Finishing *Drosophila grimshawi* Fosmid Clone 05E01

Jimmy Ma
Bio 434W
Professor Elgin
March 25, 2010

Abstract

The *Drosophila grimshawii* fourth chromosome, or dot chromosome, is unique because it actively expresses genes despite being largely heterochromatic. Recently, the dot chromosome underwent whole genome shotgun sequencing. This project serves to finish the 05E01 clone to a high quality level appropriate for publishing. The initial difficulties encountered included one gap and a possible tandem repeat region. As the project continued, the gap was removed through a force join that introduced 226 high quality discrepancies. By calling new reads in this discrepant region, these discrepancies were found to be caused by a few, salient reads that Consed had inaccurately set as consensus. The final assembly of clone 05E01 achieved here ended with one gap remaining and all low quality or discrepant regions commented or tagged.

Analysis

The fourth chromosome, or dot chromosome, in *Drosophila grimshawi* is especially relevant to understanding how differences in chromatin packaging influence gene expression. The dot chromosome is odd because many of its genes are actively expressed despite having much of its structure tightly packaged as heterochromatin. In order to clarify how the dot chromosome expresses its genes, it is important to have high quality DNA sequence of the chromosome. This project serves to provide high quality sequence for a 40 kb fragment of the chromosome that will eventually be combined with other similar projects to give a full sequence of the *D. grimshawi* dot chromosome.

Figure 1 shows the starting state of my project after running cross_match. Initial inspection shows two contigs (4c and 5), especially high level of read coverage at the start of contig 4c, some areas of repetitious DNA, and several inconsistent forward/reverse pairs. For this project, Consed assumes that the largest insert size to be 4838 bp. The reads spanning the gap underneath the contig axis are reads that have inconsistent forward/reverse pairs that are most often due to the read ends being too far apart. The larger triangles above the contig axis represent the clone ends of the fosmid. The two pairs of clone ends follow the multiple data sets combined to make this project, showing that one of the original clones is shifted slightly in relation to the other clone.

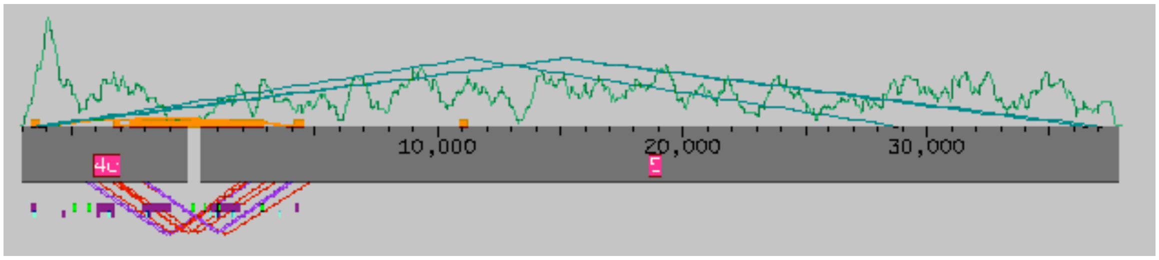
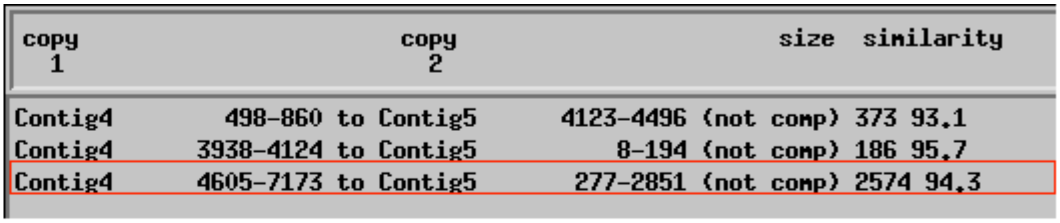


Figure 1. The initial assembly after running cross_match. Orange lines show tandem

repeat regions.

From Consed's initial analysis, contig 4c needed to be complemented in order to properly align with the rest of the project. After complementing contig 4, I used the longer fosmid ends to set the clone ends to the project, setting the fosmid between base pairs 1066 and 41962. The large amount of high quality reads downstream of the shorter fosmid ends suggested that the longer fosmid ends were the real ends to the clone. These ends were later supported by a BLAST analysis that showed microbial or vector sequence adjacent to these ends.

Next, I attempted to close the gap between contigs 4 and 5. The inconsistent read pairs suggested that the gap was small enough that single clones could span the region. Furthermore, the results from cross_match show a large tandem repeat across this area. This could mean that the area either represents one unique region with no gap or two or more sequentially repeating elements. However, the cross_match data along with the length of the inconsistent reads implied that the repeated region most likely represented the same sequence.

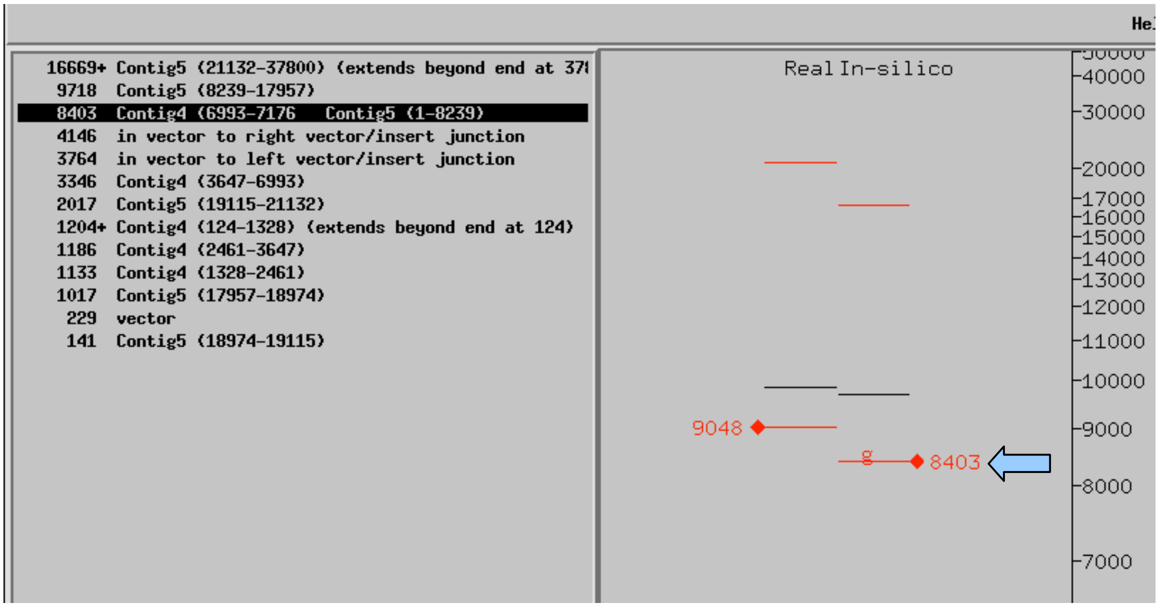


```
copy                    copy                      size  similarity
  1                       2

Contig4      498-860 to Contig5     4123-4496 (not comp) 373 93.1
Contig4    3938-4124 to Contig5         8-194 (not comp) 186 95.7
Contig4    4605-7173 to Contig5      277-2851 (not comp) 2574 94.3
```

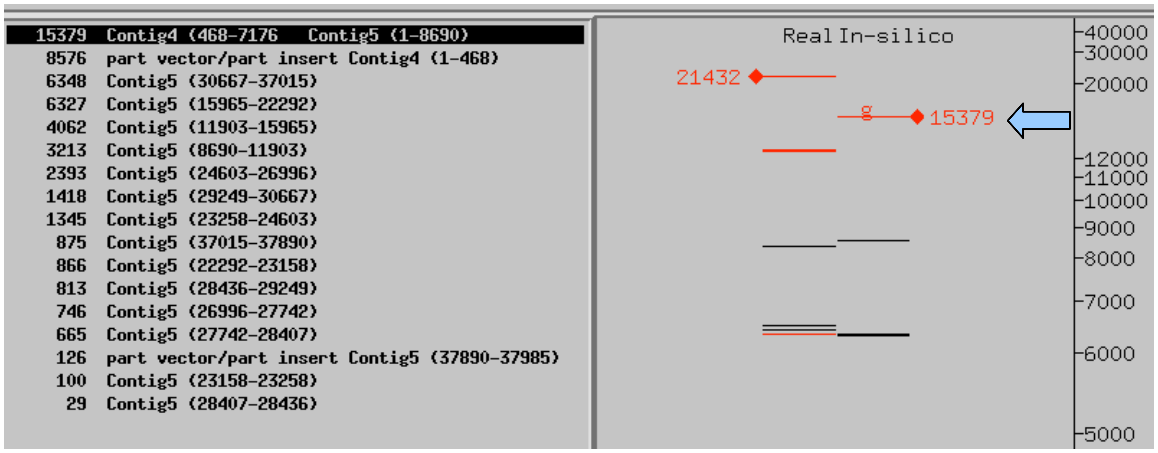
Figure 2. Cross_match sequence matches show repeats across the gap. Boxed entry is the long repeat region of interest.

In order to solidify whether or not the gap truly existed, I used an *in silico* restriction digest to compare the size of fragments. However, when the digest tested for end-to-end contigs, the results were mixed among the different restriction enzymes—the EcoRV digest found a 645 base pair gap; the EcoRI digest found a 6053 bp gap; and the HindIII and the SacI digests found gap sizes greater than 3 kb or none at all (Figures 3 and 4). As such, because the predicted gap size varied so much among all the digests, these initial digests were inconclusive as to whether or not the gap existed.



```
16669+ Contig5 (21132-37800) (extends beyond end at 37
 9718  Contig5 (8239-17957)
 8403  Contig4 (6993-7176   Contig5 (1-8239)
 4146  in vector to right vector/insert junction
 3764  in vector to left vector/insert junction
 3346  Contig4 (3647-6993)
 2017  Contig5 (19115-21132)
 1204+ Contig4 (124-1328) (extends beyond end at 124)
 1186  Contig4 (2461-3647)
 1133  Contig4 (1328-2461)
 1017  Contig5 (17957-18974)
  229  vector
  141  Contig5 (18974-19115)
```

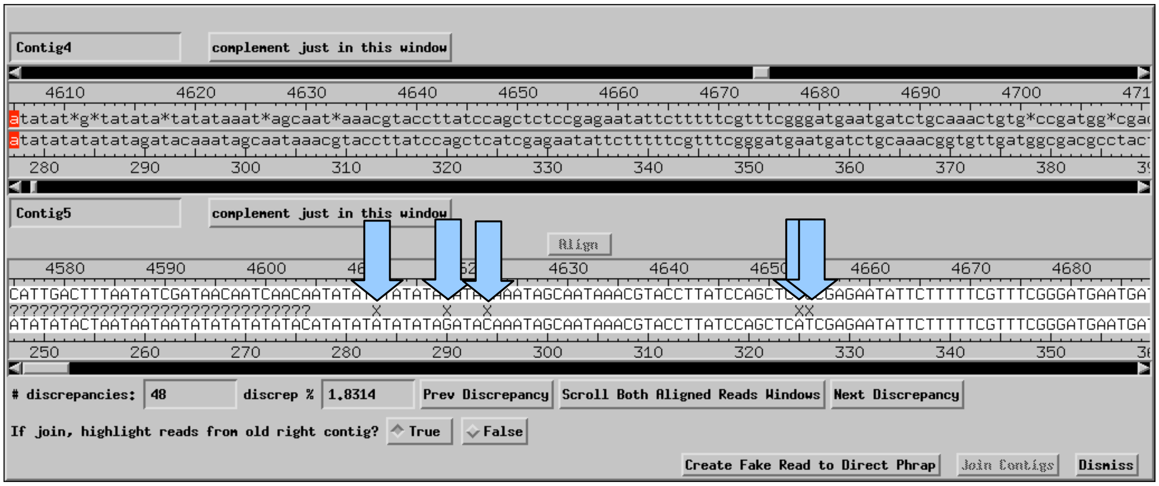
Figure 3. Representative EcoRV digest of end-to-end alignment using fragSizes.txt showing an approximate 6 kilobase gap. Arrow marks the fragment with the gap.



```
15379  Contig4 (468-7176   Contig5 (1-8690)
 8576  part vector/part insert Contig4 (1-468)
 6348  Contig5 (30667-37015)
 6327  Contig5 (15965-22292)
 4062  Contig5 (11903-15965)
 3213  Contig5 (8690-11903)
 2393  Contig5 (24603-26996)
 1418  Contig5 (29249-30667)
 1345  Contig5 (23258-24603)
  875  Contig5 (37015-37890)
  866  Contig5 (22292-23158)
  813  Contig5 (28436-29249)
  746  Contig5 (26996-27742)
  665  Contig5 (27742-28407)
  126  part vector/part insert Contig5 (37890-37985)
  100  Contig5 (23158-23258)
   29  Contig5 (28407-28436)
```

Figure 4. Representative EcoRI digest of end-to-end alignment using fragSizes.txt showing an approximate 6 kb gap. Arrow marks the fragment with the gap.

Cross_match found 48 discrepancies (1.8314%) across the bases in the tandem repeat regions. These discrepancies and the small bit of unaligned read (question marks in Figure 5) suggest that there were still some differences between the tandem repeat regions. However, because this region only had one read in contig 5 and many matching reads in contig 4, the differences could not be completely trusted because they were based only on the single read in contig 5; this read simply may have been misplaced. Despite these differences, the cross_match results were compelling enough for me to initially ignore the errors, force join contigs 4 and 5 together (Figures 5 and 6), and then examine the digests again to see if they would improve. At the time, the force join seemed to be my best direction and later on this would prove to be the right choice. Figures 7 and 8 show much more consistent results among the various digests. In order to better confirm the force join and check the sequence, I ordered two primers within the overlap region. At this point, I needed to confirm the existence of the gap and, if the force join was correct, the sequence of the clone within that same region.



```
Contig4   complement just in this window
   4610      4620      4630      4640      4650      4660      4670      4680      4690      4700      471
atatat*g*tatata*tatataaat*agcaat*aaacgtaccttatccagctctccgagaatattctttttcgtttcgggatgaatgatctgcaaactgtg*ccgatgg*cga
atatatatatatagatacaaatagcaataaacgtaccttatccagctcatcgagaatattctttttcgtttcgggatgaatgatctgcaaacggtgttgatggcgacgcctac
   280       290       300       310       320       330       340       350       360       370       380       3
Contig5   complement just in this window
                                                   Align
   4580      4590      4600      46        62        4630      4640      4650      4660      4670      4680
CATTGACTTTAATATCGATAACAATCAACAATATA  ATATA  AT  AATAGCAATAAACGTACCTTATCCAGCTC  GAGAATATTCTTTTTCGTTTCGGGATGAATGA
??????????????????????????????     X     X  X                           XX
ATATATACTAATAATAATATATATATATACATATATATATATAGATACAAATAGCAATAAACGTACCTTATCCAGCTCATCGAGAATATTCTTTTTCGTTTCGGGATGAATGA
   250       260       270       280       290       300       310       320       330       340       350       3
# discrepancies: 48   discrep % 1.8314   Prev Discrepancy   Scroll Both Aligned Reads Windows   Next Discrepancy
If join, highlight reads from old right contig?  ◇ True  ◇ False
                                   Create Fake Read to Direct Phrap   Join Contigs   Dismiss
```

Figure 5. Alignment before force join of contigs 4 and 5. Arrows show representative mismatches. Question marks show unmatched/unaligned sequence that was initially ignored because the region was covered by only one read in contig 5. This single read may have been chimeric or misplaced.

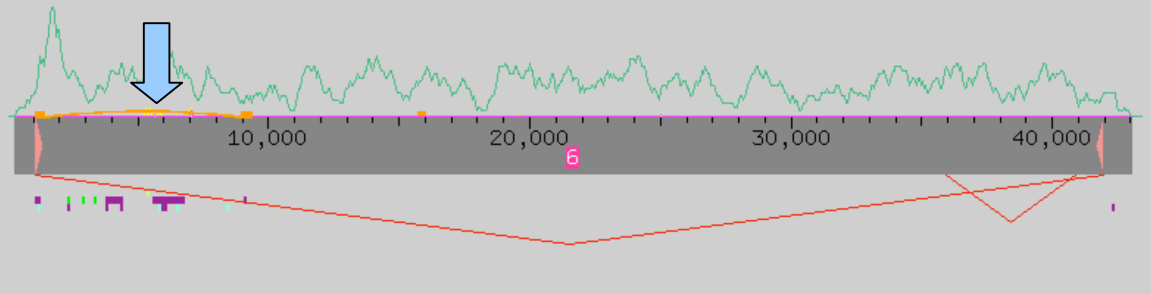
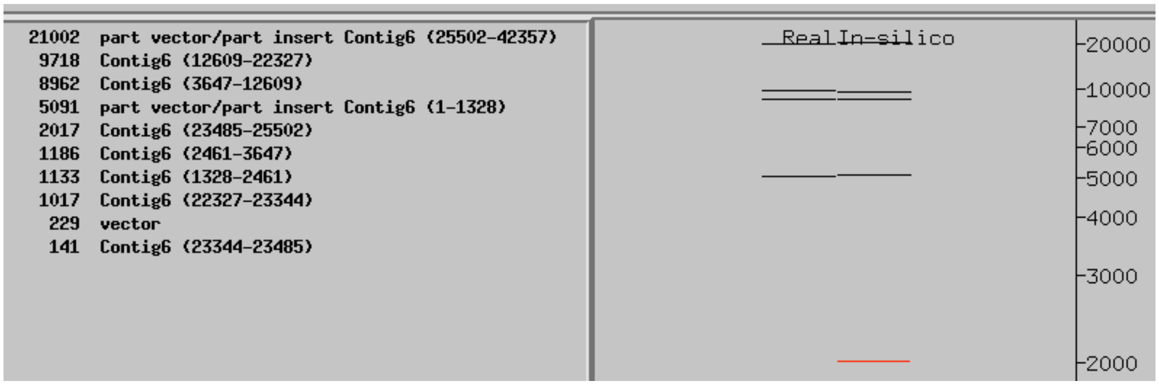
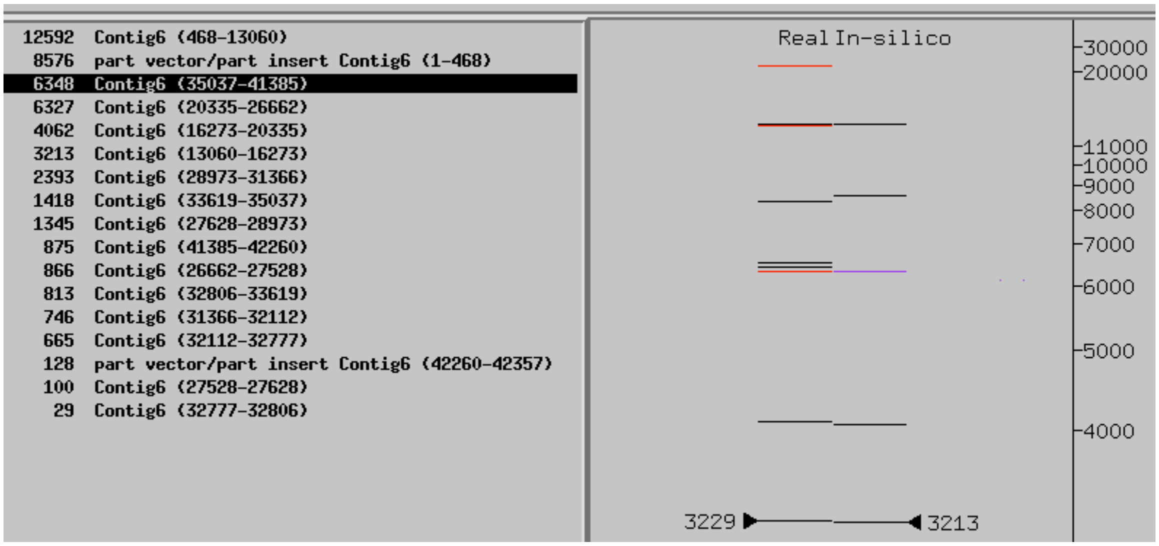


Figure 6. Assembly view after force join of contigs 4 and 5 to make contig 6. Arrow points to high level of mismatches. Clone ends and the other read are consistent. Clone ends become redefined after force join and inconsistent read is within error and only 131 bases longer than library limit.



```
21002  part vector/part insert Contig6 (25502-42357)
 9718  Contig6 (12609-22327)
 8962  Contig6 (3647-12609)
 5091  part vector/part insert Contig6 (1-1328)
 2017  Contig6 (23485-25502)
 1186  Contig6 (2461-3647)
 1133  Contig6 (1328-2461)
 1017  Contig6 (22327-23344)
  229  vector
  141  Contig6 (23344-23485)
```

Figure 7. Representative EcoRV digest using fragSizes.txt in contig 6 (after force join).



```
12592  Contig6 (468-13060)
 8576  part vector/part insert Contig6 (1-468)
 6348  Contig6 (35037-41385)
 6327  Contig6 (20335-26662)
 4062  Contig6 (16273-20335)
 3213  Contig6 (13060-16273)
 2393  Contig6 (28973-31366)
 1418  Contig6 (33619-35037)
 1345  Contig6 (27628-28973)
  875  Contig6 (41385-42260)
  866  Contig6 (26662-27528)
  813  Contig6 (32806-33619)
  746  Contig6 (31366-32112)
  665  Contig6 (32112-32777)
  128  part vector/part insert Contig6 (42260-42357)
  100  Contig6 (27528-27628)
   29  Contig6 (32777-32806)
```

Figure 8. Representative EcoRI digest using fragSizes.txt in contig 6 (after force join).

Autofinish did not call any reads for my project either prior to complementing contig 4 or after force joining contigs 4 and 5. Table 1 shows the reads that Autofinish called after reverse complementing contig 4. These differ slightly from my reads (Table 2). Autofinish's calls made the assumption that a gap existed and approached the ends of contigs 4 and 5 as areas of low coverage. My calls, on the other hand, focused on confirming not only the existence of the gap but also the sequence of the clone if the force join were correct. This approach differs from Autofinish's because I specifically chose primers that were located *inside* the repeat region that would simultaneously report both results. A prior search for string for the oligos showed that they were unique to the potential terminal repeat region. Because the unlikeliness of clean sequencing on a clone with multiple sites for oligo binding (i.e., repeated sequence), the successful sequencing of the clone at this region would suggest that there is no terminal repeat and, with the same logic, no gap. The primers I called were as close as possible to the mismatched bases in the cross_match alignment (Figure 5). As a result of the difference in rationale, Autofinish seemed to choose safer, cheaper oligos flanking the region. I found some of the oligos that Autofinish chose unusable due to the uncertainty of the region from low quality and opted for custom oligos.

Contig	**Start**	**End**	**Dir.**	**Oligo**	**Reason**
Contig 4	-741	195	<=	Ctgctctgatgccgc	Cover low quality
Contig 4	6035	6971	=>	Gcctcctacttgtgtaccattat	Cover single strand
Contig 4	6401	7337	=>	Agtatggaaacaacgttgtagg	Cover single strand
Contig 4	7096	8032	=>	Cgtcaaacggttgcttttc	Cover single strand
Contig 5	-823	113	<=	Ccaaccaaacccagatgtc	Cover low quality
Contig 5	37647	38583	=>	Aaacggcatgatgaacct	Cover single strand

Table 1. Autofinish calls after complementing contig 4.

Contig	Start End	Dir.	Oligo	Chemistry	Reason
Contig 4	5674 5690	=>	ggtggcgaggtccctac	BigDye, 4:1, dGTP	Confirm force join
Contig 5	1874 1894	<=	agcagttgtcaattgaatacg	BigDye, 4:1, dGTP	Confirm force join

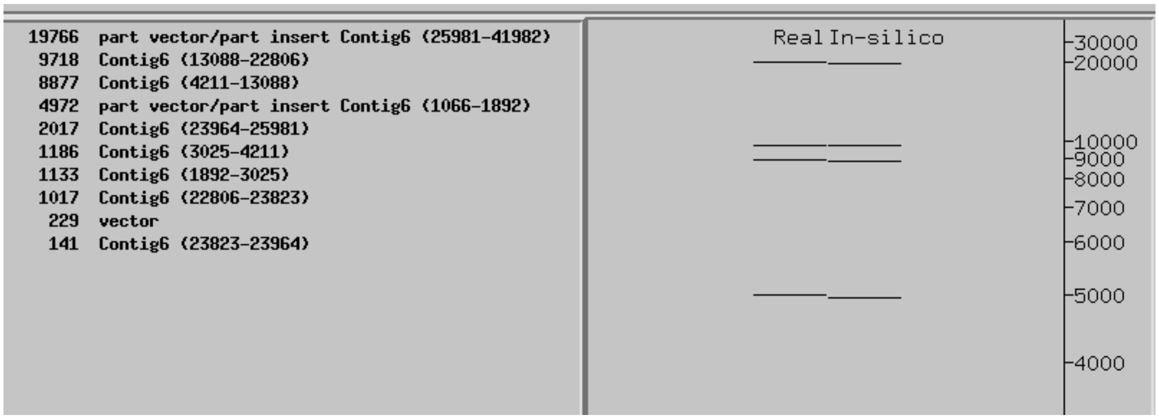
Table 2. My first round of calls.

Because of sequencing problems at the Genome Center, I called a second round of reads to cover roughly the same area in potential force joined region. I shifted the target area to confirm in a second area the real sequence of the clone. The second round of calls was made from contig 6, the product of the force join of contigs 4 and 5, to better estimate what the expected sequence would be if the force join were correct. All three chemistries were used in both rounds because repetitious regions are typically difficult to sequence. All three chemistries from the four oligos eventually incorporated into the assembly. From the previously described rationale, their successful incorporation indirectly confirmed the correctness of the force join (and nonexistence of the gap) as well as the correct sequence of the clone in the force join region. These reads reconfirmed that the sequence of the fosmid was the same as those reads that were previously called by another student who worked on this clone.

Contig	Start End	Direction	Oligo	Chemistry	Reason
Contig 6	5235 5254	=>	Gtttcgcaaacgccg	BigDye, 4:1, dGTP	Confirm force join
Contig 6	7076 7090	<=	Cgtttcgggatgaatgatct	BigDye, 4:1, dGTP	Confirm force join

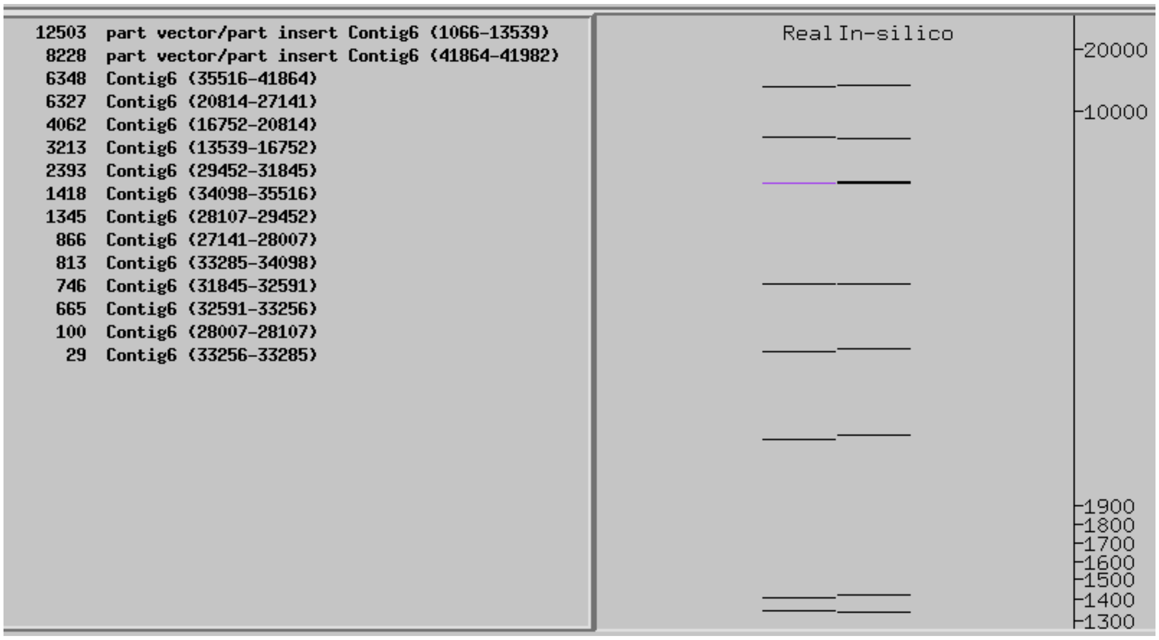
Table 3. My second round of calls.

During this time, several new sets of restriction digest data became available. These new digests were better quality than the initial digests and confirmed the force join through correct fragment sizes for all the digests after complementing the vector (Figures 9 and 10).



```
19766  part vector/part insert Contig6 (25981-41982)
 9718  Contig6 (13088-22806)
 8877  Contig6 (4211-13088)
 4972  part vector/part insert Contig6 (1066-1892)
 2017  Contig6 (23964-25981)
 1186  Contig6 (3025-4211)
 1133  Contig6 (1892-3025)
 1017  Contig6 (22806-23823)
  229  vector
  141  Contig6 (23823-23964)
```

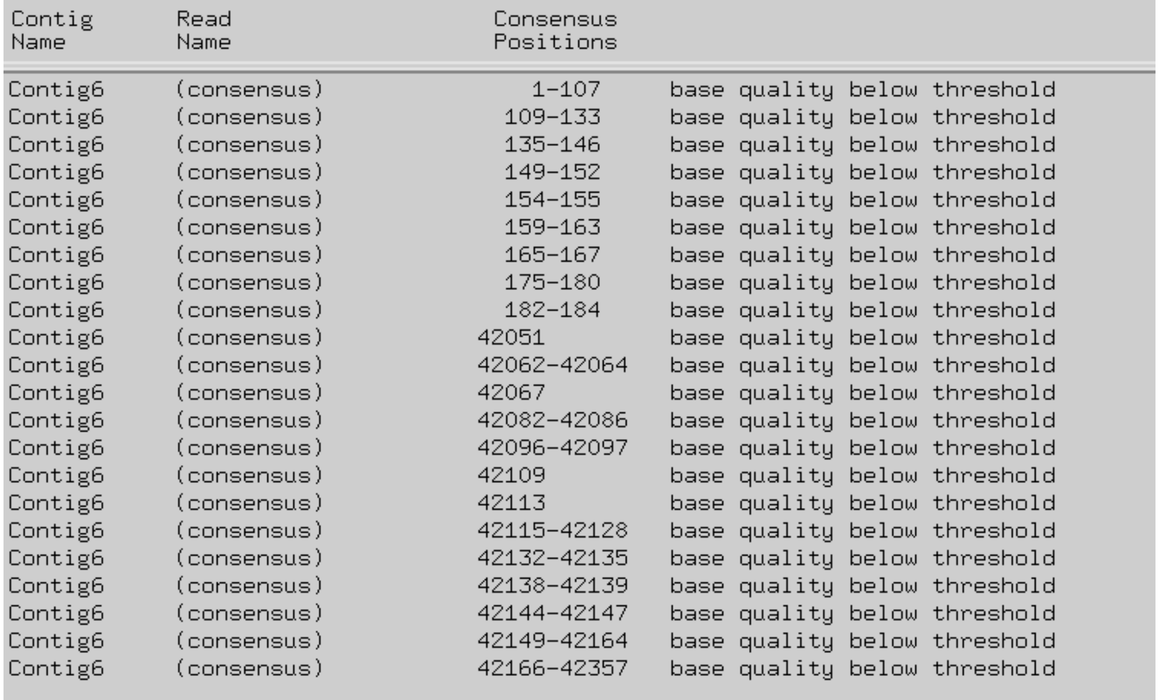
Figure 9. Representative EcoRV digest with fragSizesc.txt in contig 6 after complementing vector (a separate digest from fragSizes.txt).



```
12503  part vector/part insert Contig6 (1066-13539)
 8228  part vector/part insert Contig6 (41864-41982)
 6348  Contig6 (35516-41864)
 6327  Contig6 (20814-27141)
 4062  Contig6 (16752-20814)
 3213  Contig6 (13539-16752)
 2393  Contig6 (29452-31845)
 1418  Contig6 (34098-35516)
 1345  Contig6 (28107-29452)
  866  Contig6 (27141-28007)
  813  Contig6 (33285-34098)
  746  Contig6 (31845-32591)
  665  Contig6 (32591-33256)
  100  Contig6 (28007-28107)
   29  Contig6 (33256-33285)
```

Figure 10. Representative EcoRI digest with fragSizesc.txt in contig 6 after complementing vector (a separate digest from fragSizes.txt).

Throughout these two rounds of sequencing, I also attempted to resolve the low quality regions and high quality discrepancies found throughout the clone. After force joining the two contigs, the navigators showed 22 low quality regions (Figure 11) that all fell outside of the clone. I tagged these and ignored them.

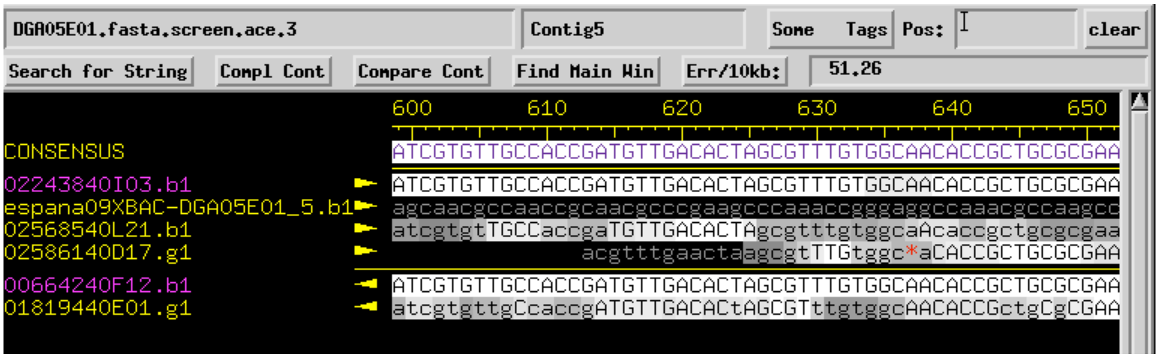


Contig Name	Read Name	Consensus Positions	
Contig6	(consensus)	1-107	base quality below threshold
Contig6	(consensus)	109-133	base quality below threshold
Contig6	(consensus)	135-146	base quality below threshold
Contig6	(consensus)	149-152	base quality below threshold
Contig6	(consensus)	154-155	base quality below threshold
Contig6	(consensus)	159-163	base quality below threshold
Contig6	(consensus)	165-167	base quality below threshold
Contig6	(consensus)	175-180	base quality below threshold
Contig6	(consensus)	182-184	base quality below threshold
Contig6	(consensus)	42051	base quality below threshold
Contig6	(consensus)	42062-42064	base quality below threshold
Contig6	(consensus)	42067	base quality below threshold
Contig6	(consensus)	42082-42086	base quality below threshold
Contig6	(consensus)	42096-42097	base quality below threshold
Contig6	(consensus)	42109	base quality below threshold
Contig6	(consensus)	42113	base quality below threshold
Contig6	(consensus)	42115-42128	base quality below threshold
Contig6	(consensus)	42132-42135	base quality below threshold
Contig6	(consensus)	42138-42139	base quality below threshold
Contig6	(consensus)	42144-42147	base quality below threshold
Contig6	(consensus)	42149-42164	base quality below threshold
Contig6	(consensus)	42166-42357	base quality below threshold

Figure 11. The 22 low quality read regions in contig 6 have a quality threshold under 30. Areas meeting single stranded quality requirements (quality of 30) will also meet quality standards for double stranded DNA (quality of 25).

Figure 6 shows a sudden increase of high quality discrepancies (yellow lines on graph) around where the force join occurred between base pairs 5000 and 7500. As seen in Figure 5, cross_match found 48 mismatches over the force join region between contig 4 and contig 5's consensus sequences. This number translated to 226 high quality discrepancies once I force joined the two contigs. Looking back at each contig individually, the navigators showed nine high quality discrepancies in contig 5 and only one in contig 4. The great increase in high quality discrepancies seemed to come from Consed mistakenly assigning the consensus value at specific bases to certain divergent reads and assigning the remaining reads as discrepant. As such, when Consed force joins two contigs, it has the potential to mix the two consensus sequences of the regions together. This results in a hybrid sequence as the final join.

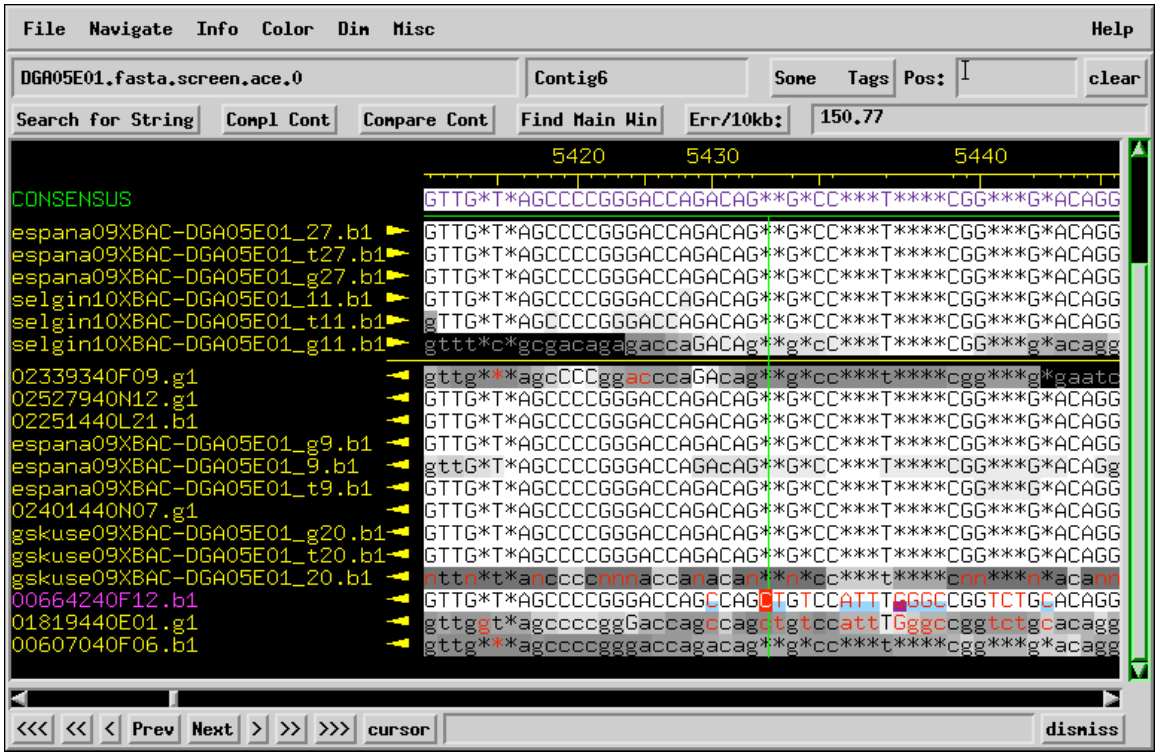
Upon closer examination, most of the high quality discrepancies arose because of only a few reads that were critical in determining the consensus in their respective contigs prior to the force join. In both cases, there were few reads at the ends and the consensus was based on a few high quality reads, even if they would later be discrepant. For example, reads 00664240F12.b1, 02586140D17.g1, and 01819440E01.g1 were critical at the start of contig 5 and read dgu27p03.g1 was important at the end of contig 4 for connecting the two contigs together (Figure 12). These reads also were the most discrepant in contig 6 when Consed relied heavily on all four for consensus base calls. As a result, most of these discrepancies were solved by relying on other high quality reads, especially reads called by previous students who worked on this clone (Figure 13). These mismatches were then labeled as a possible polymorphism (if > 3 high quality reads matched), possible growth difference, or possible misassembly depending on the quality of the reads at the base. All other high quality discrepancies were classified as compressions, expansions, or any of the previously listed categories.



```
DGA05E01.fasta.screen.ace.3    Contig5
Err/10kb: 51.26

                               600       610       620       630       640       650
CONSENSUS                      ATCGTGTTGCCACCGATGTTGACACTAGCGTTTGTGGCAACACCGCTGCGCGAA
02243840I03.b1               ► ATCGTGTTGCCACCGATGTTGACACTAGCGTTTGTGGCAACACCGCTGCGCGAA
espana09XBAC-DGA05E01_5.b1   ► agcaacgccaaccgcaacgcccgaagcccaaaccgggaggccaaacgccaagcc
02568540L21.b1               ► atcgtgtTGCCaccgaTGTTGACACTAgcgtttgtggcaAcaccgctgcgcgaa
02586140D17.g1               ►               acgtttgaactaagcgtTTGtggc*aCACCGCTGCGCGAA
00664240F12.b1               ◄ ATCGTGTTGCCACCGATGTTGACACTAGCGTTTGTGGCAACACCGCTGCGCGAA
01819440E01.g1               ◄ atcgtgttgCcaccgATGTTGACACtAGCGTttgtggcAACACCGctgCgCGAA
```

Figure 12. Close to the start of contig 5 prior to force join. Reads 00664240F12.b1 and 02586140D17.g1 are in purple and are the only high quality reads at the start. Consed relies on these reads to connect contig 5 with contig 4, which also only has a few reads at the end.



```
DGA05E01.fasta.screen.ace.0    Contig6
Err/10kb: 150.77

                                      5420          5430                5440
CONSENSUS                        GTTG*T*AGCCCCGGGACCAGACAG**G*CC***T****CGG***G*ACAGG
espana09XBAC-DGA05E01_27.b1    ► GTTG*T*AGCCCCGGGACCAGACAG**G*CC***T****CGG***G*ACAGG
espana09XBAC-DGA05E01_t27.b1   ► GTTG*T*AGCCCCGGGACCAGACAG**G*CC***T****CGG***G*ACAGG
espana09XBAC-DGA05E01_g27.b1   ► GTTG*T*AGCCCCGGGACCAGACAG**G*CC***T****CGG***G*ACAGG
selgin10XBAC-DGA05E01_11.b1    ► GTTG*T*AGCCCCGGGACCAGACAG**G*CC***T****CGG***G*ACAGG
selgin10XBAC-DGA05E01_t11.b1   ► gTTG*T*AGCCCCGGGACCAGACAG**G*CC***T****CGG***G*ACAGG
selgin10XBAC-DGA05E01_g11.b1   ► gttt*c*gcgacagagaccaGACAg**g*cC***T****CGG***g*acagg
02339340F09.g1                 ◄ gttg***agcCCCggacccaGAcag**g*cc***t****cgg***g*gaatc
02527940N12.g1                 ◄ GTTG*T*AGCCCCGGGACCAGACAG**G*CC***T****CGG***G*ACAGG
02251440L21.b1                 ◄ GTTG*T*AGCCCCGGGACCAGACAG**G*CC***T****CGG***G*ACAGG
espana09XBAC-DGA05E01_g9.b1    ◄ GTTG*T*AGCCCCGGGACCAGACAG**G*CC***T****CGG***G*ACAGG
espana09XBAC-DGA05E01_9.b1     ◄ gttG*T*AGCCCCGGGACCAGAcAG**G*CC***T****CGG***G*ACAGg
espana09XBAC-DGA05E01_t9.b1    ◄ GTTG*T*AGCCCCGGGACCAGACAG**G*CC***T****CGG***G*ACAGG
02401440N07.g1                 ◄ GTTG*T*AGCCCCGGGACCAGACAG**G*CC***T****CGG***G*ACAGG
gskuse09XBAC-DGA05E01_g20.b1   ◄ GTTG*T*AGCCCCGGGACCAGACAG**G*CC***T****CGG***G*ACAGG
gskuse09XBAC-DGA05E01_t20.b1   ◄ GTTG*T*AGCCCCGGGACCAGACAG**G*CC***T****CGG***G*ACAGG
gskuse09XBAC-DGA05E01_20.b1    ◄ nttn*t*ancccnnnaccanacan**n*cc***t****cnn***n*acann
00664240F12.b1                 ◄ GTTG*T*AGCCCCGGGACCAGCCAGCTGTCCATTTGGGCCGGTCTGCACAGG
01819440E01.g1                 ◄ gttggt*agccccggGaccagccagctgtccattTGggccggtctgcacagg
00607040F06.b1                 ◄ gttg***agccccgggaccagacag**g*cc***t****cgg***g*acagg
```

Figure 13. Arrows point to discrepant bases in read 00664240F12.b1 (one of reads highlighted in Figure 12) when compared to the consensus after force join in contig 6. Note that read 01819440E01.g1 has the same mismatches here (same matches in figure 12) but is considered low quality so was not listed in the high quality discrepancies navigator. Consensus was changed to reflect the actual clone based on previously called reads on the same clone (reads with DGA05E01).

For other high quality discrepancy regions, the reads were too divergent from the actual consensus. Reads espana09XBAC-DGA05E01_5.b1, 00664240F12.b1, and 02568540L21.b1 all differed significantly from the consensus at periodic short intervals as well as had long unaligned regions. By taking them out of the assembly, I was able to eliminate almost 30 different mismatches.

After tagging all the high quality discrepancies and making sure the digests matched in size, I submitted the consensus sequence to BLAST to check for any microbial or vector DNA (Figure 14). BLAST analysis showed that there was significant sequence that matched microbial DNA at the ends of the fosmid. These matches all occurred outside of the clone ends. I tagged these regions as possible vector DNA.

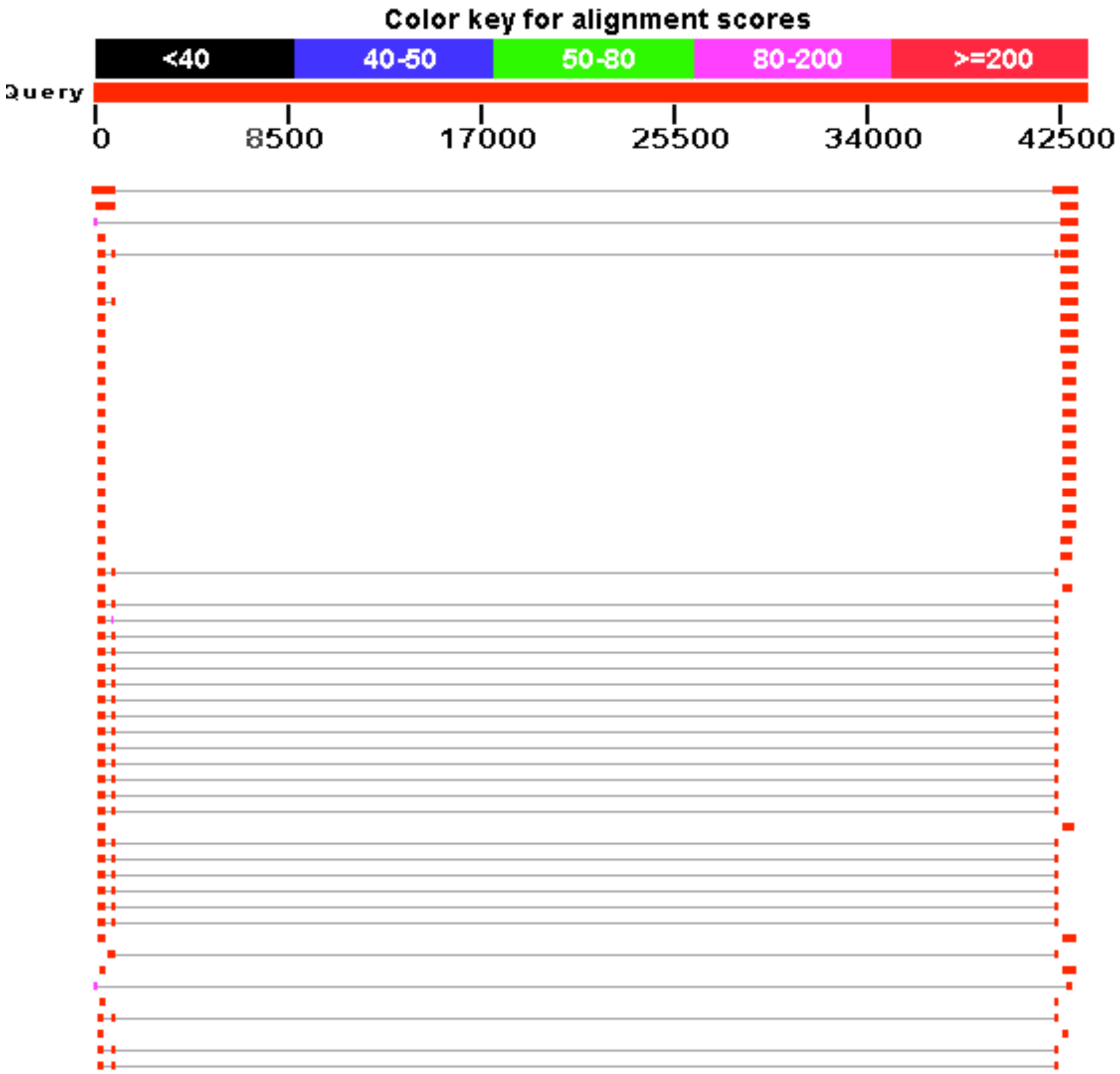


Figure 14. BLAST analysis of consensus sequence against microbial DNA. Red areas show matches at the ends of the clone.

After checking in all the navigators, the only areas of single strand or subclone were outside the clone ends. Single chemistry regions all had high quality levels that matched among many reads. Finally, there were no ambiguous vector or unknown bases in the consensus. There was one mononucleotide run of 15 As at base pair 17875 to 17889. This run, however, is supported by four reads of high quality that were manually verified to show correct base calling in and around the run. This run was later tagged and recorded.

In conclusion, Figure 6 represents the final assembly for this project. I was able to combine a project with two contigs into one contig. The high quality discrepancies were all labeled, despite having a high count between base pair 5000 and 7500. The digests show proper length and the called reads reconfirmed the past called reads for the fosmid. This work will continue through annotation.