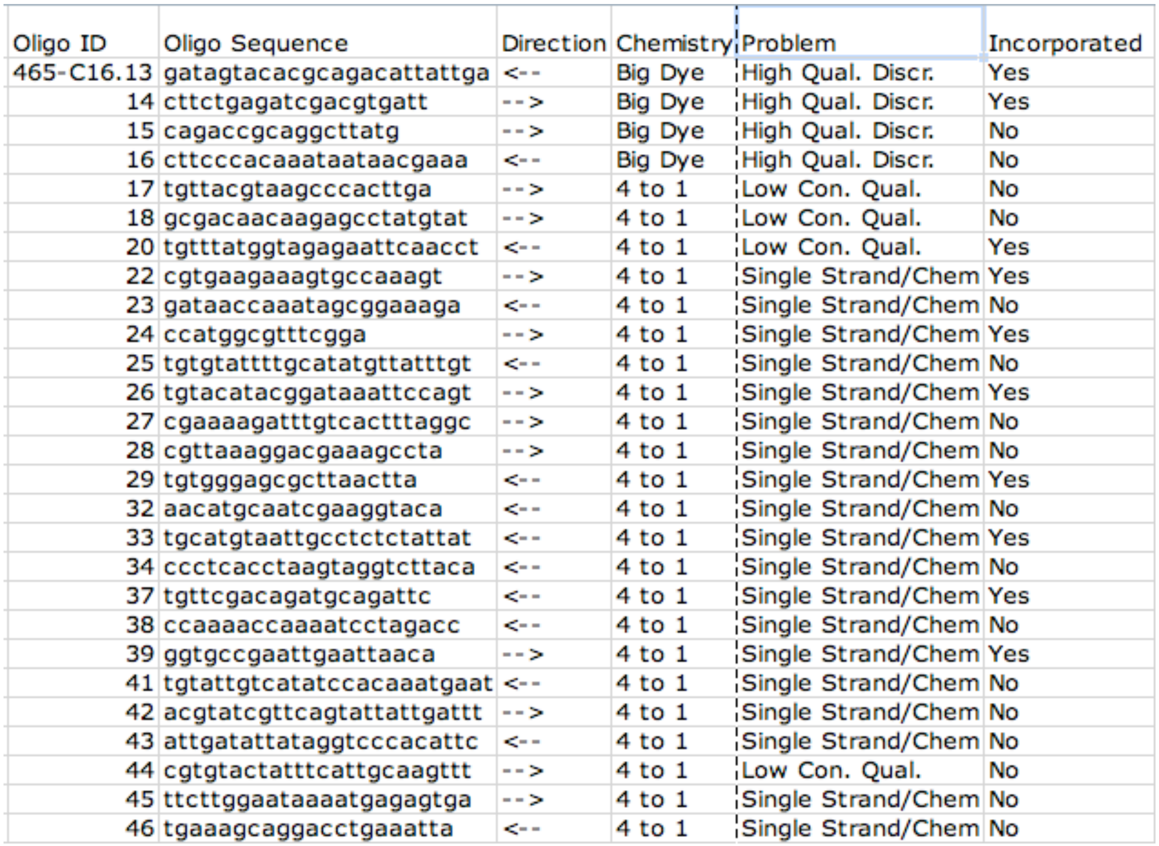
Figure 13: The HindIII digest. The red line (noted by the arrow) is actually a doublet that was incorrectly called a single band by the program that analyzed the gel.

Figure 14: The gel with all four restriction enzyme digests run out. The arrow under lane 465C16h is pointing to the suspected doublet from the restriction enzyme HindIII. It is very dark and is slightly thicker than the bands beneath it. The band in lane 465C16r represents the discrepant reading of the EcoRI restriction enzyme digest (not shown). Again, it is darker and slightly thicker than the surrounding bands, suggesting a doublet

Third Round of Reactions:

Now that the fosmid was reduced to a single contig and the high quality discrepancies were confirmed to be polymorphisms, it was time to look at some of the other problems present. The regions with low consensus quality and regions with single strand, single chemistry, and single subclone coverage were addressed in the next round of sequencing reactions (table 3). Because of all of the problems surrounding the discrepant restriction enzyme digests, the second round of reactions had been missed. These discrepancies were not resolved until the following class period when the vector sequence was recognized and tagged. This round of reactions was also used to determine which bases would be called at the two loci with high quality discrepancies. The reads that were assembled into the fosmid were taken from sequence that was generated from a population of flies which had polymorphisms at these two locations. The fosmid that the

GSC has been using as a template to sequencing from, however, represents only one DNA sequence, with a single base at the two locations. Two reactions for each of the regions with high quality discrepancies were called, one from each direction. If the reactions came back with only a single base called at each location, and the traces were clean, then it would be highly probable that these two loci are simply polymorphic in the fly population used to generate this sequence. However, if the sequence came back with two bases called at these locations, then it would suggest that there might actually be a duplication of this region. Because all four restriction enzyme digests suggested that the assembly without the duplication was correct, and the fosmid size (about 42 kb) was within an acceptable size range, the latter outcome seemed highly unlikely. If it did occur, the traces would have to be clean at all locations except for the two discrepant bases to cast significant doubt on the current assembly.



Oligo ID	Oligo Sequence	Direction	Chemistry	Problem	Incorporated
465-C16.13	gatagtacacgcagacattattga	<--	Big Dye	High Qual. Discr.	Yes
14	cttctgagatcgacgtgatt	-->	Big Dye	High Qual. Discr.	Yes
15	cagaccgcaggcttatg	-->	Big Dye	High Qual. Discr.	No
16	cttcccacaaataataacgaaa	<--	Big Dye	High Qual. Discr.	No
17	tgttacgtaagcccacttga	-->	4 to 1	Low Con. Qual.	No
18	gcgacaacaagagcctatgtat	-->	4 to 1	Low Con. Qual.	No
20	tgtttatggtagagaattcaacct	<--	4 to 1	Low Con. Qual.	Yes
22	cgtgaagaaagtgccaaagt	-->	4 to 1	Single Strand/Chem	Yes
23	gataaccaaatagcggaaaga	<--	4 to 1	Single Strand/Chem	No
24	ccatggcgtttcgga	-->	4 to 1	Single Strand/Chem	Yes
25	tgtgtattttgcatatgttatttgt	<--	4 to 1	Single Strand/Chem	No
26	tgtacatacggataaattccagt	-->	4 to 1	Single Strand/Chem	Yes
27	cgaaaagatttgtcactttaggc	-->	4 to 1	Single Strand/Chem	No
28	cgttaaaggacgaaagccta	-->	4 to 1	Single Strand/Chem	No
29	tgtgggagcgcttaactta	<--	4 to 1	Single Strand/Chem	Yes
32	aacatgcaatcgaaggtaca	<--	4 to 1	Single Strand/Chem	No
33	tgcatgtaattgcctctctattat	<--	4 to 1	Single Strand/Chem	Yes
34	ccctcacctaagtaggtcttaca	<--	4 to 1	Single Strand/Chem	No
37	tgttcgacagatgcagattc	<--	4 to 1	Single Strand/Chem	Yes
38	ccaaaaccaaaatcctagacc	<--	4 to 1	Single Strand/Chem	No
39	ggtgccgaattgaattaaca	-->	4 to 1	Single Strand/Chem	Yes
41	tgtattgtcatatccacaaatgaat	<--	4 to 1	Single Strand/Chem	No
42	acgtatcgttcagtattattgattt	-->	4 to 1	Single Strand/Chem	No
43	attgatattataggtcccacattc	<--	4 to 1	Single Strand/Chem	No
44	cgtgtactatttcattgcaagttt	-->	4 to 1	Low Con. Qual.	No
45	ttcttggaataaaatgagagtga	-->	4 to 1	Single Strand/Chem	No
46	tgaaagcaggacctgaaatta	<--	4 to 1	Single Strand/Chem	No

Table 3: The third round of reactions.

In Conclusion:

Unfortunately, about half of the reactions called failed to produce usable sequence data. Included in these failed reactions were both of the reactions that were called for the second site with a high quality discrepancy. This was most likely due to the chemistry chosen for these reactions. Because there was relatively good sequence surrounding both of the discrepant locations, Big Dye chemistry was used in all four sequencing reactions. 4:1 would have most likely been a better choice. For the first aberrant location both sequencing reactions were successful and I was able to resolve this discrepancy. A single base (T) came back for both of the reactions.

Many of the single strand/chemisty/subclone issues were fully resolved or significantly improved. Because many of these regions were quite large, some extending for 1000 bp or more, multiple reactions were called within these regions. Due to their large size some of these problem areas still remain, yet with a few more rounds of reactions, these problems should be completely resolved. Most of these regions had Phred quality scores above 30. But for those that did not, a decent depth of high quality reads in at least one direction suggested that the sequence is correct in that region. The traces from each of the low consensus quality regions were also examined. There was only one location in which the traces were insufficient in supporting the consensus sequence with confidence. This region was tagged for more data needed.

Mononucleotide runs, X's, and N's were also searched for in the consensus sequence of the fosmid. Mononucleotide runs can be difficult to sequence through. The sequencing reactions often fail after these runs. These runs can also obscure single nucleotides that are incorporated within or around those sequences. No mononucleotide runs were found. The X's refer to vector sequence. Vector sequence was identified at the right end of the fosmid, but not the left. More sequencing will have to be carried out to ensure completeness on the left end of the fosmid. There was no vector sequence within the fosmid. The N's refer to positions that do not have sequence of a high enough quality to make a confident base call. There were no N's found in the fosmid sequence.

The final problem investigated was a small contig that was not assembled into the final contig. It contains two reads and is 824 bases long. After using the *search for string* function in multiple locations and finding no matches in the main contig or with any of the other single reads that were not incorporated, a BLAST search was conducted with the sequence, hits matching sequences in *Drosophila arizonae* and *Drosophila buzzatii* were found. These species were never cloned for sequencing, so these BLAST results could not be due to contamination. It is more likely that there are repeats in the fosmid that do not match to *D. melanogaster* sequence. Because no matches were found within the fosmid even under relaxed search parameters, it is unlikely that this contig belongs in the consensus sequence.

In conclusion, there are still a few remaining problems with the fosmid, but I am confident that the finishing staff at the GSC can easily solve them. I would like to thank all of the finishers who helped me along the way. I greatly appreciate their patience, explanations, and advice that they offered throughout this finishing process.

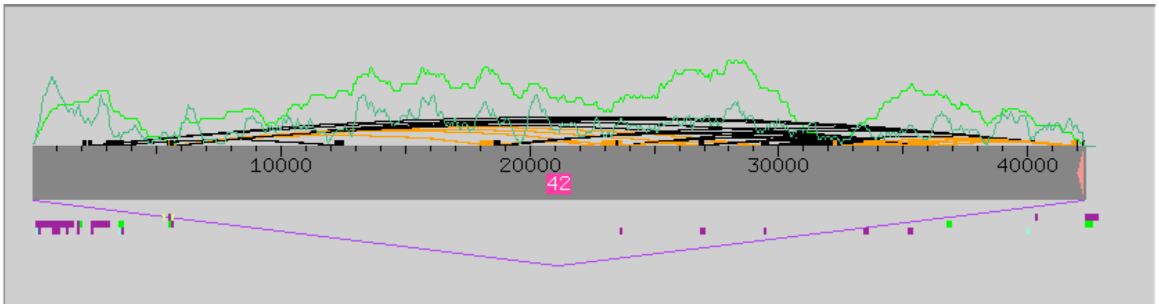


Figure 15: Project 465-C16: Final Assembly View