

**FINISHING PROJECT:
XBAA-205-H09
[Second Revision]**

Sarah Spencer
March 21, 2008

Abstract:

*For my project I attempted high quality finishing of the fosmid XBAA-205-H09 from *Drosophila mojavensis* using the finishing program Consed. My clone, along with others from a larger project, will form a high quality sequence of the *D. mojavensis* dot chromosome. Once the *D. mojavensis* dot chromosome sequence is complete, it will be available for comparison with the dot chromosome sequences of other species such as *D. melanogaster* and *D. virilis*. With this enriched data set, a greater understanding of gene expression and heterochromatin formation in the variably structured dot chromosome will be possible. My personal project required closing two sequence gaps and addressing minor internal sequence issues. To close the gaps, I ordered my own reactions, incorporated data from the *D. mojavensis* Whole Genome Shotgun database, and added data from the adjacent subclone 200-N19. I achieved a single contig which is finished to standard criteria. Note there are 14 single strand and single chemistry regions, but these meet the Phred 30 threshold for data quality. Future work should verify a TA repeat region which was polymorphic between reads generated from subclones 205-H09 and 200-N19.*

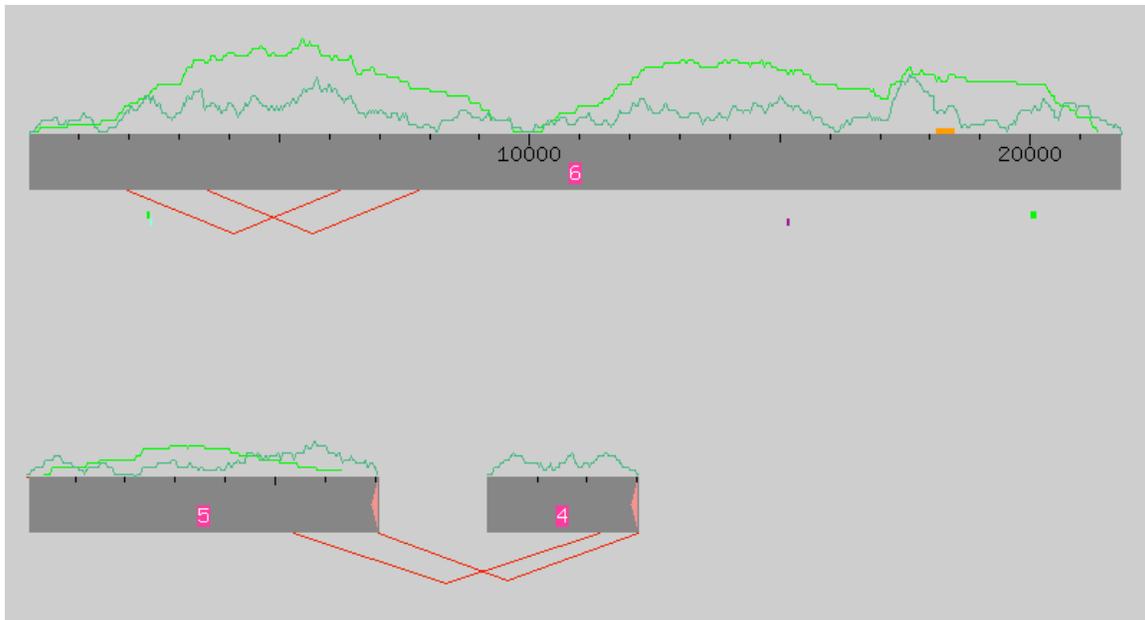


Figure 1. Assembly View of my initial fosmid clone assembly with clone ends labeled and Crossmatch results displayed.

Initial Assessment

My initial assembly consisted of three contigs without any gap-spanning subclones (Figure 1). To finish the sequence, I needed to obtain sequence data across the

sequence gaps and/or to forcibly join the contigs into a single continuous contig; this process would eliminate (close) the sequence gaps. I tagged the clone ends on contigs 4 and 5, verifying that the sequence gaps fell between these contigs and both ends of the large, central contig 6. I then used *Search for String* with sequence from the contig ends (excluding the clone ends), but found no matches to indicate in what order the three contigs should fit together.

The program Crossmatch only revealed a 160 bp tandem repeat in the midst of contig 6 (two regions with 93.2% similarity). Since I received no obvious signs of missassembly from Crossmatch, I continued my analysis by navigating for low consensus quality, high quality discrepancies, and single subclone regions within each individual contig. During this navigation process, I discounted sequence problems within 20 bp of the two clone ends. Also, the regions leading into and separated by gaps (contig gap ends) invariably harbored low consensus quality and single subclone sequence. I did not address these issues individually because I intended to call reads sequencing into the gaps; these would work to improve the sequence of the gap ends as the gaps closed.

I found two low consensus quality regions that did not fall near clone ends or gap regions. Contig 5 held the more serious problem, with a dramatic loss of quality over 78 bp in the midst of a single subclone region (Figure 2). In contig 6, a smaller 7 bp region encompassed a string of low quality G bases (Figure 3).

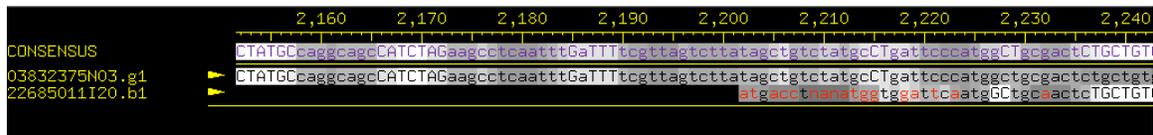


Figure 2. Aligned Reads window of contig 5 encompassing a 78 bp low consensus quality, single subclone region (2158-2235).



Figure 3. Aligned Reads window of contig 6 displaying a 7 bp low consensus quality region (1450-1456).

The results of my high quality discrepancy search presented a greater concern (Figure 4). Consed only identified one high quality discrepancy in contig 5, but contig 6 included eight locations where a high quality base disagreed with the consensus sequence.

Contig Name	Read Name	Consensus Positions	
Contig6	07729475N18.b1	2770	high quality base disagrees with consensus
Contig6	03966475I09.g1	4081	high quality base disagrees with consensus
Contig6	09329875J04.b1	6506	high quality base disagrees with consensus
Contig6	04118075G08.g1	7694	high quality base disagrees with consensus
Contig6	07690375J15.g1	10442	high quality base disagrees with consensus
Contig6	33839411A08.g1	10536	high quality base disagrees with consensus
Contig6	22682911L14.g1	13866	high quality base disagrees with consensus
Contig6	04065075K16.g1	17505	high quality base disagrees with consensus

Figure 4. Contig 6 navigation results for high quality (≥ 40) discrepancies, > 5 bp from an unaligned region.

In addition to these problems, I identified two large single subclone regions. The 403 bp single subclone region in contig 5 maintained Phred scores above 30 except in the low quality region I described above (Figure 2). The second single subclone region of 234 bp in contig 6 maintained Phred scores above 30 along its entire length. Single-stranded and single chemistry regions did not pose a problem for my assembly because their consensus base calls all surpassed the same threshold Phred score of 30.

Considering these issues together, my objective for the project appeared straightforward. I had to address a few internal sequence problems and close the two gaps between my three contigs. I worked toward these goals through multiple rounds of sequencing reaction orders in which I selected primers to address (cover) target areas. In this report, the entire process of primer selection and reaction ordering is referred to as reaction or read calling.

Round 1 Analysis and Reaction Ordering

As I began my work, I found that I could resolve many of my initial problems without calling additional reads. This held true for all of the high quality discrepancies in contigs 5 and 6. The single high quality discrepancy in contig 5 serves an example for the method I extended to the eight high quality discrepancies in contig 6. In contig 5, I found that only one read of four displayed a discrepancy and it appeared to be a compression of two A bases (Figure 5). The discrepant A trace had an abnormally large width compared to its neighboring traces and as such most likely encompassed two A bases. This was consistent with the three other high quality reads covering the region, which recorded two separate A bases at the same location.



Figure 5. (a) A high quality discrepancy in Contig 5 from read 04064375L18.g1. The Trace Window shows that the A trace is abnormally wide compared to neighboring traces. (b) At the same position, there are three high quality reads which identify two A bases. The Trace Window of read 09246075D09.g1 is shown. This would suggest the discrepant read displays a compression of two A bases.

Faced with the long list of high quality discrepancies in Contig 6, I feared a misassembly or some other alignment problem. Instead, visual analysis confirmed that each high quality discrepancy involved single compressions or expansions similar to the example above. I resolved these discrepancies by studying the sequence spacing and comparing the discrepant trace with other high quality reads. Instead of manually changing the discrepant base calls, I commented on the consensus sequence to indicate that I had resolved each discrepant region with visual analysis.

For the remaining problems in my sequence, visual analysis could not provide resolution and I called additional sequencing reads as part of the first round of reaction ordering (Table 1). To order sequencing reactions, I identified oligonucleotides (oligos) that would serve as primers for new reads by choosing from a list generated by Consed (Figure 6). I ensured that these oligos were between 70 and 400 bp from target regions, had melting temperatures greater than 50°C, and generated only one sequence match with *Search for String*.

Oligo	Contig	Sequence	Dir.	Chemistry	Problem
2	5	aagtgatgaagactgtggca	<--	All 3	Gap
3	4	aaccatagggcaaagcttat	<--	All 3	Gap
4	6	ccggtttgcgaatga	<--	All 3	Left Gap
5	6	cgcaatcacaggtttgac	-->	All 3	Right Gap
7	5	ccggaactaccagaactgt	<--	All 3	78bp LQ
13	6	ttagcgagtatgtaaagcca	<--	4:1	7bp LQ
14	6	tccttcggctctccatta	-->	4:1	LQ reads
15	5	tcatgttgctgtgaataggag	-->	All 3	78bp LQ

All 3 = Big Dye + dGTP + 4:1; LQ = Low Consensus Quality

Table 1. List of Round 1 reads called to address initial sequence and assembly problems.

primer number	primer bases	dis to melt cursor temp	left pos	right pos	len	-----matches----- self false vector	min qua
1	gctgtgaataggagctaattgtatc	74 58	2059	2083	25	5 14	16 68
2	gctgtgaataggagctaattgta	76 56	2059	2081	23	4 17	17 68
3	tgctgtgaataggagctaattg	78 57	2058	2079	22	3 16	16 68
4	tcatgttgctgtgaataggag	85 56	2052	2072	21	5 16	15 68
5	cgtttggggttcatgttg	98 58	2042	2059	18	5 17	17 59
6	tccgtttggggttcat	102 55	2040	2055	16	4 17	17 59
7	tcgtctctccttcaactagtttt	118 57	2017	2039	23	6 17	15 57
8	gtttccctcgtctctcctt	129 56	2010	2028	19	0 17	17 68

Figure 6. Primers window generated by Consed.

I used my identified primers to order single sequencing reactions with Big Dye chemistry or multiple reactions with Big Dye, dGTP, and 4:1 chemistries. To address the gaps between my contigs, I ordered reactions sequencing into the gaps from each of the four gap ends. In contig 5 I ordered reactions from both directions to sequence the 78 bp low consensus quality, single subclone region. In contig 6 I ordered one primer to cover the 7 bp low quality string of G bases. I also ordered a primer in contig 6 to cover a double stranded region (16155-16181) where the consensus stayed above Phred 25 but all the individual reads had low quality scores.

Round 1 Comparison with Autofinish

I ran the program Autofinish to assess how my primer selection compared to a computer-generated output (Table 2). I found that 6 out of 12 computer-generated reactions addressed the same problem regions I identified. Both Autofinish and I called four reactions to sequence into the two sequence gaps. In contig 5, Autofinish called a single reaction to cover the 78 bp low quality, single subclone region that I called two reactions to cover. In contig 6, both Autofinish and I called a reaction to cover the 7 bp low quality region.

I only called one reaction that did not have a comparable primer in the Autofinish results. This reaction covered the double-stranded region (16155-16181) in contig 6 that met Phred consensus standards but relied entirely on low quality reads. In retrospect, I realize this reaction was unnecessary and Autofinish made a better choice in not calling for a sequencing reaction here.

Autofinish designed six sequencing reactions that did not have a comparable reaction in my order list. In contigs 4 and 5, Autofinish called two reads to cover low quality regions at the clone ends, which do not need to meet threshold Phred quality in our study. In contigs 4, 5, and 6, Autofinish designed three reads to address single stranded and single chemistry regions. These reads were unnecessary because the single stranded and single chemistry regions met the Phred 30 threshold we are using. Autofinish also called a read in contig 5 to cover the 403 bp single subclone region. I only addressed the part of this single subclone region below Phred 30 with my first round of reactions. In this last case Autofinish made the better choice, since in subsequent rounds of reaction ordering, I did call reads for single subclone regions.

Overall, my analysis of Phred quality in single stranded and single chemistry regions reduced the number of reactions I ordered compared to Autofinish. Autofinish, however, addressed a single subclone region that I did not identify until subsequent

rounds. Neither Autofinish nor I ordered a reaction to cover the 234 bp single subclone region in contig 6.

Oligo	Contig	Autofinish Primer Sequence	Dir.	Problem (Similar Primer)
1	4	gacaattgaaaatgtacacatagaa	<--	Gap (3)
2	4	aagacggattaccgcaac	<--	489 bp SS/SC
3	4	cggcgactattcacagtatcc	-->	clone end
4	5	tgatgaagactgtggaattat	<--	Gap (2)
5	5	gaatcagagtacgggcga	<--	121 bp SS/SC
6	5	cttagagcagagatagtcaaaggt	-->	646 bp SS/SC, 403 bp SS
7	5	gacttaaacattctcacgataca	<--	78bp LQ (7, 15)
8	5	gatgctcagtcgagacataag	-->	clone end
9	6	cgctgacaacttctctctg	<--	Left Gap (4)
10	6	agtgcgctaatactgatagata	<--	136 bp SS/SC
11	6	accattttctgaaataatgacg	<--	7bp LQ (13)
12	6	tgtattttcgatatgcattatgta	-->	Right Gap (5)

SS/SC = Single Strand/Single Chemistry; SS = Single Subclone; LQ = Low Consensus Quality

Table 2. Comparison of Autofinish primers with my Round 1 reaction calls. Note my primer 14 (Table 1) was not included in calls made by Autofinish.

Round 2 Analysis and Reaction Ordering

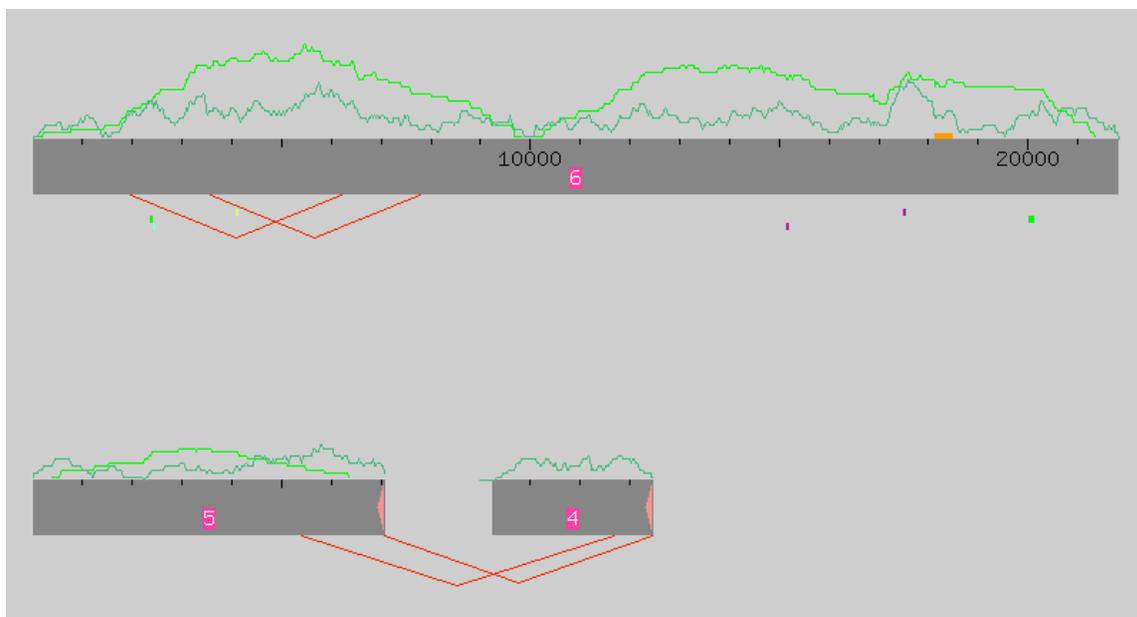
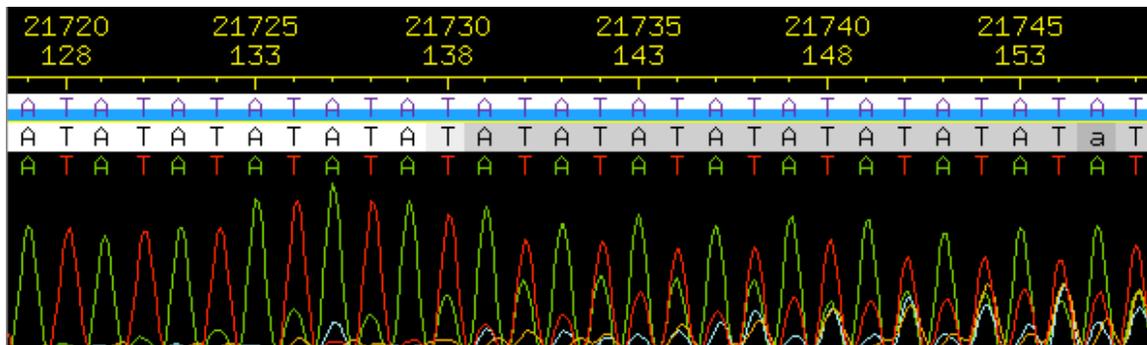


Figure 7. Assembly View after the incorporation of Round 1 reads. dGTP reactions are missing.

Many of the problems in my sequence remained after the incorporation of Round 1 reads (Figure 7). This was partially due to the fact that the dGTP reactions from Round 1 were missing from the read data that was returned to me.

The reactions I called to cover the sequence gaps had mixed results. In fact, none of the reactions I personally ordered contributed meaningful data. Another student working on the same project generated some reads with high quality data, occasionally with the same primers I identified. One such read extended the sequence of contig 5, but did not provide enough data to close a gap. In contrast, no sequencing reactions extended the left gap end of contig 6. Sequencing on the right gap end of contig 6 began successfully, but failed in a TA repeat region (Figure 8.a). In a similar manner, the gap end sequencing of contig 4 failed in a mononucleotide run of 23 C's (Figure 8.b).

Round 1 reads did not resolve either of my low consensus quality regions. Neither of the two reactions covering the 78 bp low consensus quality, single subclone region in contig 5 contributed meaningful data. The reactions I called on the 7 bp low consensus quality region in contig 6 provided good data, but a dye blob interrupted the traces exactly at the region of low quality (Figure 9.a). Increasing the vertical gain of the traces did not reveal additional sequence data in the target area (Figure 9.b).



(a)



(b)

Figure 8. (a) Trace Window of 08XBAB-205H09_5a.b1 shows a sequencing failure in a TA repeat region (51 repeats, 21665-21780) at the right gap end of contig 6. (b) Trace Window of read 08XBAB-205H09_3a.b1 shows a sequencing failure in a mononucleotide run of 23 C's at the gap end of contig 4.

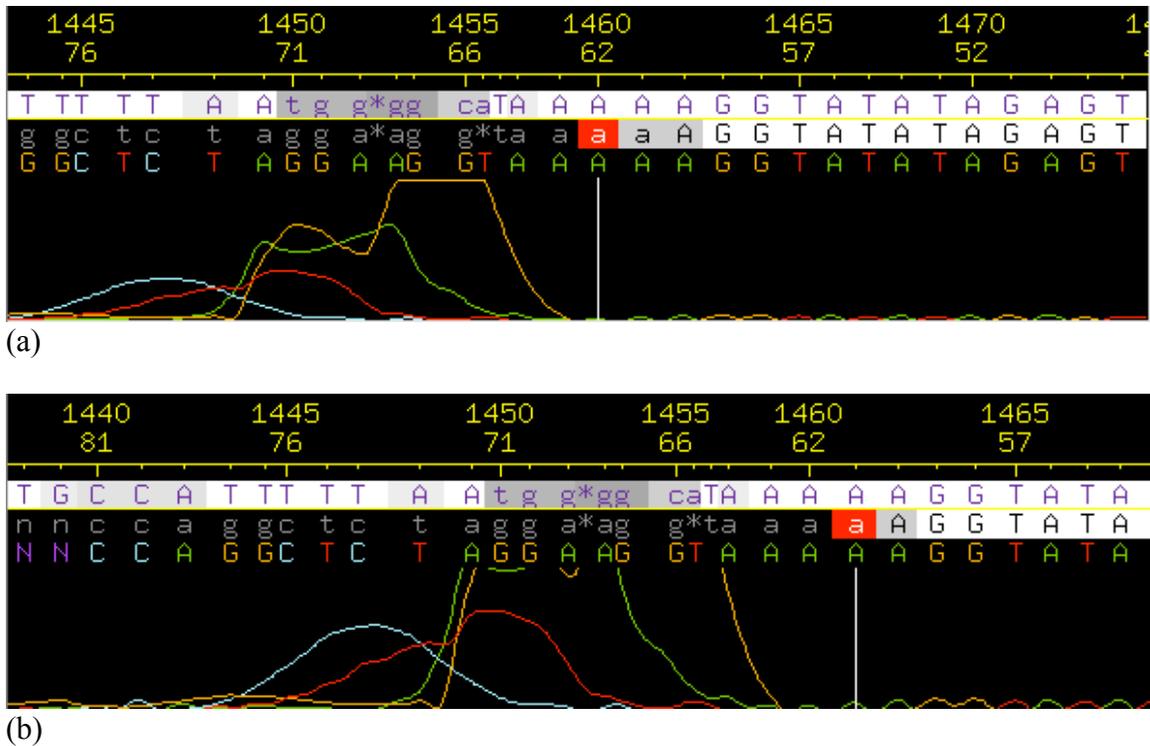


Figure 9. (a) Trace Window of read 08XBAB-205H09_t13a.b1 showing the effects of a sequencing dye blob in contig 6. (b) Increasing the vertical gain of the traces did not reveal additional sequence data.

The somewhat unsatisfactory results of Round 1 helped me develop a finer strategy for Round 2 reaction ordering (Table 3). I knew I could not sequence past either the gap end TA repeat of contig 6 or the C mononucleotide run of contig 4. Thus, to close the gaps I ordered two new primers on the extended gap end of contig 5 and two primers on the left gap end of contig 6. For the 78 bp low quality, single subclone region in contig 5 I ordered two new primers from both directions, but backed them to a greater distance from the low quality sequence. Similarly, in contig 6 I ordered a new primer for the 7 bp low quality region, but backed it farther from the region to avoid dye blob interference.

It was also during Round 2 of reaction ordering that I began to address my single subclone regions. Both regions maintained the threshold of Phred 30 (except in the 78 bp low quality region in contig 5), but I called reads in order to improve the consensus sequence and simplify annotation. I ordered one primer to cover the 234 bp single subclone region in contig 6. I did not order one specifically for the 403 bp single subclone region in contig 5 because the primers for the 78 bp low quality region in contig 5 covered the same sequence.

Oligo	Contig	Sequence	Dir.	Chemistry	Problem
16	6	caggataggacacgcaaa	<--	All 3	Left Gap
17	6	taaccggttgaacgctt	<--	All 3	Left Gap
18	6	acggaccgcttttcac	<--	All 3	7bp LQ
19	5	cggtatcaatacagcttctgtagt	<--	All 3	Gap
20	5	aaaatgctgcttagccataa	<--	All 3	Gap
21	5	ccgtcaatgccgatagat	-->	All 3	78bp LQ
22	5	caactgttggtccgtctagtct	<--	All 3	78bp LQ
23	6	cgctgacatcaaccgaa	<--	All 3	234bp SS

All 3 = Big Dye + dGTP + 4:1; LQ = Low Consensus Quality; SS = Single Subclone

Table 3. List of Round 2 reads called to address remaining sequence and assembly problems.

Round 3 Analysis and Reaction Ordering

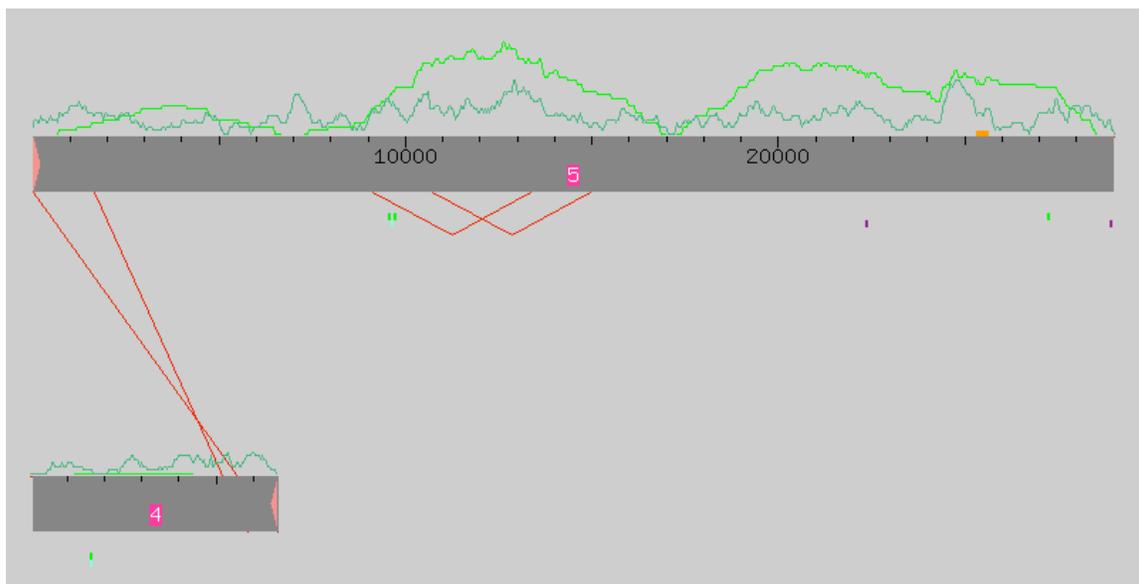


Figure 10. Assembly View after the incorporation of missing Round 1 dGTP reads, Round 2 reads, missing Whole Genome Shotgun reads, and a missing Whole Genome Shotgun contig.

For my final round of reaction ordering, I had much more data to work with. First, the missing dGTP reactions from Round 1 accompanied the results of my Round 2 reactions. Also, to supplement my own reaction orders I incorporated additional data related to my fosmid from the Whole Genome Shotgun (WGS) database. We found that some WGS reads did not appear in my initial assembly even though they mapped to this region. Approximately 10% of the WGS read files originated in sequencing centers with unconventional file nomenclature. These files were not included when my 205-H09 project file was first constructed. Before Round 3, these files were identified and incorporated into my project to improve and extend my sequence data.

Another way I incorporated additional WGS reads into my project involved adding an entire missing contig into my sequence. A BLAST search was conducted on the WGS data to look for small contigs that might fit into my sequence gaps. After a careful search, one 1.2 kb contig was found which shared a single forward-reverse read pair with contig 4. Since two or more forward-reverse pairs are required for PhredPhrap to assemble contigs together, this smaller contig did not appear in my assembly. I incorporated it, however, because my assembly and the 1.2 kb contig shared another connection. Exons for a gene called UNC-13 appeared in each of my contigs as well as in the 1.2 kb contig. The single forward-reverse pair, the connected exons, and the appropriate size of the 1.2 kb contig compared to the gap provided enough encouragement for me to incorporate the 1.2 kb contig into my assembly.

With this influx of missing data, my project finally took a step forward (Figure 10). After running PhredPhrap, the gap between my contigs 5 and 6 closed and formed a single new contig 5. At the same time, contig 4 dramatically increased in size due to the incorporation of the 1.2 kb contig from the WGS database. This incorporation reduced the remaining gap between contig 4 and the larger contig 5.

With only one round of reactions left, I selected multiple primers to cover each remaining sequence problem (Table 4). The gap end of the new contig 5 still included a confounding TA repeat sequence, but I used a small amount of low quality sequence beyond the TA repeat to call two primers sequencing into the gap. The gap end of contig 4, now including the added 1.2 kb sequence, no longer had any repetitive obstacle to sequencing. Taking advantage of this, I ordered three primers at different intervals that would all sequence into my gap region.

I also ordered multiple reactions to address the remaining internal sequence problems. The new 1.2 kb extension of contig 4 included a 140 bp single subclone, low consensus quality region. I ordered two primers to sequence this region from the opposite strand. Contig 4 also contained another single subclone region (112 bp) that I addressed with a single primer. The new contig 5 had two single subclone regions (94 bp and 220 bp) and I ordered two primers to cover these regions.

Oligo	Contig	Sequence	Dir.	Chem.	Problem
24	5	tgtaagtacgaatgcttggaa	-->	All 3	Gap
25	5	cgaatgcttgaagggtt	-->	All 3	Gap
26	4	tcgttcaacaagcagttagtatt	<--	All 3	Gap
27	4	ccaccataatcaggtattgta	<--	All 3	Gap
28	4	agaaactccaaccattcaatta	<--	All 3	Gap
29	5	cggcggatacaattagc	-->	All 3	94bp SS
30	5	gcttctccggcaggatt	<--	All 3	220bp SS
31	4	ttgcccggtacctgtaa	<--	All 3	140bp LQ
32	4	ctgtgcaagttcgatattatg	<--	All 3	140bp LQ
33	4	tgtacgaatagaaataaccggttag	<--	All 3	112bp SS

All 3 = Big Dye + dGTP + 4:1; LQ = Low Consensus Quality; SS = Single Subclone

Table 4. List of Round 3 reads called to address remaining sequence and assembly problems.

Final Analysis

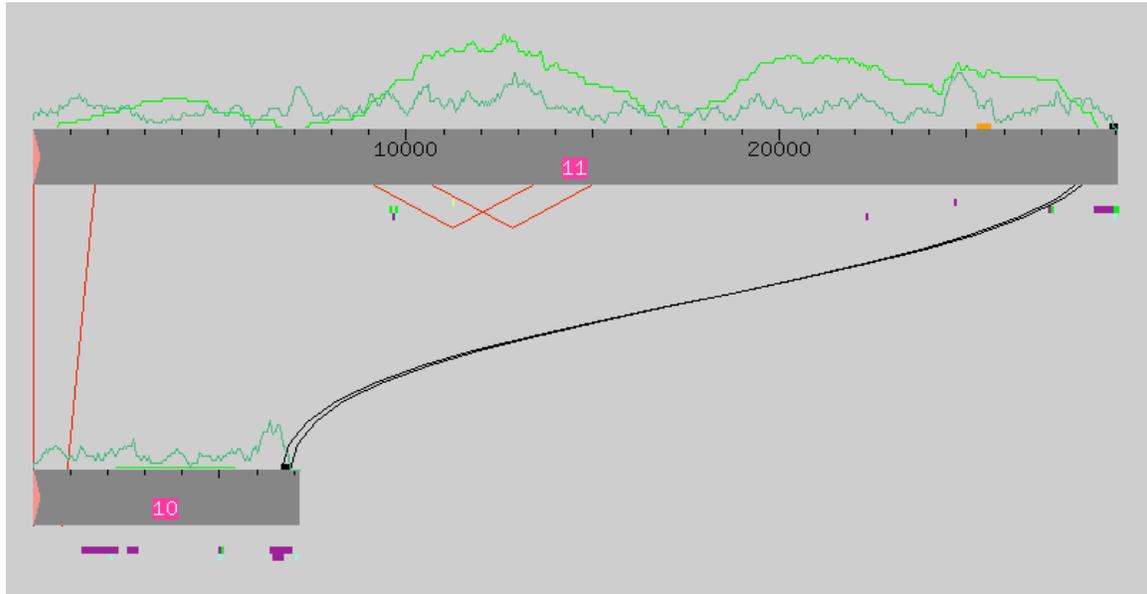


Figure 11. Assembly View after incorporation of Round 3 reads and reads from adjacent clone 200-N19.

In my final iteration of PhredPhrap, I once again pulled in additional information to supplement my own Round 3 reaction results. I included the reads from a green clone, 200-N19, which overlapped my own fosmid. About half the 200-N19 reads did not map into my two main contigs (now contigs 10 and 11), either because they no longer overlapped or because of polymorphisms. I found, for example, a polymorphism involving my gap end TA repeat sequence. Subclone 200-N19 has only 11 repeats of the TA dinucleotide, while my subclone 205-H09 has 51 repeats.

Despite the new data, my assembly view still maintained a gap between the main contigs 10 and 11 (Figure 11). A new sequence match identified by Crossmatch, however, indicated that a forced join might be possible. The match included the 51 dinucleotide TA repeat sequence, but also 61 bp of unique sequence. I began exploring the possibility of a forced join by complementing the sequence of contig 10 so that the clone ends would face each other and a single contig could be formed (Figure 12).

Final Status

For a clone that others have given up on and I did not think I would finish, I am proud of the progress I made. I reached my finishing goal of forming a single contig with all bases in the consensus having Phred scores greater than 25. There are still fourteen regions with only one strand and one chemistry, but all meet the threshold of Phred 30 (Figure 17). There is one region covered by only a single subclone. It encompasses 13 bp at the right clone end of the sequence (35696-35708) and has a comment tag. It also meets the Phred 30 requirement. All of the high quality discrepancies I encountered were resolved by visual analysis have comments in the consensus sequence.

Contig		Start Pos	End Pos	
Contig12	(consensus)	1	18	18 bp
Contig12	(consensus)	5862	6076	217 bp
Contig12	(consensus)	9286	9469	189 bp
Contig12	(consensus)	15653	15740	89 bp
Contig12	(consensus)	16343	16483	142 bp
Contig12	(consensus)	17890	18034	153 bp
Contig12	(consensus)	18633	19145	524 bp
Contig12	(consensus)	28952	29007	56 bp
Contig12	(consensus)	29855	29919	67 bp
Contig12	(consensus)	31296	31306	11 bp
Contig12	(consensus)	32045	32085	41 bp
Contig12	(consensus)	32251	32279	29 bp
Contig12	(consensus)	32601	33152	567 bp
Contig12	(consensus)	35696	35708	13 bp

Figure 17. Regions covered by sequence from only one strand and one chemistry in contig 12.

There are no contigs over 2 kb outside of the assembly and I did not use fake reads or run PCR reactions. The clone ends are identified and delineate a total clone length of 35708 bp. This relatively short clone length may be due to a failure in library construction. In the final assembly view, there are two inconsistent forward reverse pairs that are not associated with the clone ends (Figure 16). Consed identifies them as too far apart in the sequence, but they exceed the maximum of 4158 bp by only 25-42 bp. These slightly larger subclones may be due to the random annealing of extra DNA to the subclone or a natural experimental variation.

When I ran the program BLAST to check for contamination from laboratory organisms, I found no significant similarity between my sequence and any common microbes.

My finished sequence had two A mononucleotide runs (16, 25), one T mononucleotide run (17), and one C mononucleotide run (23). I checked the sequence adjacent to each and the bases were called correctly. *Search for String* identified no X's or N's in the consensus.

The restriction digests I completed had varied results. The outputs of *Hind*III and *Eco*RI displayed one or more discrepant bands between real and in silico digests. Many of the discrepancies occurred around the region of the contig where I made the forced join. This made me consider that the join might be inappropriate. Still, digestion by *Eco*RV and *Sac*I showed perfect correspondence between real and in silico bands 1 kb to 10 kb in size (Figure 18). To be sure my join was correct, I cross-checked the discrepant regions of the *Hind*III and *Eco*RI digests with the same regions in my *Eco*RV and *Sac*I digests, making sure the forced join region was correctly banded in at least two digests. I also checked the consensus sequence of discrepant bands to make sure there were no significant sequence errors. The *Eco*RV and *Sac*I digests appear to be correct representations of the region and confirm my assembly of reads and former contigs in the final sequence.

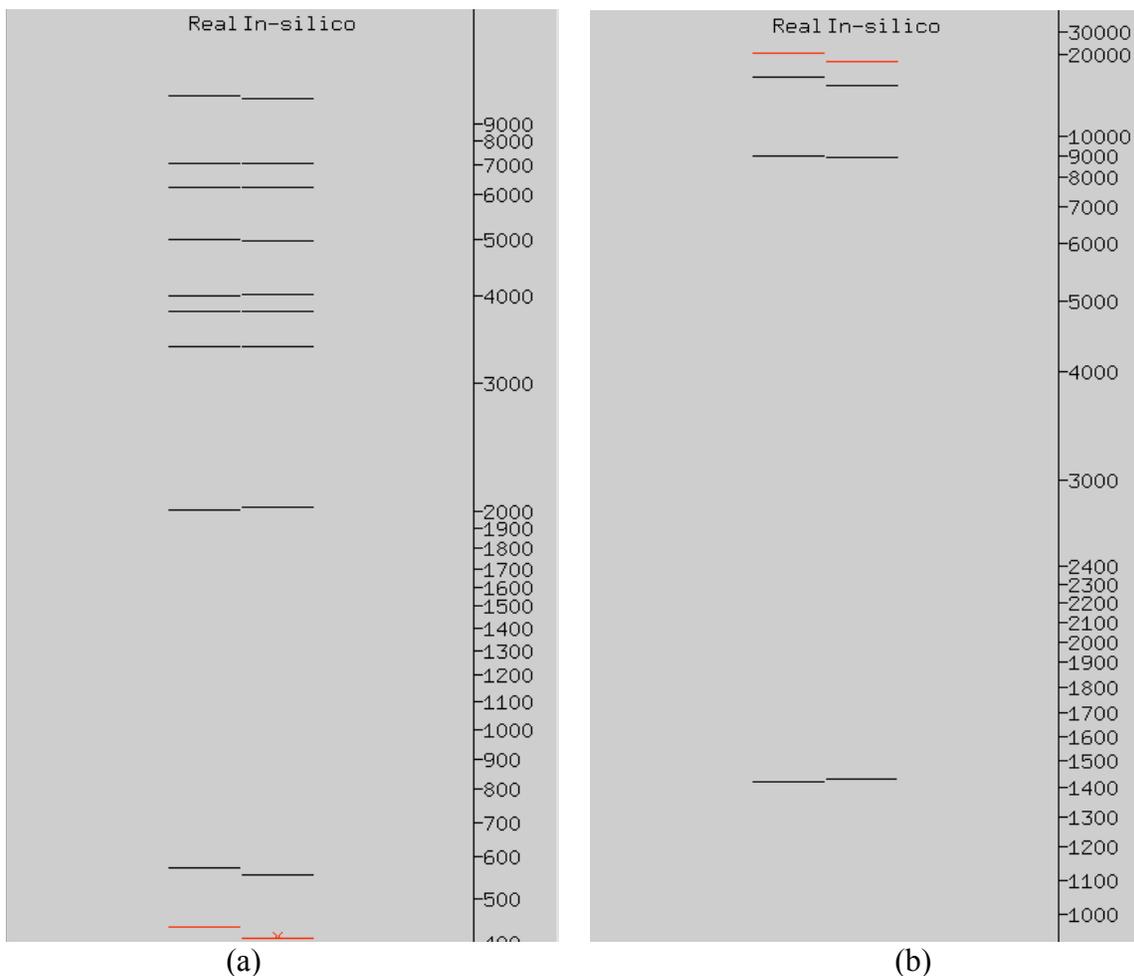


Figure 18. (a) *Sac*I restriction digest. (b) *Eco*RV restriction digest.

Further actions to improve my finished sequence include possibly covering the single strand and single chemistry regions as well as double checking the forced join region and restriction digests for accuracy. It may also be rewarding to investigate the TA repeat region (28850-28951) and the implications of polymorphisms between clones.