Louis Lo
Finishing paper
14P24
April 10, 2006

Finishing fosmid 14p24

Chromosome four of the *Drosophila* species *melanogaster* and *virilis* is of interest to investigators because the first is mostly heterochromatic while the second is mostly euchromatic. Obtaining high quality sequences for both species will enable comparative analysis of chromosome four, providing sequence information about the basis of gene packaging. My project contributes to the final sequence data of *D. virilis* by finishing fosmid 14p24.

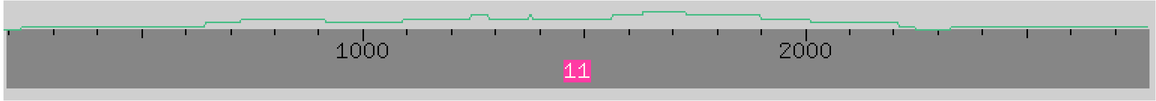


Figure 1. Initial assembly view with 96 reads

Figure 1 shows the initial assembly of fosmid 14p24. This view represents the information obtained from a 96 well plate. The contig is only 3kb out of the expected 40kb; because this is an initial assembly with few reads, Phred/Phrap was not able to assemble the entire 40kb. Since this assembly is a preliminary view of my project, I am not too concerned that so little data has been assembled.

Figure 2 shows the initial assembly view after Phred/Phrap was given the 956 reads produced by the production pipeline. The main contig of the assembly is 39kb long. From this view we see no major problems with Phred/Phrap's assembly; there are no gaps, and given its length, the main contig likely represents the entire insert.

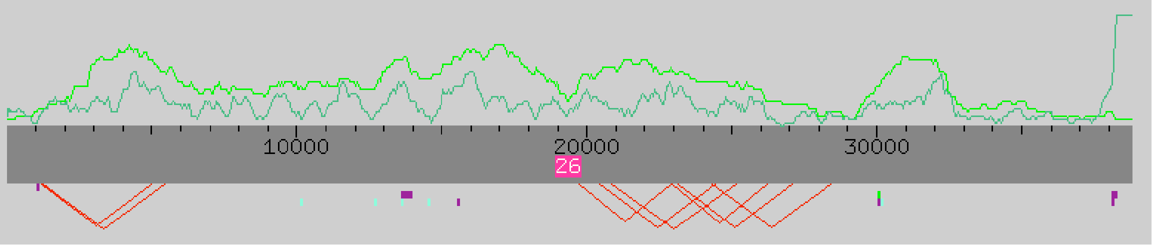


Figure 2. Initial assembly view with full production reads

The main problems that I had in my assembly were a large region of low quality consensus, two large regions that were labeled *high quality match elsewhere* at 38103-38292 and 13650-14005, and various large regions covered by only one strand or chemistry. The minor problems included some inconsistent forward/reverse reads, small regions of low quality consensus and small regions with high quality discrepancies.

With my first round of reads I attempted to resolve the large low quality region. It spanned from 26653 to 26954. Much of this region is also only covered by reads in one direction. So I called one read using all three chemistries in the other direction (Table 1). Autofinish also called a read to address this issue. Because the Autofinish report was generated from an earlier assembly, it states that the read is on contig 8, although the template XBAA-aad92g06 corresponds to the same low quality region that I am trying to resolve on the main contig. A search for the oligo sequence used by Autofinish reveals that Autofinish called a read starting from position 26,431 on the main contig. I called a read from position 26,523. Because the target region was only 300 bp's long, Autofinish's oligo would be as effective as mine, but my strategy would have better success if the reads were shorter than the expected 1kb.

Table 1. Reads called for first round of reactions.

Oligo	Sequence	Contig	Direction	Template	Goal
1	**cgcttgctatcgtaatcgg**	**26(main)**	**-->**	**XBAA-aaf75g02**	Resolve Low Quality
Autofinish	**cagtcgttcttcgatgtctgt**	**8**	**-->**	**XBAA-aad92g06**	Resolve Low Quality

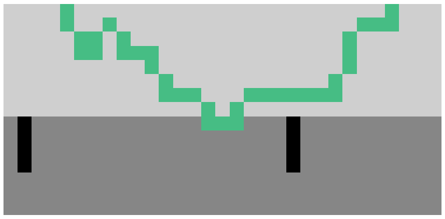


Figure 3. Zoomed Assembly View of low quality region *before* first round of reads

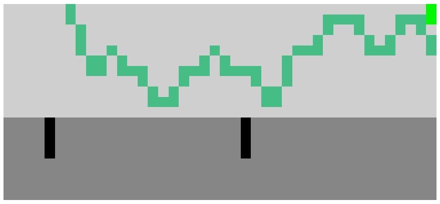
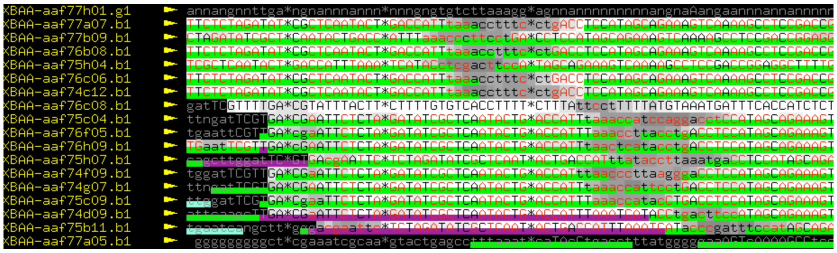


Figure 4. Zoomed Assembly View of low quality region *after* first round of reads

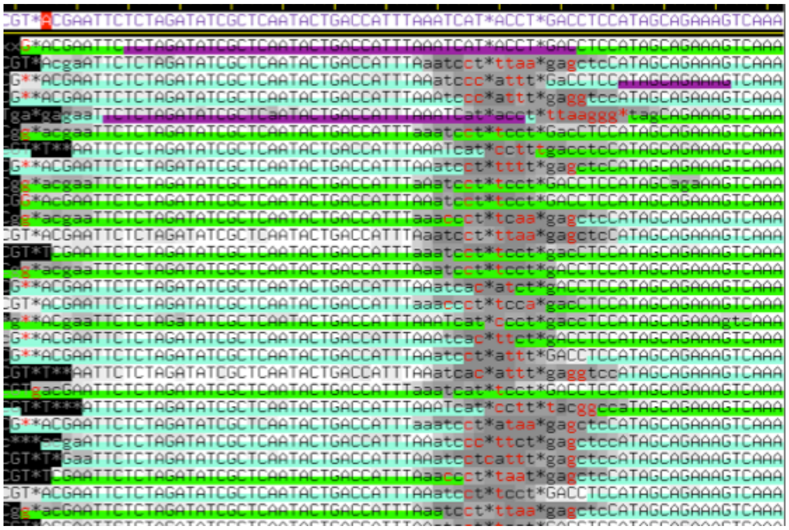
Figure 3 shows a "zoomed in" picture from Assembly View of the low quality region before the first round of reads. There is a portion where there is no read depth at all. With the first round of reads I was able to resolve this region (Figure 4). As I only targeted one region with my first round of reactions, all the problems that were present before were still present after the first round, with the exception of this large region of low quality reads.

The next problems I tried to resolve were the two regions that were labeled *high quality match elsewhere*. The first of these regions spanned from 38104-38348; Figure 5 shows a portion of this region.



```
XBAA-aaf77h01.g1   ►  annangnnttga*ngnannnannn*nnngngtgtcttaaagg*agnnannnnnnnnnangnaAangaannnannannnnn
XBAA-aaf77a07.b1   ►  TTCTCTAGATAT*CGCTCAATACT*GACCATTtaaacctttc*ctgACCTCCATAGCAGAAAGTCAAAGCCTCCGACCG
XBAA-aaf77b09.b1   ►  CTAGATATCGCT*CAATACTGACC*ATTTaaacccttcctGA*CCTCCATAGCAGAAAGTCAAAGCCTCCGACCGGAGG
XBAA-aaf76b08.b1   ►  TTCTCTAGATAT*CGCTCAATACT*GACCATTtaaacctttc*ctGACCTCCATAGCAGAAAGTCAAAGCCTCCGACCG
XBAA-aaf75h04.b1   ►  TCGCTCAATACT*GACCATTTAAA*TCATACctcgacttccA*TAGCAGAAAGTCAAAAGCCTCCGACCGGAGGCTTTTG
XBAA-aaf76c06.b1   ►  TTCTCTAGATAT*CGCTCAATACT*GACCATTtaaacctttc*ctGACCTCCATAGCAGAAAGTCAAAGCCTCCGACCG
XBAA-aaf74c12.b1   ►  TTCTCTAGATAT*CGCTCAATACT*GACCATTtaaacctttc*ctgACCTCCATAGCAGAAAGTCAAAGCCTCCGACCG
XBAA-aaf76c08.b1   ►  gatTCGTTTTGA*CGTATTTACTT*CTTTTGTGTCACCTTTT*CTTTAttcctTTTTATGTAAATGATTTCACCATCTCT
XBAA-aaf75c04.b1   ►  ttngatTCGTGA*CGAATTCTCTA*GATATCGCTCAATACTG*ACCATTtaaaccatccaggacctCCATAGCAGAAAGT
XBAA-aaf76f05.b1   ►  tgaattCGTTGA*cgaATTCTCTA*GATATCGCTCAATACTG*ACCATTTaaaccttacctgACCTCCATAGCAGAAAGT
XBAA-aaf76h09.b1   ►  TGaatTCGTTGA*cGAATTCTCTA*GATATCGCTCAATACTG*ACCATTtaactcatacctgACCTCCATAGCAGAAAGT
XBAA-aaf75h07.b1   ►  cagcttggatTC*GTGAcgAATTC*TCTAGATATCGCTCAAT*ACTGACCATttataccttaaatgaCCTCCATAGCAGA
XBAA-aaf74f09.b1   ►  tggatTCGTTGA*CGaATTCTCTA*GATATCGCTCAATACTG*ACCATttaacccttaagggaCCTCCATAGCAGAAAGT
XBAA-aaf74g07.b1   ►  ttngatTCGTGA*CGAATTCTCTA*GATATCGCTCAATACTG*ACCATTtaaaccattcctGACCTCCATAGCAGAAAGT
XBAA-aaf75c09.b1   ►  ttggatTCGTGA*CgaaTTCTCTA*GATATCGCTCAATACTG*ACCATTTaaaccatacCTGACCTCCATAGCAGAAAGT
XBAA-aaf74d09.b1   ►  attcaagcTTGA*CGaATTCTCTA*GATATCGCTCAATACTG*ACCATTTAAATCATACCtgacttccATAGCAGAAAGT
XBAA-aaf75b11.b1   ►  tgaatcangctt*gggacgaattc*TCTAGATATCGCTCAAT*ACTGACCATTTAAATCATacccgatttccatAGCAGA
XBAA-aaf77a05.b1   ►   ggggggggggct*cgaaatcgcaa*gtactgagcctttaaat*caTAcCtgacctttatgggggaaAGTcAAAAGCCtcc
```

Figure 5. Aligned Reads view of misaligned high quality region (above) and corresponding region found using Search (below).

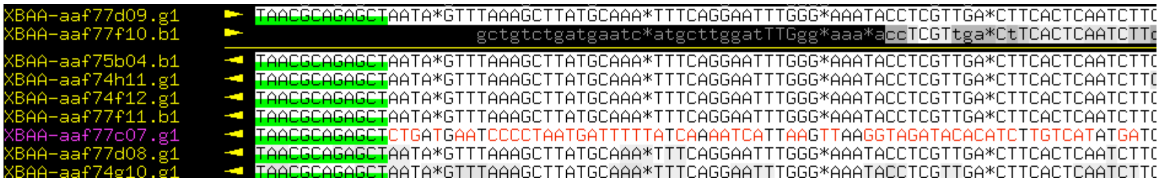


```
CGT*ACGAATTCTCTAGATATCGCTCAATACTGACCATTTAAATCAT*ACCT*GACCTCCATAGCAGAAAGTCAAA

 G*ACGAATTCTCTAGATATCGCTCAATACTGACCATTTAAATCAT*ACCT*GACCTCCATAGCAGAAAGTCAAA
CGT*AcgaATTCTCTAGATATCGCTCAATACTGACCATTTAaatcct*ttaa*gagctcCATAGCAGAAAGTCAAA
 G**ACGAATTCTCTAGATATCGCTCAATACTGACCATTTAAatccc*attt*GaCCTCCATAGCAGAAAGTCAAA
 G**ACGAATTCTCTAGATATCGCTCAATACTGACCATTTAAAtccc*attt*gaggtccATAGCAGAAAGTCAAA
Tga*gagaaTTCTCTAGATATCGCTCaATACTGACCATTTAAATCat*acct*ttaaggg*tagCAGAAAGTCAAA
Cgg*acgaaTTCTCTAGATATCGCTCAATACTGACCATTTaaatcct*tcct*GacCTCCATAGCAGAAAGTCAAA
CGT*T**AATTCTCTAGATATCGCTCAATACTGACCATTTAAATcat*ccttgacctcCATAGCAGAAAGTCAAA
 G**ACGAATTCTCTAGATATCGCTCAATACTGACCATTTAAatcct*tttt*gagctcCATAGCAGAAAGTCAAA
Cgg*acgaaTTCTCTAGATATCGCTCAATACTGACCATTTaAatcct*tcct*GACCTCCATAGCagaAAGTCAAA
CGG*AcGAATTCTCTAGATATCGCTCAATACTGACCATTTAAatcct*tcct*GACCTCCATAGCAGAAAGTCAAA
Cgg*acgaaTTCTCTAGATATCGCTCAATACTGACCATTTaaaccct*tcaa*gagctcCATAGCAGAAAGTCAAA
CGT*ACGAATTCTCTAGATATCGCTCAATACTGACCATTTAaatcct*ttaa*gagctcCATAGCAGAAAGTCAAA
CGT*TCGAATTCTCTAGATATCGCTCAATACTGACCATTTaaatcct*tcct*gacCTCCATAGCAGAAAGTCAAA
Cgg*acgaaTTCTCTAGATATCGCTCAATACTGACCATTTAaatcct*tcct*gACCTCCATAGCAGAAAGTCAAA
 G**ACGAATTCTCTAGATATCGCTCAATACTGACCATTTAAatcac*atct*gACCTCCATAGCAGAAAGTCAAA
CGT*ACGAATTCTCTAGATATCGCTCAATACTGACCATTTaaaccct*tcca*gacCTCCATAGCAGAAAGTCAAA
 g**ACgaaTTCTCTAGaTATCGCTCAATACTGACCATTTAAATcat*ccct*gaccTCCATAGCAGAAAgtcAAA
 G**ACGAATTCTCTAGATATCGCTCAATACTGACCATTTAAAtcac*ttct*gACCTCCATAGCAGAAAGTCAAA
 G**ACGAATTCTCTAGATATCGCTCAATACTGACCATTTAAatcct*attt*GACCTCCATAGCAGAAAGTCAAA
CGT*T**AATTCTCTAGATATCGCTCAATACTGACCATTTAAatcac*attt*gaggtccATAGCAGAAAGTCAAA
CGTgacGAATTCTCTAGATATCGCTCAATACTGACCATTTaaatcat*tcct*GACCTCCATAGCAGAAAGTCAAA
CGT*T***ATTCTCTAGATATCGCTCAATACTGACCATTTAAATcat*cctt*tacggccaTAGCAGAAAGTCAAA
 G**ACGAATTCTCTAGATATCGCTCAATACTGACCATTTaaatcct*ataa*gagctcCATAGCAGAAAGTCAAA
C***acgaATTCTCTAGATATCGCTCAATACTGACCATTTAAatccc*ttct*gagctccATAGCAGAAAGTCAAA
CGT*T*GaaTTCTCTAGATATCGCTCAATACTGACCATTTAAatcctcattt*gagctccATAGCAGAAAGTCAAA
CGT*TCGAATTCTCTAGATATCGCTCAATACTGACCATTTAaaccct*taat*gagctcCATAGCAGAAAGTCAAA
CGT*ACGAATTCTCTAGATATCGCTCAATACTGACCATTTAAatcct*tcct*GACCTCCATAGCAGAAAGTCAAA
Cgg*acGAATTCTCTAGATATCGCTCAATACTGACCATTTAaatcct*ttaa*gagctcCATAGCAGAAAGTCAAA
```

Using Search For String, I found that the region matched to a different contig that consisted of many similarly tagged reads. This made me think that the region did not belong in my main contig. I then ran a Blast search on the sequence and found that the region corresponds to part of the *E. coli* genome. From this evidence I concluded that this region was contaminated by the vector sequence. To remedy this I pulled out all the affected reads. This effectively eliminated this problem in my main contig. Phred/Phrap may have aligned the vector sequence to that location because of various repeats that exist in the region.

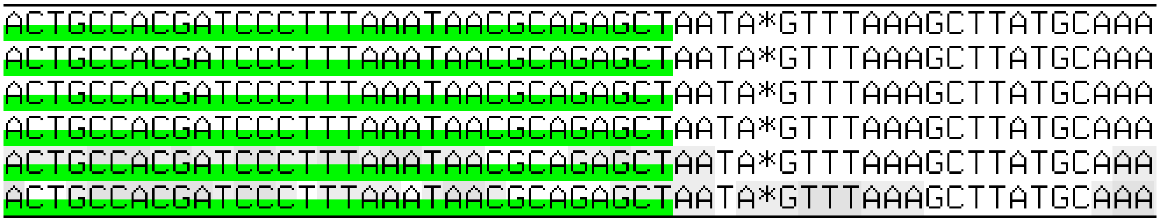
The second large region where I had sequence tagged as *high quality match elsewhere* was 13650-14005. For this region I ran Search For String and found that it corresponded to a different contig that was composed of a single 800 bp read. About 400 bps of this read was also marked *high quality match elsewhere*; using Force Join, I was able to incorporate this region perfectly with the region in my main contig. However the sequence of the untagged regions of the read did not match my main contig (Figure 6). Because of this I removed the single read from my main contig.



```
XBAA-aaf77d09.g1   ►  TAACGCAGAGCTAATA*GTTTAAAGCTTATGCAAA*TTTCAGGAATTTGGG*AAATACCTCGTTGA*CTTCACTCAATCTTC
XBAA-aaf77f10.b1   ►                       gctgtctgatgaatc*atgcttggatTTGgg*aaa*accTCGTtga*CtTCACTCAATCTTc
XBAA-aaf75b04.b1   ◄  TAACGCAGAGCTAATA*GTTTAAAGCTTATGCAAA*TTTCAGGAATTTGGG*AAATACCTCGTTGA*CTTCACTCAATCTTC
XBAA-aaf74h11.g1   ◄  TAACGCAGAGCTAATA*GTTTAAAGCTTATGCAAA*TTTCAGGAATTTGGG*AAATACCTCGTTGA*CTTCACTCAATCTTC
XBAA-aaf74f12.g1   ◄  TAACGCAGAGCTAATA*GTTTAAAGCTTATGCAAA*TTTCAGGAATTTGGG*AAATACCTCGTTGA*CTTCACTCAATCTTC
XBAA-aaf77f11.b1   ◄  TAACGCAGAGCTAATA*GTTTAAAGCTTATGCAAA*TTTCAGGAATTTGGG*AAATACCTCGTTGA*CTTCACTCAATCTTC
XBAA-aaf77c07.g1   ◄  TAACGCAGAGCTCTGATGAATCCCCTAATGATTTTTATCAAAATCATTAAGTTAAGGTAGATACACATCTTGTCATATGATC
XBAA-aaf77d08.g1   ◄  TAACGCAGAGCTAATA*GTTTAAAGCTTATGCAAA*TTTCAGGAATTTGGG*AAATACCTCGTTGA*CTTCACTCAATCTTC
XBAA-aaf74g10.g1   ◄  TAACGCAGAGCTAATA*GTTTAAAGCTTATGCAAA*TTTCAGGAATTTGGG*AAATACCTCGTTGA*CTTCACTCAATCTTC
```

Figure 6. Aligned Reads View of attempted force join

I then used Blast to search for tagged regions in my main contig to see if the mismatch was due to contamination. I found that this region corresponded to *D. melanogaster* with an E value of .005. This strongly suggests that this sequence is not due to any contamination. Closer examination of the tagged region along with the flanking regions revealed that they are very similar in sequence, consisting of sequential sets of identical nucleotides, for example CCC, TTT, AAA (Figure 7). Although there are no obvious repeats in this region the sequence does not seem entirely random. It will be interesting to examine this region during annotation to see if it is functional. This region is tagged because it matches to another contig, but because the contig is comprised of only one read the tagged region in the main contig is likely in the correct position.



```
ACTGCCACGATCCCTTTAAATAACGCAGAGCTAATA*GTTTAAAGCTTATGCAAA
ACTGCCACGATCCCTTTAAATAACGCAGAGCTAATA*GTTTAAAGCTTATGCAAA
ACTGCCACGATCCCTTTAAATAACGCAGAGCTAATA*GTTTAAAGCTTATGCAAA
ACTGCCACGATCCCTTTAAATAACGCAGAGCTAATA*GTTTAAAGCTTATGCAAA
ACTGCCACGATCCCTTTAAATAACGCAGAGCTAATA*GTTTAAAGCTTATGCAAA
ACTGCCACGATCCCTTTAAATAACGCAGAGCTAATA*GTTTAAAGCTTATGCAAA
```

Figure 7. Aligned Reads View exemplifying possible order

On the second ordering of reads, I ordered reads to generate more data for all of the regions that were covered in only one direction, and for a small region of low quality consensus. Conveniently, many of the regions of low quality consensus were within the single stranded regions, allowing me to reduce the number of reads called (Table 2). I ordered all reactions using only the Big Dye chemistry because the trace data did not show any evidence of possible complications. Despite the lack of budget constraint on our project, I felt that frugality was still important.

Table 2. Reads ordered for second round

Oligo	Sequence	Contig	Direction	Template	Target
2	**gataaagaatggagcgtattagg**	**main**	**<--**	**XBAA-aaf77d05**	Single Strand
3	**caactcccatacattacgct**	**main**	**-->**	**XBAA-aaf75g02**	Single Strand
4	**catacattttattggtttgagca**	**main**	**-->**	**XBAA-aaf77g09**	Single Strand
5	**gctgctgtggcaataacaat**	**main**	**<--**	**XBAA-aaf75g02**	Low Consensus Quality
6	**ctaaatccaaatgtatgtaagaagg**	**main**		**XBAA-**	Single

			-->	aaf76d12	Strand
7	**aattgtgcccaaaagcat**	**main**	-->	**XBAA-aaf76g01**	Single Strand
8	**caagaagcgccttgtgta**	**main**	-->	**XBAA-aaf77h05**	Single Strand
9	**ggcacagctgaagtccaa**	**main**	-->	**XBAA-aaf74f01**	Single Strand
10	**cctgaaccagccaaacaac**	**main**	-->	**XBAA-aad98a09**	Single Strand
11	**gatctttgaaaatgaattgaataaa**	**main**	<--	**XBAA-aaf74c05**	Single Strand
12	**aatcggccgtgactgt**	**main**	<--	**XBAA-aaf74c05**	Single Strand

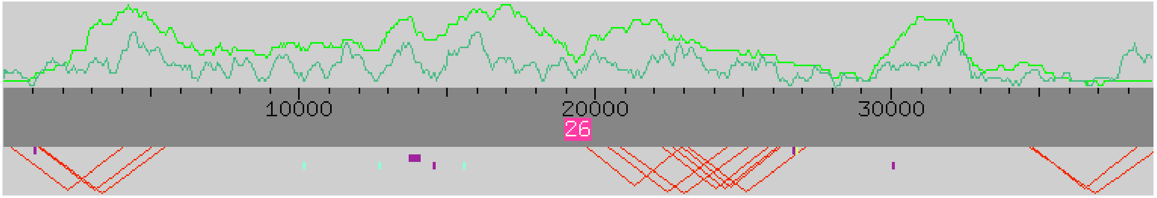
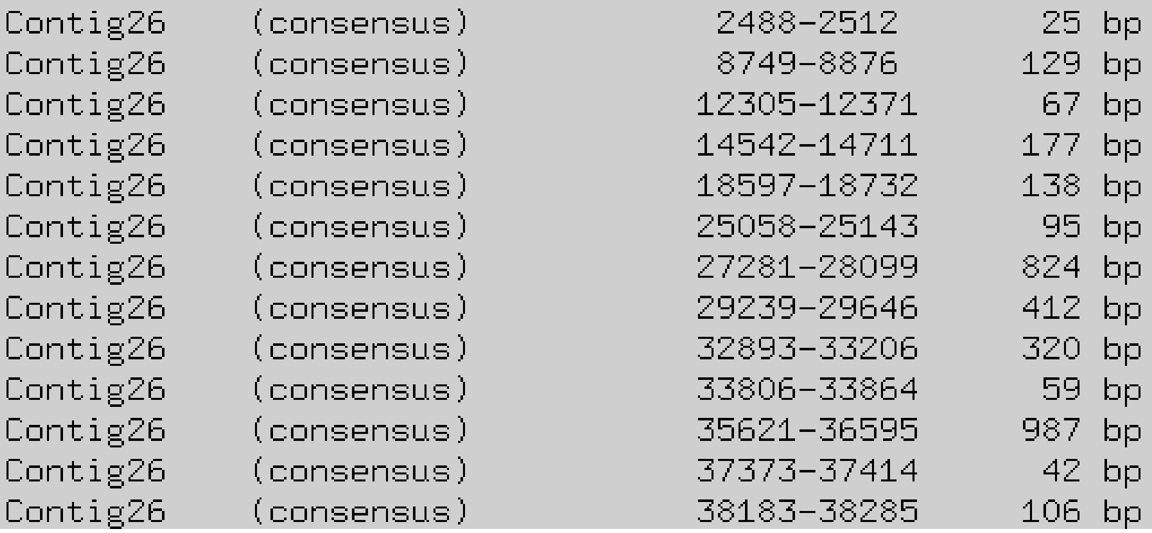


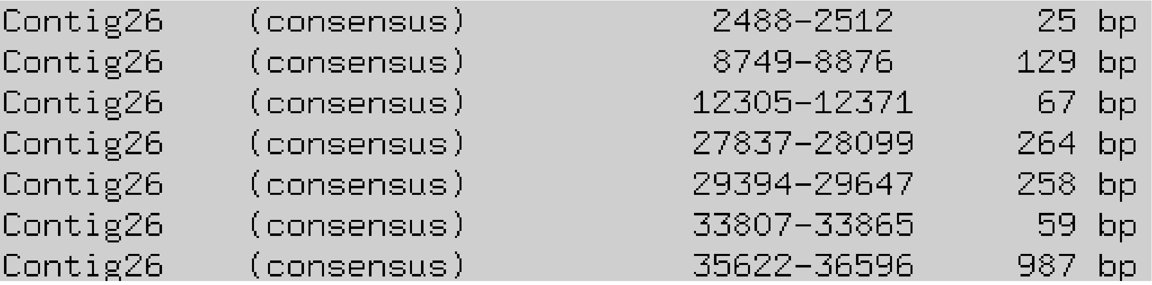
Figure 8. Assembly View after the second round of reads

With this round of reads I was able to cover many of the regions that consisted of reads in only one direction. Of the 13 original single stranded regions, I was able to completely resolve five, and two regions were partially resolved (Figures 9 and 10).



```
Contig26     (consensus)            2488-2512        25 bp
Contig26     (consensus)            8749-8876       129 bp
Contig26     (consensus)           12305-12371       67 bp
Contig26     (consensus)           14542-14711      177 bp
Contig26     (consensus)           18597-18732      138 bp
Contig26     (consensus)           25058-25143       95 bp
Contig26     (consensus)           27281-28099      824 bp
Contig26     (consensus)           29239-29646      412 bp
Contig26     (consensus)           32893-33206      320 bp
Contig26     (consensus)           33806-33864       59 bp
Contig26     (consensus)           35621-36595      987 bp
Contig26     (consensus)           37373-37414       42 bp
Contig26     (consensus)           38183-38285      106 bp
```

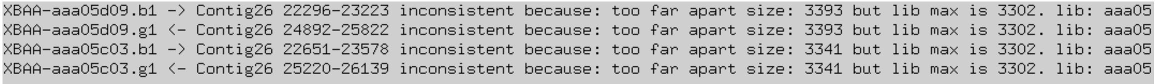
Figure 9. Regions covered by only a single strand before the second round of reads



```
Contig26     (consensus)           2488-2512      25 bp
Contig26     (consensus)           8749-8876     129 bp
Contig26     (consensus)          12305-12371     67 bp
Contig26     (consensus)          27837-28099    264 bp
Contig26     (consensus)          29394-29647    258 bp
Contig26     (consensus)          33807-33865     59 bp
Contig26     (consensus)          35622-36596    987 bp
```

Figure 10. Regions covered by a single strand after second round of reads

After the second round of reactions I examined the *inconsistent forward/reverse reads*. I found that Consed had labeled these forward/reverse reads inconsistencies because they were determined to be too far apart. Figure 11 shows a sample of these forward/reverse reads, and we can see that they are slightly further apart than the limit set by Consed. Because forward/reverse read lengths are variable and the read lengths are so close to the limits, I chose to ignore Consed on this issue.



```
XBAA-aaa05d09.b1 -> Contig26 22296-23223 inconsistent because: too far apart size: 3393 but lib max is 3302. lib: aaa05
XBAA-aaa05d09.g1 <- Contig26 24892-25822 inconsistent because: too far apart size: 3393 but lib max is 3302. lib: aaa05
XBAA-aaa05c03.b1 -> Contig26 22651-23578 inconsistent because: too far apart size: 3341 but lib max is 3302. lib: aaa05
XBAA-aaa05c03.g1 <- Contig26 25220-26139 inconsistent because: too far apart size: 3341 but lib max is 3302. lib: aaa05
```

Figure 11. Inconsistent forward/reverse base pairs

Although not my primary goal, the second round of reads was able to resolve many of the regions with low consensus quality. For the remaining low quality regions I attempted to edit the reads by viewing the traces. A few short regions of low consensus quality, however, could not be edited because the traces were of insufficient quality to support any editing (Figure 12). There are three such regions and because they are each only a few bps long I did not call for additional reads at these regions. In retrospect I should have attempted to cover these areas with additional reads so that the entire contig was finished to high quality.

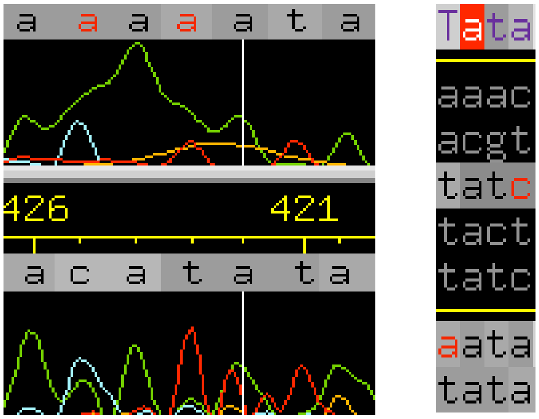
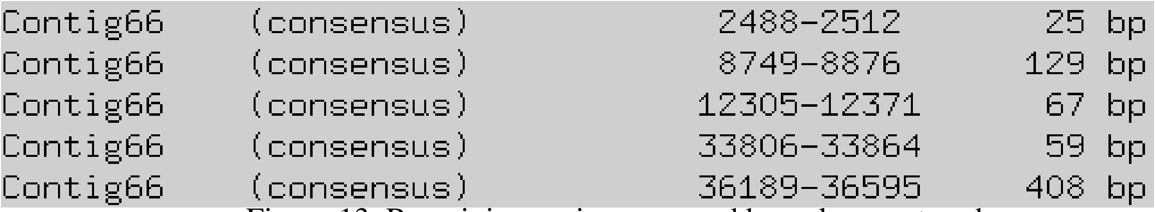


Figure 12. Example of un-editable low quality regions

For my third and final round of reads, I attempted to cover three of the seven regions that had reads from only one strand. I did not attempt to call reads for the other four regions because they were relatively short regions. However, like my earlier discussion I should have attempted to cover these regions with additional reads. Because this was my last chance to order reads, I used the 4:1 chemistry, giving me the best possible chance of resolving these regions (Table 3). Figure 13 shows that my last round of reads was successful. I was able to cover two regions completely, and cover 406 bps of the 987 bp region.

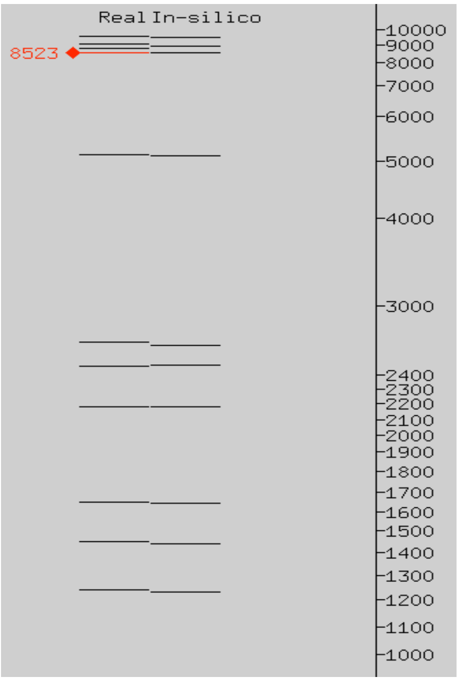
Table 3. Third round of reactions

Oligo	Sequence	Contig	Direction	Template	Target
13	**cgtcacattatcattactctagcag**	**Main**	**-->**	**XBAA-aaa05c10**	Single Strand
14	**ttgccagtatttatcacgaag**	**Main**	**<--**	**XBAA-aaf77d05**	Single Strand
15	**ggttgtgcaattgtgagtattatta**	**Main**	**-->**	**XBAA-aaf75go2**	Single Strand



```
Contig66     (consensus)           2488-2512        25 bp
Contig66     (consensus)           8749-8876       129 bp
Contig66     (consensus)          12305-12371       67 bp
Contig66     (consensus)          33806-33864       59 bp
Contig66     (consensus)          36189-36595      408 bp
```

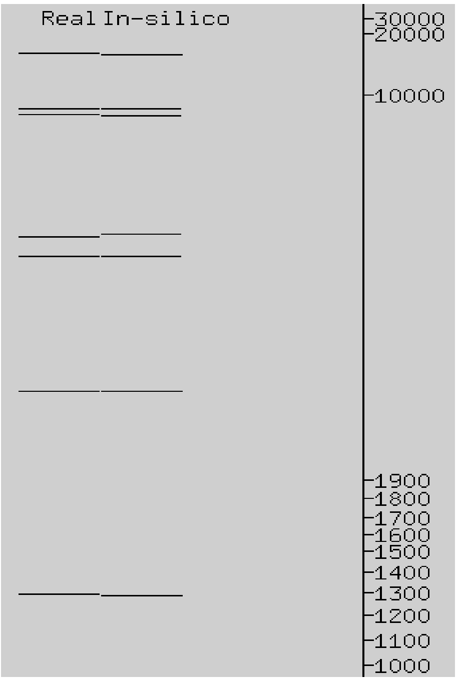
Figure 13. Remaining regions covered by only one strand



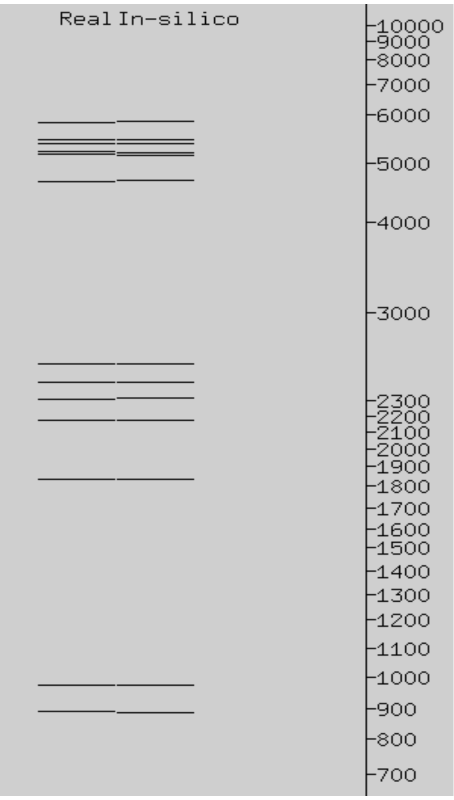
***Eco*RI**

Restriction Digests

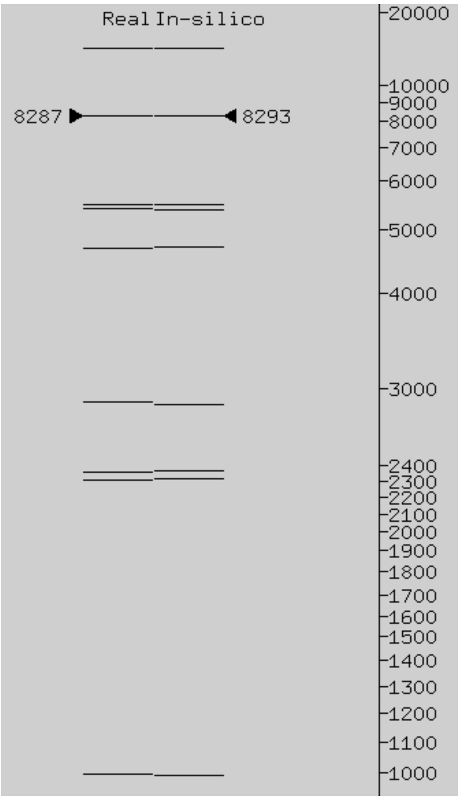
Figure 14



***Sac*I**



***Eco*RV**



***Hind*III**

With the exception of the *Eco*RI restriction digest, a comparison of the real and *in silico* digests shows good agreement. In *Eco*RI there is an inconsistent band. The real digest has a fragment of length of 8523, while the in silico digest has a fragment of length 8531. This is only a difference of eight base pairs; this is well within the range of error for such a broad spectrum of fragment lengths (Figure 14).

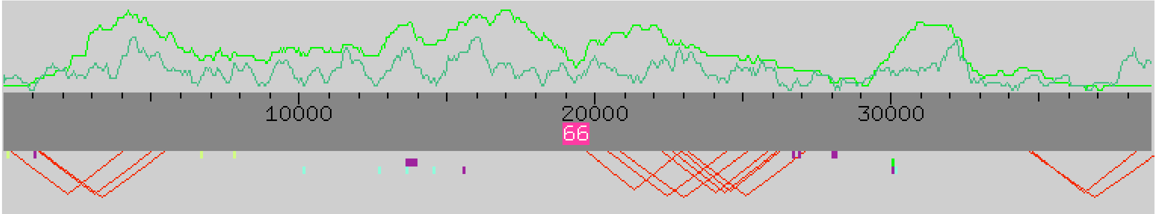


Figure 15. Final Assembly View

Figure 15 is the final Assembly View of my project. Because I initially had one contig, and my problems were quite minor, the final and the initial assembly views look very much alike. Using Search For String I found that the consensus sequence of my contig contains no X's or N's. Using the same technique I also found that there are no mononucleotide runs greater than 15 bps in my sequence. Unfortunately, I was unable to resolve all of the problem areas in my contig with the three rounds of reads (Table 4).

Table 4. Remaining Problems

Problem	Region
Single Strand/ Single Chemistry	2488-2512 8749-8876 12305-12371 33806-33864 36189-36595
Low Consensus Quality	26659-26661 26668-26671 34289-34292 34294-34295

These regions have all been tagged in my final assembly, even though I did not attempt to resolve some of these minor problems. My project was straightforward because it was already assembled into one contig when I received it. My work involved improving the quality of the minor regions that did not meet the required standards. With additional rounds of reads I could resolve the listed problems.