Michelle Miller
Finishing Paper, Revision 2
February 29, 2008

Finishing 230-A21 and 235-I16

Abstract:

The dot chromosome in *Drosophila melanogaster* contains approximately 80 genes and yet is primarily heterochomatic. Genes within heterochromatin are usually silenced in cells, and so the genes on the dot chromosome are interesting to researchers studying transcriptional regulation. To further examine this chromosome, the genomes from several species in this genus are being sequenced. This collaborative group project aims to finish sequence data from the dot chromosome of *D. mojavensis* to high quality.

The goal of my individual project was to finish sequence data from fosmids 230-A21 and 235-I16. The first thing to examine was the amount of overlap between the two fosmids so that the same problems would not have to be solved twice. The 5' region of 230-A21 mapped at about the 27,500 bp position in fosmid 235-I16. Therefore, I decided to finish the sequence for the first 27.5 kb of 235-I16 and then finish all of the sequence for 230-A21.

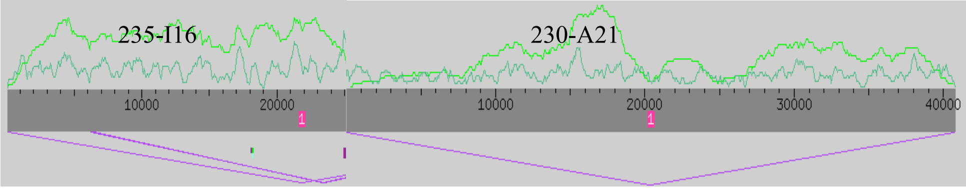


Figure 1: Initial Assembly Views of both fosmids showing regions to be finished

First, I will discuss finishing of the first 27.5 kb of fosmid 235-I16:

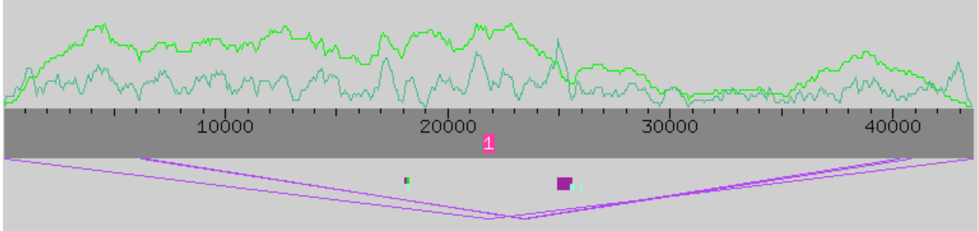


Figure 2: Initial Assembly View: 235-I16

Figure 2 shows the initial Assembly View of the fosmid. It is all in one contig; so, barring any misassemblies, finishing this sequence at first glance seemed to be straightforward. The dark green line indicates the level of high quality reads over a given region. There were a few noticeable regions of low quality, which needed more sequencing reactions to increase confidence in accuracy for the sequences obtained. Also, there are a couple of purple/red discrepant forward/reverse pairs that span the whole length of the contig. These were not of concern; Consed just flagged them because they were above its threshold value, being as they mark fosmid ends rather than subclone ends.

I first ran Crossmatch to check if the contig was assembled correctly and to see the repeats. Crossmatch identifies regions of strong sequence matches that reveal repeats, which are potential areas where the initial assembly may contain similar sequences in places they do not actually belong.

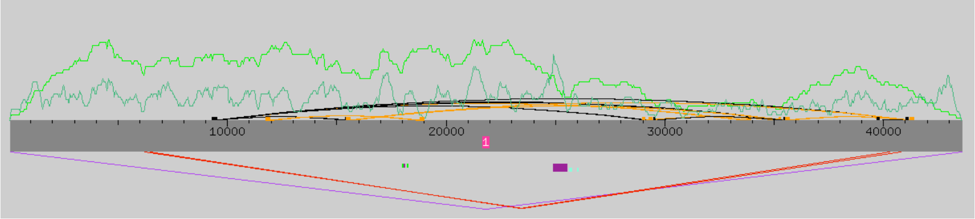
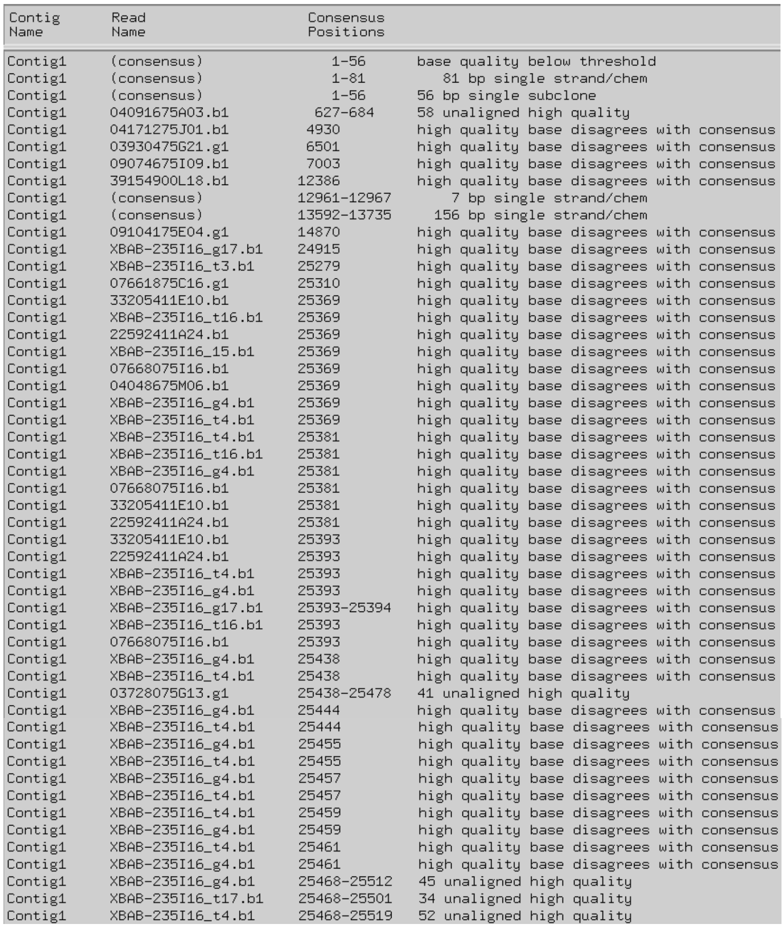


Figure 3: Assembly View for 235-I16 after running Crossmatch

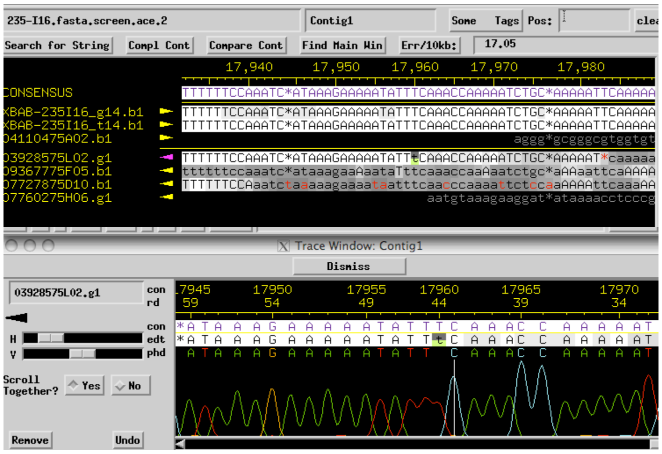
There are clearly several repeats within the fosmid; however, there are no other discrepant read pairs, there is no reason to suspect a global misassembly. The major problems then that needed to be examined were the regions of low quality, high quality discrepant bases, and regions covered only by one strand or chemistry or by a single subclone. These regions are less reliable. The goal for the project was to finish all of the sequence to a quality rating of Phred 30 or higher (called mouse standard). The only region that was low quality was the 5' region. This is usually of little concern, since most of the fosmids chosen in the “golden path” linking together all the fosmids over the length of the chromosome overlap by a few kilobases. However, a reaction was called for this region in this instance, and this did improve the overall Phred score for the sequence.



Contig Name	Read Name	Consensus Positions	
Contig1	(consensus)	1-56	base quality below threshold
Contig1	(consensus)	1-81	81 bp single strand/chem
Contig1	(consensus)	1-56	56 bp single subclone
Contig1	04091675A03.b1	627-684	58 unaligned high quality
Contig1	04171275J01.b1	4930	high quality base disagrees with consensus
Contig1	03930475G21.g1	6501	high quality base disagrees with consensus
Contig1	09074675I09.b1	7003	high quality base disagrees with consensus
Contig1	39154900L18.b1	12386	high quality base disagrees with consensus
Contig1	(consensus)	12961-12967	7 bp single strand/chem
Contig1	(consensus)	13592-13735	156 bp single strand/chem
Contig1	09104175E04.g1	14870	high quality base disagrees with consensus
Contig1	XBAB-235I16_g17.b1	24915	high quality base disagrees with consensus
Contig1	XBAB-235I16_t3.b1	25279	high quality base disagrees with consensus
Contig1	07661875C16.g1	25310	high quality base disagrees with consensus
Contig1	33205411E10.b1	25369	high quality base disagrees with consensus
Contig1	XBAB-235I16_t16.b1	25369	high quality base disagrees with consensus
Contig1	22592411A24.b1	25369	high quality base disagrees with consensus
Contig1	XBAB-235I16_15.b1	25369	high quality base disagrees with consensus
Contig1	07668075I16.b1	25369	high quality base disagrees with consensus
Contig1	04048675M06.b1	25369	high quality base disagrees with consensus
Contig1	XBAB-235I16_g4.b1	25369	high quality base disagrees with consensus
Contig1	XBAB-235I16_t4.b1	25369	high quality base disagrees with consensus
Contig1	XBAB-235I16_t4.b1	25381	high quality base disagrees with consensus
Contig1	XBAB-235I16_t16.b1	25381	high quality base disagrees with consensus
Contig1	XBAB-235I16_g4.b1	25381	high quality base disagrees with consensus
Contig1	07668075I16.b1	25381	high quality base disagrees with consensus
Contig1	33205411E10.b1	25381	high quality base disagrees with consensus
Contig1	22592411A24.b1	25381	high quality base disagrees with consensus
Contig1	33205411E10.b1	25393	high quality base disagrees with consensus
Contig1	22592411A24.b1	25393	high quality base disagrees with consensus
Contig1	XBAB-235I16_t4.b1	25393	high quality base disagrees with consensus
Contig1	XBAB-235I16_g4.b1	25393	high quality base disagrees with consensus
Contig1	XBAB-235I16_g17.b1	25393-25394	high quality base disagrees with consensus
Contig1	XBAB-235I16_t16.b1	25393	high quality base disagrees with consensus
Contig1	07668075I16.b1	25393	high quality base disagrees with consensus
Contig1	XBAB-235I16_g4.b1	25438	high quality base disagrees with consensus
Contig1	XBAB-235I16_t4.b1	25438	high quality base disagrees with consensus
Contig1	03728075G13.g1	25438-25478	41 unaligned high quality
Contig1	XBAB-235I16_g4.b1	25444	high quality base disagrees with consensus
Contig1	XBAB-235I16_t4.b1	25444	high quality base disagrees with consensus
Contig1	XBAB-235I16_g4.b1	25455	high quality base disagrees with consensus
Contig1	XBAB-235I16_t4.b1	25455	high quality base disagrees with consensus
Contig1	XBAB-235I16_g4.b1	25457	high quality base disagrees with consensus
Contig1	XBAB-235I16_t4.b1	25457	high quality base disagrees with consensus
Contig1	XBAB-235I16_t4.b1	25459	high quality base disagrees with consensus
Contig1	XBAB-235I16_g4.b1	25459	high quality base disagrees with consensus
Contig1	XBAB-235I16_t4.b1	25461	high quality base disagrees with consensus
Contig1	XBAB-235I16_g4.b1	25461	high quality base disagrees with consensus
Contig1	XBAB-235I16_g4.b1	25468-25512	45 unaligned high quality
Contig1	XBAB-235I16_t17.b1	25468-25501	34 unaligned high quality
Contig1	XBAB-235I16_t4.b1	25468-25519	52 unaligned high quality

Figure 4: All of the initial problems in the first 27.5 kb of 235-I16

Initially, I believed that many of the high quality discrepancies were polymorphisms. The others I edited manually. One example is illustrated in Figure 5 where there are clearly three T's when only two were called. The rest of the discrepant bases that I did not feel comfortable manually editing were tagged. Many of them are clustered in the same region. I checked over all of the unaligned high quality regions, pulled out the three reads at position 25468 shown above, but ended up putting them back right in the same spot (data not shown). I left two other reads out that had also showed unaligned high quality bases, putting them in separate contigs because the majority of their sequence was of low quality. After completing pulling out the reads and rejoining the contigs, the majority of the consensus sequence was in contig 9.



```
                        17,940    17,950    17,960    17,970    17,980
CONSENSUS            TTTTTTCCAAATC*ATAAAGAAAAATATTTCAAACCAAAAATCTGC*AAAAATTCAAAAA
XBAB-235I16_g14.b1   TTTTTTCCAAATC*ATAAAGAAAAATATTTCAAACCAAAAATCTGC*AAAAATTCAAAAA
XBAB-235I16_t14.b1   TTTTTTCCAAATC*ATAAAGAAAAATATTTCAAACCAAAAATCTGC*AAAAATTCAAAAA
04110475A02.b1                                                 aggg*gcgggcgtggtgt
03928575L02.g1       TTTTTTCCAAATC*ATAAAGAAAAATATTtCAAACCAAAAATCTGC*AAAAAT*caaaaa
09367775F05.b1       ttttttccaaatc*ataaagaaAaataTttcaaaccaaAaatctgc*aAAaattcaAAAA
07727875D10.b1       TTTTTTCCAaatctaaaaagaaaataatttcaacccaaaattctccaaAAAAttcaaaAA
07760275H06.g1                                      aatgtaaagaaggat*ataaaacctcccg
```

```
con  .7945    17950    17955   17960   17965    17970
rd    59       54       49      44      39       34
con  *A T A  A  A G  A  A  A  A  A T A T  T T C  A  A  A C  C  A  A  A  A  A T
edt  *A T A  A  A G  A  A  A  A  A T A T  T t C  A  A  A C  C  A  A  A  A  A T
phd   A T A  A  A G  A  A  A  A  A T A T  T   C  A  A  A C  C  A  A  A  A  A T
```

Figure 5: Example of a high quality discrepancy edited

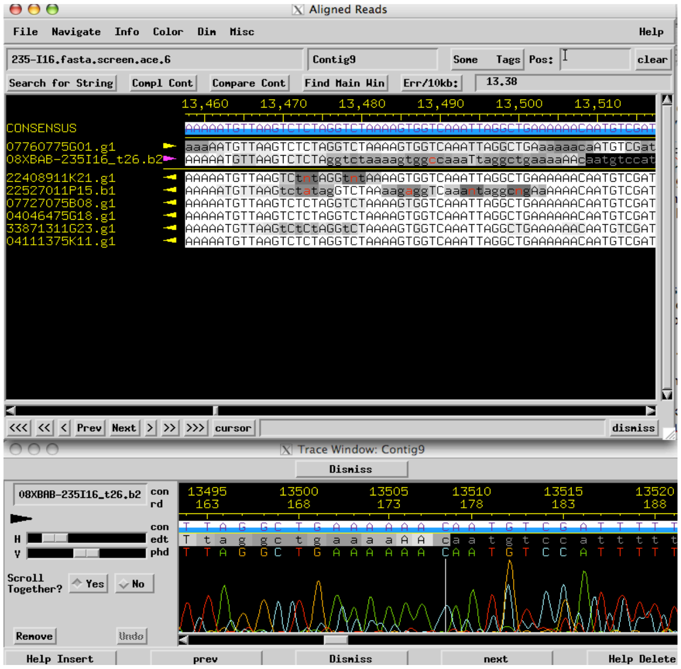
I called reactions for two of the three single stranded/single chemistry regions. The third region (7 bp) was in between two repeats, and there were no unique oligos to use. I called these two reactions before running Autofinish. Autofinish is a program that generates oligos from sequence data to obtain more data for weak areas. I called one additional reaction that Autofinish neglected (see Table 1). The sequences in the region 13592-13735 were above Phred 30, so the reaction was not entirely necessary, and Autofinish neglected it. As it turns out, that reaction did not yield data, but that oligo was the only one the program offered because the region was sandwiched between two repeats. I decided that because the contig was finished to our quality standard (all bases above Phred 30) that I would not waste money on another attempt.

SCR Elgin 4/20/08 6:10 PM
Deleted: t

Table 1: Oligos called by finisher and by Autofinish

Finisher	Oligo Sequence	Direction	Chemistry	Position	Reason
Miller	caatcatcgtcaacatcattat	←	Bigdye, dGTP, 4:1	158-179	To cover low quality at end
Miller	gcatacacacgcatctacac	→	Bigdye, dGTP, 4:1	13127-13146	To cover single strand 13592-13735
Autofinish	aattttaaatttcttattttggca	←	N/A	101-124	To cover low quality at end
Autofinish	ccccatatgagtcctaataatatga	→	N/A	43481-43506	To cover low quality at end

After adding the new reads, the 5' end of the primary contig improved the most. My other read tapered off about 90 bases before the region I was hoping to cover (Figure 6).



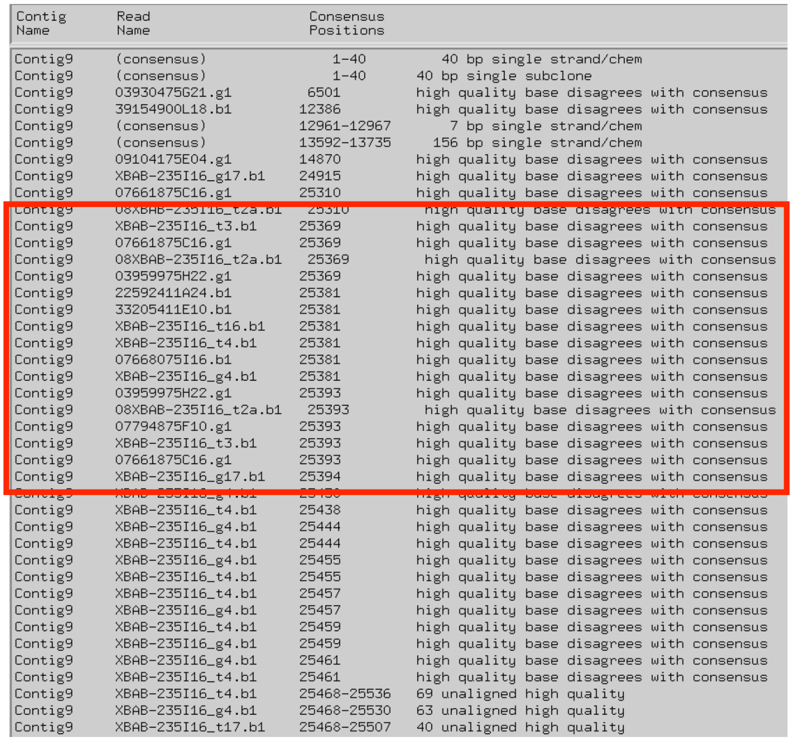
```
235-I16.fasta.screen.ace.6    Contig9    Pos:    Err/10kb: 13.38

                         13,460    13,470    13,480    13,490    13,500    13,510
CONSENSUS                AAAAATGTTAAGTCTCTAGGTCTAAAAGTGGTCAAATTAGGCTGAAAAAACAATGTCGAT
07760775G01.g1           aaaAATGTTAAGTCTCTAGGTCTAAAAGTGGTCAAATTAGGCTGAaaaaacaATGTCGat
08XBAB-235I16_t26.b2     AAAAATGTTAAGTCTCTAggtctaaaagtggccaaaTtaggctgaaaaAAcaatgtccat
22408911K21.g1           AAAAATGTTAAGTCtntAGGtntAAAAGTGGTCAAATTAGGCTGAAAAAACAATGTCGAT
22527011P15.b1           AAAAATGTTAAGtctatagGTCTAAaagaggTCaaantaggcngAaAAAACAATGTCGAT
07727075B08.g1           AAAAATGTTAAGTCTCTAGGTCTAAAAGTGGTCAAATTAGGCTGAAAAAACAATGTCGAT
04046475G18.g1           AAAAATGTTAAGTCTCTAGGTCTAAAAGTGGTCAAATTAGGCTGAAAAAACAATGTCGAT
33871311G23.g1           AAAAATGTTAAGtCtCtAGGtCTAAAAGTGGTCAAATTAGGCTGAAAAAACAATGTCGAT
04111375K11.g1           AAAAATGTTAAGTCTCTAGGTCTAAAAGTGGTCAAATTAGGCTGAAAAAACAATGTCGAT
```

```
08XBAB-235I16_t26.b2
con  13495     13500     13505     13510     13515     13520
rd   163       168       173       178       183       188
con  T T A G G C T G A A A A A A C A A T G T C G A T T T T T
edt  T t a g g c t g a a a a A A c a a t g t c c a t t t t t
phd  T T A G G C T G A A A A A A C A A T G T C C A T T T T T
```

Figure 6: The new read ends before the single stranded region begins

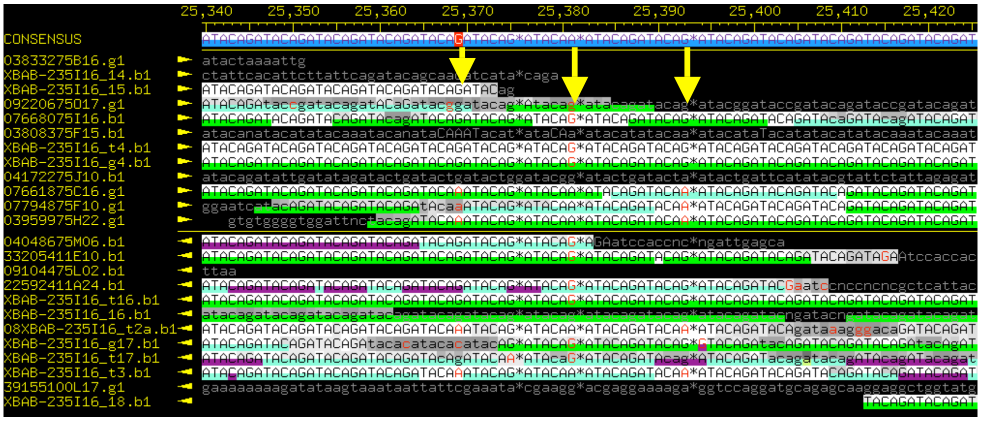
I then generated a list of remaining problems by using the Navigate feature in Consed (Figure 7). This list still contained a number of high quality discrepancies. Some of these remaining high quality discrepancies from the whole genome shotgun reads could be polymorphic with the fosmid DNA in the repeated region or could derive from copies of repetitious sequences elsewhere in the genome (Figure 8).



```
Contig     Read                     Consensus
Name       Name                     Positions

Contig9    (consensus)                 1-40        40 bp single strand/chem
Contig9    (consensus)                 1-40      40 bp single subclone
Contig9    03930475G21.g1           6501         high quality base disagrees with consensus
Contig9    39154900L18.b1          12386         high quality base disagrees with consensus
Contig9    (consensus)             12961-12967      7 bp single strand/chem
Contig9    (consensus)             13592-13735    156 bp single strand/chem
Contig9    09104175E04.g1          14870         high quality base disagrees with consensus
Contig9    XBAB-235I16_g17.b1      24915         high quality base disagrees with consensus
Contig9    07661875C16.g1          25310         high quality base disagrees with consensus
Contig9    08XBAB-235I16_t2a.b1     25310         high quality base disagrees with consensus
Contig9    XBAB-235I16_t3.b1       25369         high quality base disagrees with consensus
Contig9    07661875C16.g1          25369         high quality base disagrees with consensus
Contig9    08XBAB-235I16_t2a.b1     25369         high quality base disagrees with consensus
Contig9    03959975H22.g1          25369         high quality base disagrees with consensus
Contig9    22592411A24.b1          25381         high quality base disagrees with consensus
Contig9    33205411E10.b1          25381         high quality base disagrees with consensus
Contig9    XBAB-235I16_t16.b1      25381         high quality base disagrees with consensus
Contig9    XBAB-235I16_t4.b1       25381         high quality base disagrees with consensus
Contig9    07668075I16.b1          25381         high quality base disagrees with consensus
Contig9    XBAB-235I16_g4.b1       25381         high quality base disagrees with consensus
Contig9    03959975H22.g1          25393         high quality base disagrees with consensus
Contig9    08XBAB-235I16_t2a.b1     25393         high quality base disagrees with consensus
Contig9    07794875F10.g1          25393         high quality base disagrees with consensus
Contig9    XBAB-235I16_t3.b1       25393         high quality base disagrees with consensus
Contig9    07661875C16.g1          25393         high quality base disagrees with consensus
Contig9    XBAB-235I16_g17.b1      25394         high quality base disagrees with consensus
Contig9    XBAB-235I16_g4.b1       25430         high quality base disagrees with consensus
Contig9    XBAB-235I16_t4.b1       25438         high quality base disagrees with consensus
Contig9    XBAB-235I16_g4.b1       25444         high quality base disagrees with consensus
Contig9    XBAB-235I16_t4.b1       25444         high quality base disagrees with consensus
Contig9    XBAB-235I16_g4.b1       25455         high quality base disagrees with consensus
Contig9    XBAB-235I16_t4.b1       25455         high quality base disagrees with consensus
Contig9    XBAB-235I16_t4.b1       25457         high quality base disagrees with consensus
Contig9    XBAB-235I16_g4.b1       25457         high quality base disagrees with consensus
Contig9    XBAB-235I16_t4.b1       25459         high quality base disagrees with consensus
Contig9    XBAB-235I16_g4.b1       25459         high quality base disagrees with consensus
Contig9    XBAB-235I16_g4.b1       25461         high quality base disagrees with consensus
Contig9    XBAB-235I16_t4.b1       25461         high quality base disagrees with consensus
Contig9    XBAB-235I16_t4.b1       25468-25536   69 unaligned high quality
Contig9    XBAB-235I16_g4.b1       25468-25530   63 unaligned high quality
Contig9    XBAB-235I16_t17.b1      25468-25507   40 unaligned high quality
```

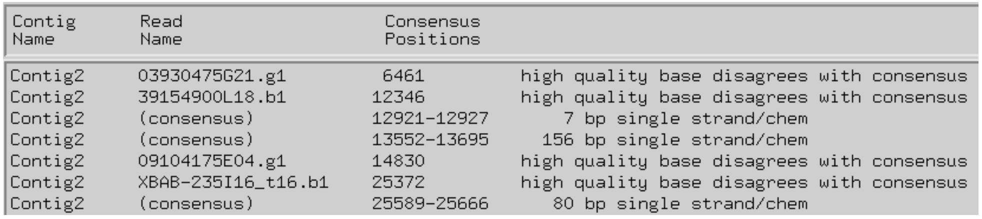
Figure 7: Remaining problems from the Navigate list before running PhredPhrap. Traces shown for those bases framed in red in Figure 8.



```
                         25,340    25,350    25,360    25,370    25,380    25,390    25,400    25,410    25,420
CONSENSUS                ATACAGATACAGATACAGATACAGATACAGATACAG*ATACAA*ATACAGATACAG*ATACAGATACAGATACAGATACAGATACAGAT
03833275B16.g1        ►  atactaaaattg
XBAB-235I16_14.b1     ►  ctattcacattcttattcagatacagcaa[illegible]atcata*caga
XBAB-235I16_15.b1     ►  ATACAGATACAGATACAGATACAGATACAGATACag
09220675O17.g1        ►  ATACAGAtaccgatacagatACaGatacggatacag*Atacag*atacagatacag*atacggataccgatacagataccgatacagat
07668075I16.b1        ►  ATACAGATACAGATACAGATAcagATACAGATACAG*ATACAG*ATACAGATACAG*ATACAGATACAGATAcaGATAcagATACAGAT
03808375F15.b1        ►  atacanatacatatacaaatacanataCAAATacat*ataCAa*atacatatacaa*atacataTacatatacatatacaaatacaaat
XBAB-235I16_t4.b1     ►  ATACAGATACAGATACAGATACAGATACAGATACAG*ATACAG*ATACAGATACAG*ATACAGATACAGATACAGATACAGATACAGAT
XBAB-235I16_g4.b1     ►  ATACAGATACAGATACAGATACAGATACAGATACAG*ATACAG*ATACAGATACAG*ATACAGATACAGATACAGATACAGATACAGAT
04172275J10.b1        ►  atacagatattgatatagatactgatactgatactggatacgg*atactgatacta*atactgattcatatacgtattctattagagat
07661875C16.g1        ►  ATACAGATACAGATACAGATACAGATACAAATACAG*ATACAA*ATACAGATACAA*ATACAGATACAGATACAGATACAGATACAGAT
07794875F10.g1        ►  ggaatcatacAGATACAGATACAGAtAcaaATACAG*ATACAA*ATACAGATACAA*ATACAGATACAGATACAGATACAGATACAGAT
03959975H22.g1        ►     gtgtggggtggattnctacagATACAAATACAG*ATACAA*ATACAGATACAA*ATACAGATACAGATACAGATACAGATACAGAT
04048675M06.b1        ◄  ATACAGATACAGATACAGATACAGATACAGATACAG*ATACAG*AGAatccaccnc*ngattgagca
33205411E10.b1        ◄  ATACAGATACAGATACAGATACAGATACAGATACAG*ATACAG*ATACAGATACAG*ATACAGATACAGATACAGATAGAAtccaccac
09104475L02.b1        ◄  ttaa
22592411A24.b1        ◄  ATACAGATACAGATACAGATACAGATACAGATACAG*ATACAG*ATACAGATACAG*ATACAGATACGaatccnccncncgctcattac
XBAB-235I16_t16.b1    ◄  ATACAGATACAGATACAGATACAGATACAGATACAG*ATACAG*ATACAGATACAG*ATACAGATACAGATACAGATACAGATACAGAT
XBAB-235I16_16.b1     ◄  atacagatacagatacagatacagatacagatacag*atacag*atacagatacag*atacagatacngatacngatacagatacagat
08XBAB-235I16_t2a.b1  ◄  ATACAGATACAGATACAGATACAGATACAAATACAG*ATACAA*ATACAGATACAA*ATACAGATACAgataaagggacaGATACAGAT
XBAB-235I16_g17.b1    ◄  ATACAGATACAGATACAGAtacacatacacatacAG*ATACAG*ATACAGATACAG*GTACAGAtacAGATACAGATACAGAtacAgAT
XBAB-235I16_t17.b1    ◄  ATACAGATACAGATACAGATACAGATAcAgATACAA*ATAcaG*ATACAGATAcag*ATACAGATacagatacagATAcagATacagat
XBAB-235I16_t3.b1     ◄  ATACAGATACAGATACAGATACAGATACAAATACAG*ATACAA*ATACAGATACAA*ATACAGATACAGATACAGATACAGATACAGAT
39155100L17.g1        ◄  gaaaaaaaagatataagtaaataattattcgaaata*cgaagg*acgaggaaaaga*ggtccaggatgcagagcaaggaggctggtatg
XBAB-235I16_18.b1     ◄                                                                              TACAGATACAGAT
```

Figure 8: Some examples of the high quality discrepancies I thought were polymorphisms.

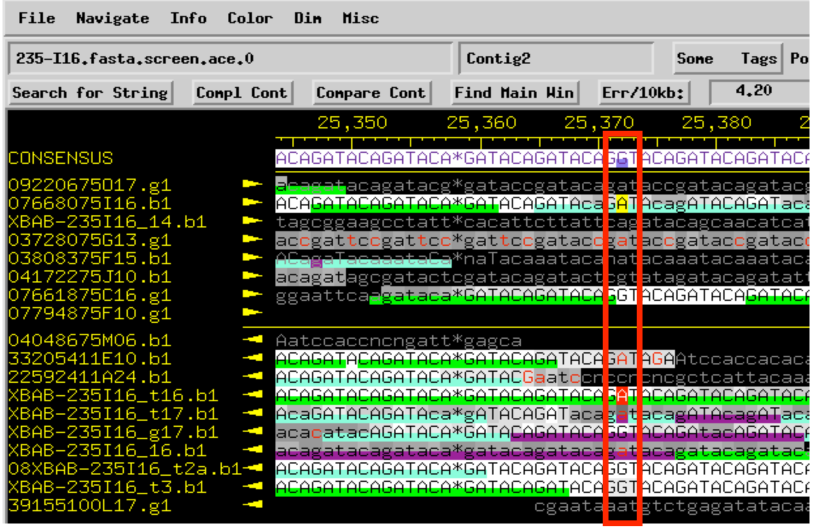
For simplicity's sake, I ran PhredPhrap again, so now there are one large (numbered contig 2) and four single-read contigs. I tore out one read that gave a high quality unaligned region and was overall of poor quality. Then I tore out two others that gave many high quality discrepant bases in a huge repeat. This process eliminated many high quality discrepancies and all of the high quality misaligned regions from the Navigate list. As shown in Figure 8, there were many repeats in the region where most of the suspected polymorphisms were (highlighted by the green tags).



Contig Name	Read Name	Consensus Positions	
Contig2	03930475G21.g1	6461	high quality base disagrees with consensus
Contig2	39154900L18.b1	12346	high quality base disagrees with consensus
Contig2	(consensus)	12921-12927	7 bp single strand/chem
Contig2	(consensus)	13552-13695	156 bp single strand/chem
Contig2	09104175E04.g1	14830	high quality base disagrees with consensus
Contig2	XBAB-235I16_t16.b1	25372	high quality base disagrees with consensus
Contig2	(consensus)	25589-25666	80 bp single strand/chem

Figure 9: Remaining problems in 235-I16

This procedure eliminated many problems from the initial Navigate list. It seems that rerunning PhredPhrap resulted in a new assembly in which several of the bases that were part of the high quality discrepancy category became part of the consensus sequence. However, the assembly was slightly different with new reads added in (some others besides mine coming from the other schools). This generated a new single strand region that was not there before. I trust this assembly, however, because the three restriction digests shown below are very exact. The first three high quality discrepancies were all checked, and there are many other reads in the region to ensure confidence in the consensus. The last high quality discrepancy is potentially a real one.



```
CONSENSUS                    ACAGATACAGATACA*GATACAGATACAGGTACAGATACAGATACA
09220675017.g1           ►   acagatacagatacg*gataccgatacagataccgatacagatacg
07668075I16.b1           ►   ACAGATACAGATACA*GATACAGATAcaGATAcagATACAGATaca
XBAB-235I16_14.b1        ►   tagcggaagcctatt*cacattcttattcagatacagcaacatca
03728075G13.g1           ►   accgattccgattcc*gattccgataccgataccgataccgatacc
03808375F15.b1           ►   ACagaTacaaataCa*naTacaaatacanatacaaatacaaataca
04172275J10.b1           ►   acagatagcgatactcgatacagatactcgtatagatacagatatt
07661875C16.g1           ►   ggaattcaagataca*GATACAGATACAGGTACAGATACAGATACA
07794875F10.g1           ►
04048675M06.b1           ◄   Aatccaccncngatt*gagca
33205411E10.b1           ◄   ACAGATACAGATACA*GATACAGATACAGATAGAAtccaccacaca
22592411A24.b1           ◄   ACAGATACAGATACA*GATACGaatccnccrcncgctcattacaaa
XBAB-235I16_t16.b1       ◄   ACAGATACAGATACA*GATACAGATACAGATACAGATACAGATACA
XBAB-235I16_t17.b1       ◄   AcaGATACAGATAca*gATACAGATacagatacagATAcagATaca
XBAB-235I16_g17.b1       ◄   acacatacAGATACA*GATACAGATACAGGTACAGAtacAGATACA
XBAB-235I16_16.b1        ◄   acagatacagataca*gatacagatacagatacagatacagatac
08XBAB-235I16_t2a.b1     ◄   ACAGATACAGATACA*GATACAGATACAGGTACAGATACAGATACA
XBAB-235I16_t3.b1        ◄   ACAGATACAGATACA*GATACAGATACAGGTACAGATACAGATACA
39155100L17.g1           ◄                      cgaataaatgtctgagatataca
```

Figure 10: A potential real polymorphism

As a final check, I examined the digests. Restriction digests compare the actual DNA template with a synthetic computer-generated digest of the assembly (*in-silico)*. I initially thought that I had some discrepancies between the *in-silico* digest and the actual digests, but I think those were most likely due to a technical issue with the vector in Consed. Several places showed two bands that added to one correct band in the real digest. This issue was resolved by selecting the option to digest a single contig *in-silico*, which does not force Consed to try to join the vector ends together. (The digest feature in Consed was not designed for Whole Genome Shotgun reads.) These clean digests are shown below. Finally, I am proud to report that the entire consensus sequence from the first 27.5 kb is Phred 52 or above.

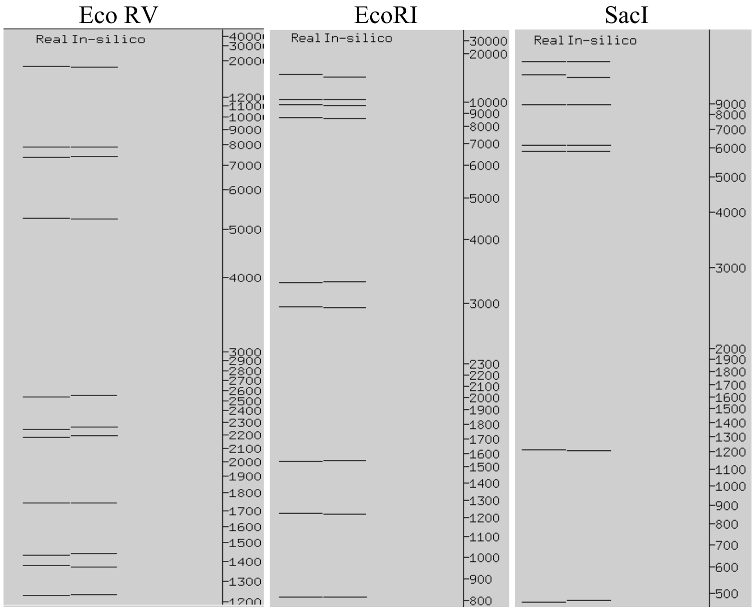


Figure 11: Restriction Digests of 235-I16

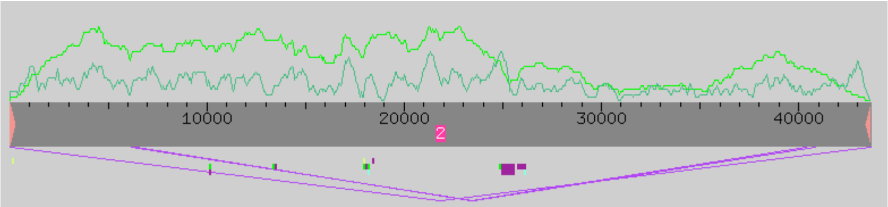


Figure 12: Final Assembly View of 235-I16

FOSMID 230-A21:

This fosmid had fewer issues to resolve than the one previously discussed. The initial assembly was of a single contig (Figure 13). There were a few regions that were single stranded, but nothing that was of low quality except for the two ends (Figure 15). Crossmatch showed few repeats throughout the assembly (Figure 14). Neither of these views indicated any major misassemblies.

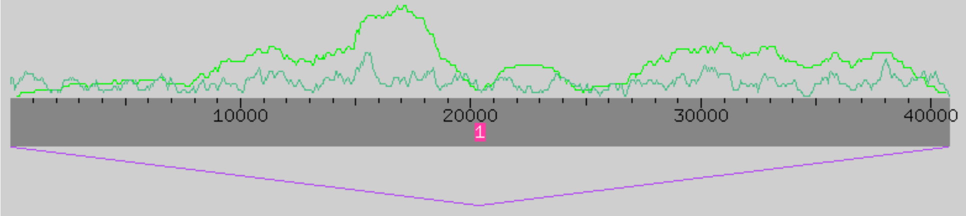


Figure 13: Initial Assembly View: 230-A21

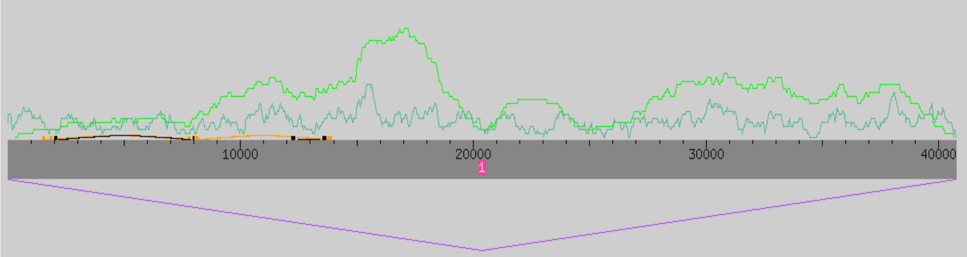
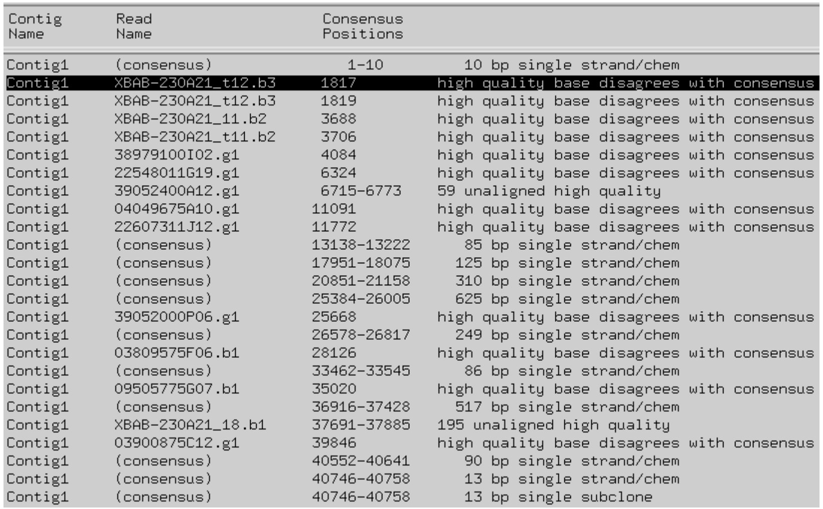


Figure 14: Crossmatch in 230-A21

There were two places in this assembly where I was able to manually correct high quality base discrepancies (the last two in Figure 15), because the reads had been miscalled due to compressions.



```
Contig     Read                     Consensus
Name       Name                     Positions

Contig1    (consensus)                 1-10        10 bp single strand/chem
Contig1    XBAB-230A21_t12.b3       1817         high quality base disagrees with consensus
Contig1    XBAB-230A21_t12.b3       1819         high quality base disagrees with consensus
Contig1    XBAB-230A21_11.b2        3688         high quality base disagrees with consensus
Contig1    XBAB-230A21_t11.b2       3706         high quality base disagrees with consensus
Contig1    38979100I02.g1           4084         high quality base disagrees with consensus
Contig1    22548011G19.g1           6324         high quality base disagrees with consensus
Contig1    39052400A12.g1           6715-6773    59 unaligned high quality
Contig1    04049675A10.g1          11091         high quality base disagrees with consensus
Contig1    22607311J12.g1          11772         high quality base disagrees with consensus
Contig1    (consensus)             13138-13222     85 bp single strand/chem
Contig1    (consensus)             17951-18075    125 bp single strand/chem
Contig1    (consensus)             20851-21158    310 bp single strand/chem
Contig1    (consensus)             25384-26005    625 bp single strand/chem
Contig1    39052000P06.g1          25668         high quality base disagrees with consensus
Contig1    (consensus)             26578-26817    249 bp single strand/chem
Contig1    03809575F06.b1          28126         high quality base disagrees with consensus
Contig1    (consensus)             33462-33545     86 bp single strand/chem
Contig1    09505775G07.b1          35020         high quality base disagrees with consensus
Contig1    (consensus)             36916-37428    517 bp single strand/chem
Contig1    XBAB-230A21_18.b1       37691-37885   195 unaligned high quality
Contig1    03900875C12.g1          39846         high quality base disagrees with consensus
Contig1    (consensus)             40552-40641     90 bp single strand/chem
Contig1    (consensus)             40746-40758     13 bp single strand/chem
Contig1    (consensus)             40746-40758     13 bp single subclone
```

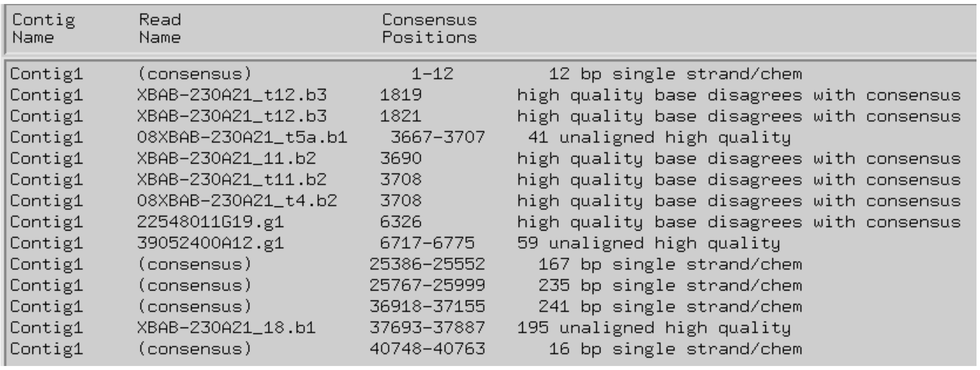
Figure 15: Initial Problems for 230-A21

After completing this task, I started ordering reactions to cover eight of the interior single stranded regions.

Table 2: Reactions called for fosmid 230-A21

Finisher	Direction	Chemistry	Position	Region	Outcome
Miller	→	4:1	12855-12877	13138-13222 (85bp)	covered
Miller	→	4:1	17834-17856	17951-18075 (125 bp)	covered
Miller	←	4:1	21266-21288	20851-21158 (310 bp)	covered
Miller	←	4:1	26071-26090	25384-26005 (625 bp)	25384-25550 (167 bp), 25765-26005 (235 bp)
Miller	→	4:1	26287-26305	26578-26817 (249 bp)	covered
Miller	→	4:1	33255-33277	33462-33545 (86 bp)	covered
Miller	←	4:1	37636-37656	36916-37428 (517 bp)	36916-37153 (241 bp)
Miller	→	4:1	40463-40481	40552-40641 (90 bp)	covered
Autofinish	→	N/A	40678-40760	To cover low quality at end	Not ordered, but covered by above reaction up to the vector sequence

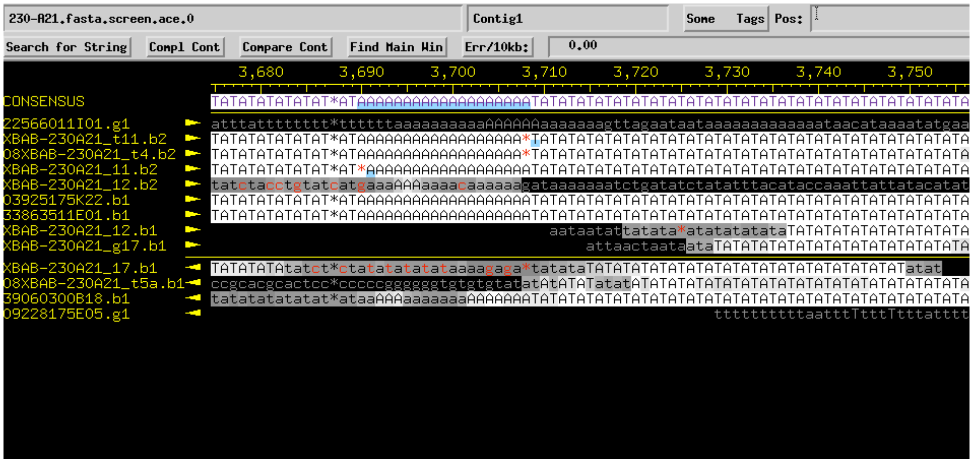
Autofinish only called one reaction, which I did not feel was necessary to include in the order, because this region was covered with one reaction I had already called for the gap spanning 40552-40641 (Table 2). I added the new reads, and there are only three remaining smaller single stranded regions. I re-ran PhredPhrap and generated the list of final problems (Figure 16). Reassembly by PhredPhrap seems to have eliminated most of the high quality unaligned regions. The high quality discrepancies were all checked and tagged, but there are sufficient reads in each place to be confident of the consensus, with one exception. The region around position 3700 is a slight concern. There is a long stretch of mononucelotide A's there in the middle of a poly AT repeat, and the reads coming directly off the fosmid suggest the consensus could be shorter by one base than the whole genome shotgun reads would indicate (Figures 17 and 18).



```
Contig       Read                          Consensus
Name         Name                          Positions

Contig1      (consensus)                     1-12         12 bp single strand/chem
Contig1      XBAB-230A21_t12.b3            1819          high quality base disagrees with consensus
Contig1      XBAB-230A21_t12.b3            1821          high quality base disagrees with consensus
Contig1      08XBAB-230A21_t5a.b1           3667-3707     41 unaligned high quality
Contig1      XBAB-230A21_11.b2             3690          high quality base disagrees with consensus
Contig1      XBAB-230A21_t11.b2            3708          high quality base disagrees with consensus
Contig1      08XBAB-230A21_t4.b2           3708          high quality base disagrees with consensus
Contig1      22548011G19.g1                6326          high quality base disagrees with consensus
Contig1      39052400A12.g1                6717-6775     59 unaligned high quality
Contig1      (consensus)                  25386-25552     167 bp single strand/chem
Contig1      (consensus)                  25767-25999     235 bp single strand/chem
Contig1      (consensus)                  36918-37155     241 bp single strand/chem
Contig1      XBAB-230A21_18.b1            37693-37887    195 unaligned high quality
Contig1      (consensus)                  40748-40763     16 bp single strand/chem
```

Figure 16: Final Problems



```
                     3,680     3,690     3,700     3,710     3,720     3,730     3,740     3,750
CONSENSUS            TATATATATATAT*ATAAAAAAAAAAAAAAAAAAATATATATATATATATATATATATATATATATATATATATATATATATA
22566011I01.g1     ► atttattttttt*tttttaaaaaaaaaaAAAAAaaaaaaaagttagaataataaaaaaaaaaaaaataacataaaatatgaa
XBAB-230A21_t11.b2 ► TATATATATATAT*ATAAAAAAAAAAAAAAAAAA*TATATATATATATATATATATATATATATATATATATATATATATATA
08XBAB-230A21_t4.b2 ► TATATATATATAT*ATAAAAAAAAAAAAAAAAAA*TATATATATATATATATATATATATATATATATATATATATATATATA
XBAB-230A21_11.b2  ► TATATATATATAT*AT*AAAAAAAAAAAAAAAAAATATATATATATATATATATATATATATATATATATATATATATATATATA
XBAB-230A21_12.b2  ► tatctacctgtatcatgaaaAAAaaaacaaaaaagataaaaaaatctgatatctatatttacataccaaattattatacatat
03925175K22.b1     ► TATATATATATAT*ATAAAAAAAAAAAAAAAAAAATATATATATATATATATATATATATATATATATATATATATATATATA
33863511E01.b1     ► TATATATATATAT*ATAAAAAAAAAAAAAAAAAAATATATATATATATATATATATATATATATATATATATATATATATATA
XBAB-230A21_12.b1  ►                                       aataatattatata*atatatatataTATATATATATATATATA
XBAB-230A21_g17.b1 ►                                           attaactaataataTATATATATATATATATATATATATA
XBAB-230A21_17.b1  ◄ TATATATAtatct*ctatatatatataaaagaga*tatataTATATATATATATATATATATATATATATATATatat
08XBAB-230A21_t5a.b1◄ ccgcacgcactcc*cccccgggggggtgtgtgtatatAtATATatatATATATATATATATATATATATATATATATATATA
39060300B18.b1     ◄ tatatatatatat*ataaAAAaaaaaaaAAAAAAATATATATATATATATATATATATATATATATATATATATATATATA
09228175E05.g1     ◄                                                        ttttttttttaatttTtttTtttatttt
```

Figure 17: Mononucelotide run of A's within a poly AT repeat

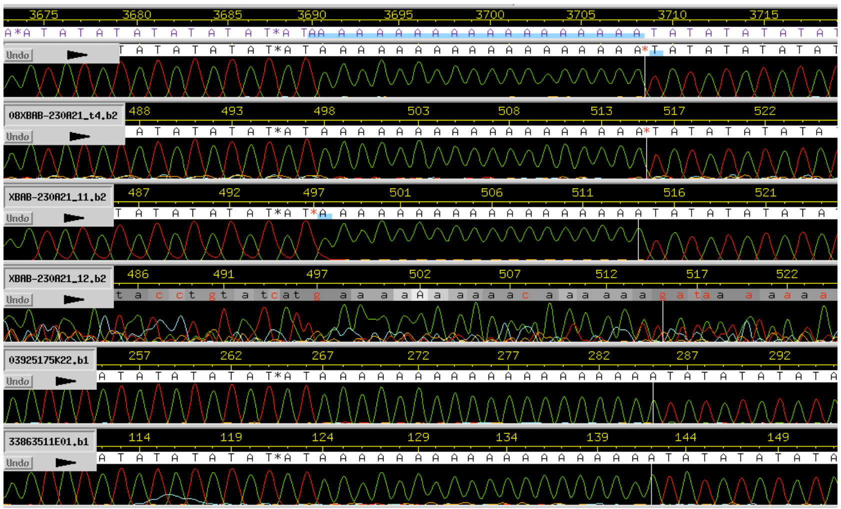
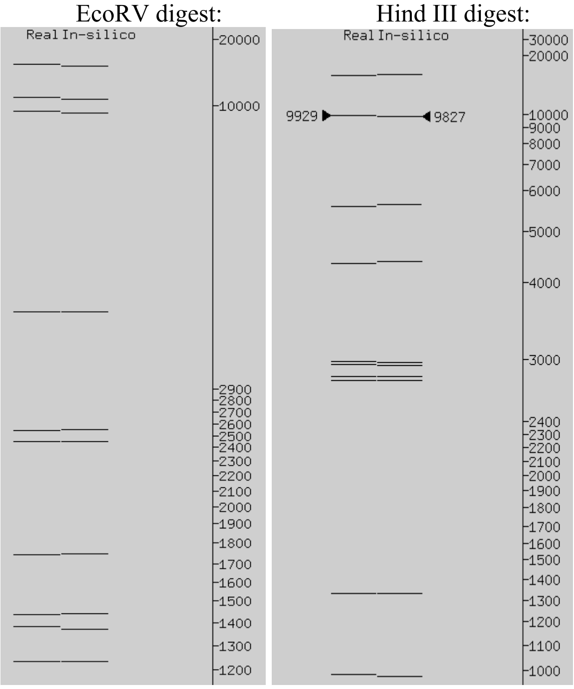


Figure 18: Traces for one of the PolyA regions of 230-A21


EcoRV digest:

Hind III digest:

Figure 19: Restriction Digests of 230-A21

Finally, the digests show no discrepancies, and the entire sequence is Phred 40 or higher.

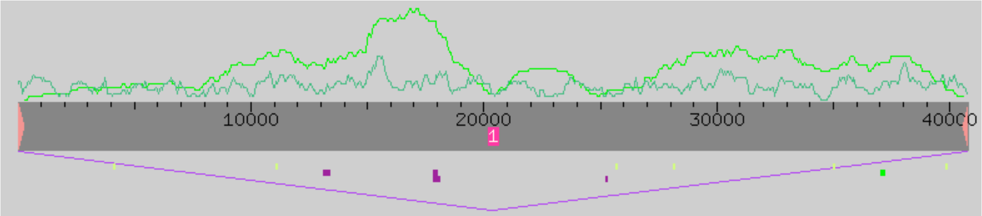


Figure 20: Final Assembly View of 230-A21

Short notes on the other fosmids:

250-J02:
I force-joined the gap and then ordered more reactions to confirm the join. There was a hit on the BLAST search against some bacterial DNA. 76% ID to a subunit of Candidatus Peligabacter ubique's NADH dehydrogenase. This is unlikely to be contamination because while this strain of bacteria has the smallest genome of any self-replicating cell yet sequenced, it lives in the ocean, and therefore, is most likely not the host strain used for sequencing. Additionally, the gene itself is important and likely highly conserved.

200-N19:
There is a polymorphism with an AT repeat with the neighboring fosmid. However, after incorporating some reads from 205 and re-running PhredPhrap, the gap in this fosmid was able to be closed. The digests are perfect.

325-M22
There was a gap with many overlapping forward/reverse pairs. However, there was a chimera at the 3' end of the left contig, yet no sequences at that end matched anything at the 5' region of the right contig. The digests indicated though that there was about 100 bp overlap between the two contigs. Performing a force join over just 11 bases did not improve the *in-silico* digest. Oligos were called to breach the gap, and many reactions worked. After re-running PhredPhrap, the chimera in that location went away, but the tag showed up again in a read that was misaligned called by students at another school. The digests now look fine, and there are no longer any high quality discrepancies or low quality regions.