Finishing *Drosophila grimshawi* Fosmid Clone DGA10F16

Jerome M Molleston

3/5/2009

Abstract
A sequence assembly of *Drosophila grimshawi* fosmid DGA10F16, initially consisting of four contigs, was assembled further. Gaps between several of the contigs were sequenced across, facilitating joins, and a polymorphic region was identified as such. In addition, a low quality sequence area was identified as a gap between the sequence consisting of ~2 kb. Sequencing across this gap was attempted but was unsuccessful, resulting in a final assembly of two contigs.

Introduction

The *Drosophila melanogaster* fourth chromosome, also known as the dot chromosome, has many features of both heterochromatin and euchromatin. It is repeat rich, does not recombine, and is associated with heterochromatic proteins, but it encodes ~80 expressed genes as well. Bio 4342 seeks to sequence dot chromosomes of various *Drosophila* species, in the hopes that comparison of this important chromosome's sequences in multiple species will assist in the understanding of chromatin organization and its use in gene regulation. This semester, sequencing work was begun on *D. grimshawi*, the Hawaiian fruit fly, including my project, DGA10F16. As it stands, DGA10F16 is mostly finished, but has one gap remaining, which will require later attention.

Workflow

Initial Assessment
My initial assembly view for DGA10F16 (Figure 1) consisted of four contigs. Contigs 15 and 13 clearly had matching forward-reverse pairs between them, as did 13 and 16, suggesting simple sequence gaps between them. However, contig 14 had discrepant forward-reverse pairs to the middle of contig 16, an area to which it had 99% sequence similarity as reported by Crossmatch. This suggested either a near-identical repeat in 16, which wasn't placed there, or a polymorphism preventing Consed from aligning the sequences.

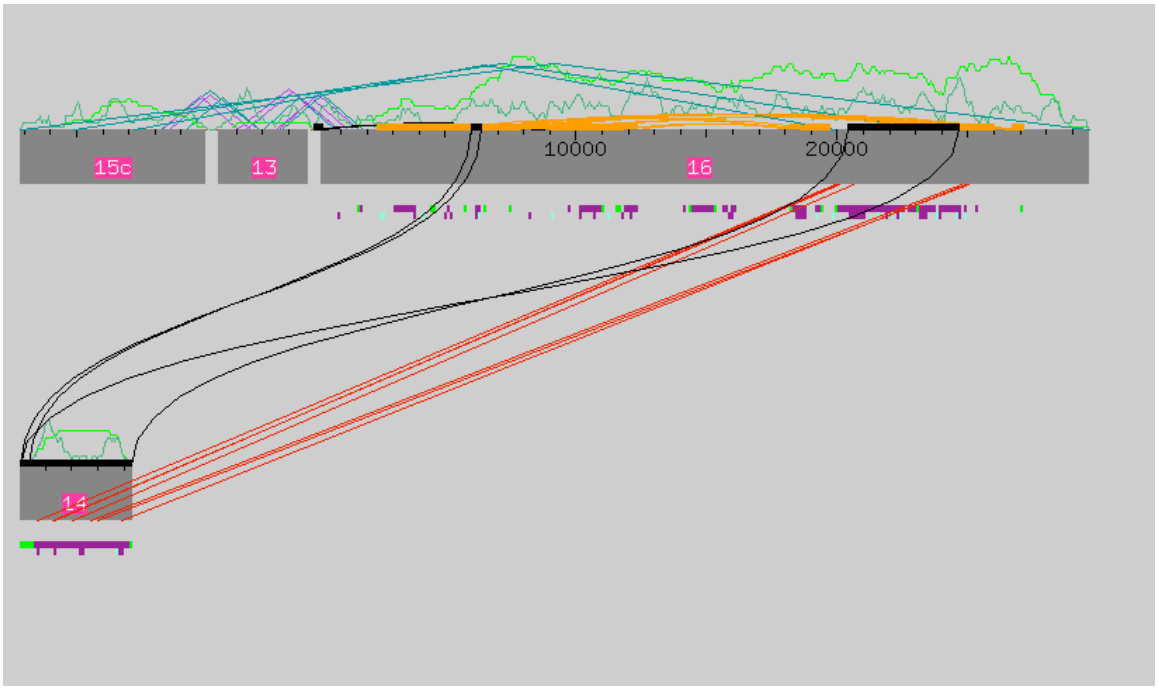

Figure 1: Initial Assembly View with Crossmatch results.

Initial Work

I began by converting contig 15 to its complement to make it match the orientation of the other contigs. I then ran an *in silico* digest in the order 15-13-16, both to estimate the size of the gaps between 15, 13, and 16, as well as to look for evidence of 4 kb missing from contig 16, which would suggest that 14 belonged there as a repeat.

Real Frag Size	In Silico Size	Position
11555	11598	Contig16 (16069-27667)
10145	10020	part vector/part insert Contig16 (27667-29580)
6683	6610	Contig16 (3856-10466)
~~5599~~	5603	Contig16 (10466-16069)
4832		
4395		
	4257	Contig15 (3649-7187 Contig13 (1-606)
	3447	Contig13 (3910-4084 Contig16 (1-3365)
2938	~~2926~~	Contig15 (723-3649)
	954	Contig13 (2956-3910)
	755	part vector/part insert Contig15 (1-723)
	671	Contig13 (2285-2956)
	580	Contig13 (832-1412)
	566	Contig13 (1719-2285)
	491	Contig16 (3365-3856)
	307	Contig13 (1412-1719)
	226	Contig13 (606-832)

Figure 2: HindIII digestion of my initial project.

The HindIII digest, shown in Figure 2, revealed mismatches in size between 15 and 13 as well as between 13 and 16, as expected. Depending on which band in the real digest

matched which *in silico* band, the gap between 15 and 13 seemed to be either about 150 bp or 600 bp, leaving the 13-16 gap to be either about 1400 bp or 1000 bp. The other digests seemed to support these figures (data not shown). Both of the gaps seemed to be amenable to relatively easy sequencing, though the 13-16 gap was wide enough that more than one run may have been necessary. In addition, the area from 20000-24000 of contig 16, where it matched contig 14, did not cause any discrepant bands, suggesting that the 4 kb contig represented a single polymorphic area in 16 rather than a repeat. Finally, the added size of all real bands (with vector sequence in the 10145 bp band subtracted) suggested a fosmid of about 37000 base pairs, which did not suggest that 4 kb were missing from my assembly.

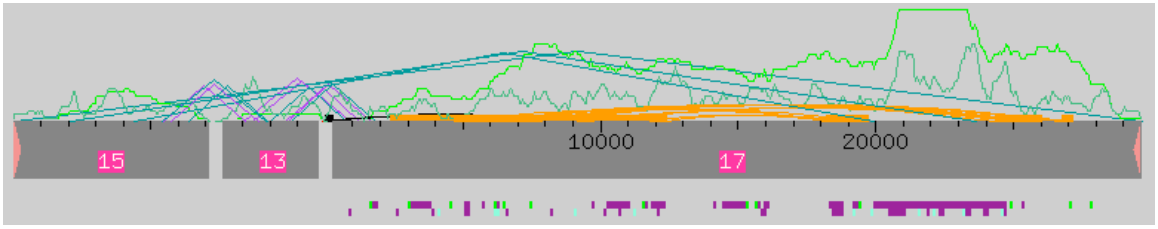

Figure 3: Experimental assembly with contig 14 on top of its match in 16.

With this digestion data in mind, I tested the hypothesis that contig 14 represented a single polymorphic region in 16 by forcing a join between the two with contig 14 assembled at the same location as its equivalent reads in contig 16. As seen in Figure 3, this satisfied the forward-reverse pairs completely. However, it did create an abnormally high read density in the putative polymorphic area, which was suggestive of a repeat.

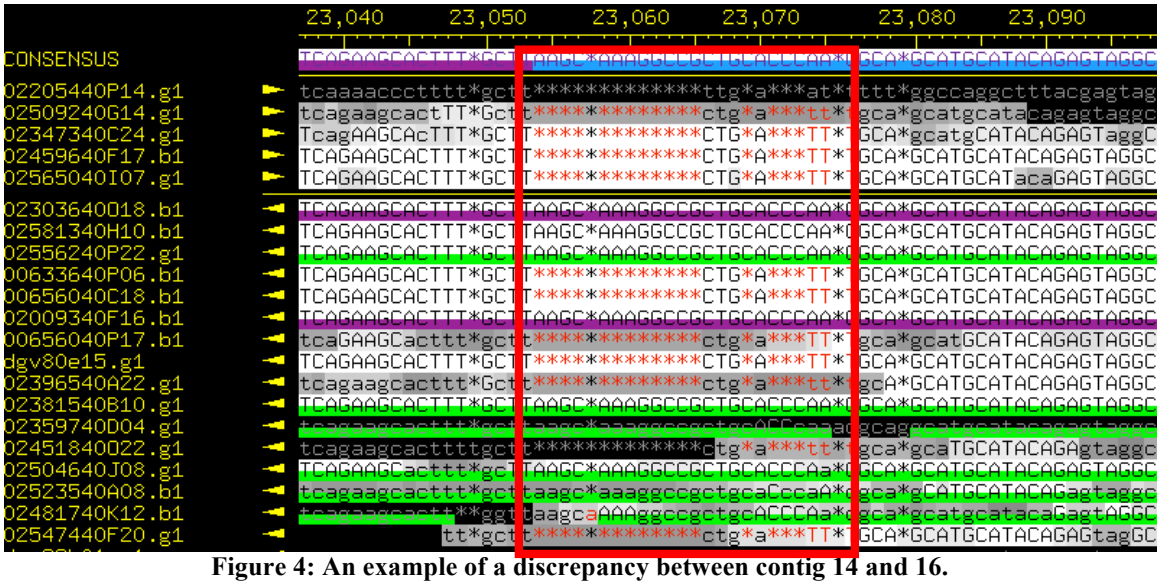

Figure 4: An example of a discrepancy between contig 14 and 16.

A number of high quality discrepancies were found in the area where contig 14 was inserted, creating in each area two populations of sequences. Most of these regions were single bases, but Figure 4 represents an area where two very different sequences are reported by the reads from the two contigs. I used this area as a "litmus test" for my hypothesis that the discrepancies represented a polymorphism between contig 14 and

contig 16. If this were true, then a sequencing reaction run across this area off of my fosmid would only report one sequence, as the fosmid must have been cloned from one or the other copy of this area of DNA. If there were a repeat, then I would get two separate populations of sequences or an uninterpretable double sequence.

First Round of Sequencing and Comparison with Autofinish

Problem addressed	Direction of sequencing	Oligo sequence	Oligo location	Chemistry	Oligo name	Special	Success?
15-13 gap	→	Cgaagcgaaaatcgaag	15:6768-6784	All	DGA10F16.1		Yes
	←	cagttgccaaattaaatgaaat	13:518-539	All	DGA10F16.3		Yes
13-16 gap	→	agcaagcgcatagattataca	13:3565-3585	All	DGA10F16.4		Yes
	←	aattccaccggccca	16:345-359	All	DGA10F16.10		Yes
Low quality 16:1396-1610	→	ccctggttcaaataatttatgac	16:1286-1308	4:1	DGA10F16.6		No
	←	gaatatttaaaatttgcaatgaaaa	16:1921-1945	4:1	DGA10F16.7		No
Low quality 16:4585-4587	→	gaaacggcgacgatgta	16:4342-4358	All	DGA10F16.8	Relaxed "match elsewhere" parameters	Yes
	←	tcaaatgaattttcagtaattgc	16:4822-4844	All	DGA10F16.9	Relaxed "match elsewhere" parameters	No
Putative polymorphism 16:23054-23076	→	gagaagagctagggtagtcttcat	16:22909-22932	All	DGA10F16.12	From subclone	No
	←	ttatcaaaggcaccttctatatcta	16:23184-23208	All	DGA10F16.13	From subclone	No

Table 1: Oligos called for first round of sequencing.

Table 1 depicts the first round of sequences which I ordered. Sequencing reads were called across the 15-13 and 13-16 gaps. In addition, several low-quality consensus areas were sequenced, as well as my litmus test sequence. The litmus test sequence required primers to be chosen using Consed's "from subclone" algorithm, as any primers chosen would have matched contig 14 and thus would have been rejected by Consed. In addition, one of the low quality areas required primer parameters to be slightly relaxed in order to find a good primer.

Contig	Left position	Right position	Sequence	Direction	Unique?
Contig13	-603	326	tcctcaaattttcaagtgtca	←	No
Contig13	-89	840	tgcactgttctcctttactgt	←	No
Contig13	3464	4393	aaacagtgattatcttaagagaccc	→	No
Contig13	3888	4817	gcattaatgttaagcagttgga	→	No
Contig14	-821	108	cgaggcggctaacaag	←	Yes
Contig14	59	988	gtgtgtctccctatgcca	→	Yes
Contig14	3876	4805	actcaaggggcacgc	→	Yes
Contig15	-542	387	cgaaggtatatttcggatttt	←	Yes
Contig15	5215	6144	ttgttattagttgttcgaattgtta	→	Yes
Contig15	7084	8013	caaaattccaatagccaataaaa	→	No
Contig16	-634	295	gggctccgcttgaaaat	←	No
Contig16	1299	2228	cgcccctggttcaaat	→	No
Contig16	3952	4881	cgggcttatttcttatctgatct	→	No
Contig16	4358	5287	gaaacggcgacgatgt	→	No
Contig16	29454	30383	gcgtcgctggactagtta	→	Yes

Table 2: Autofinish reads.

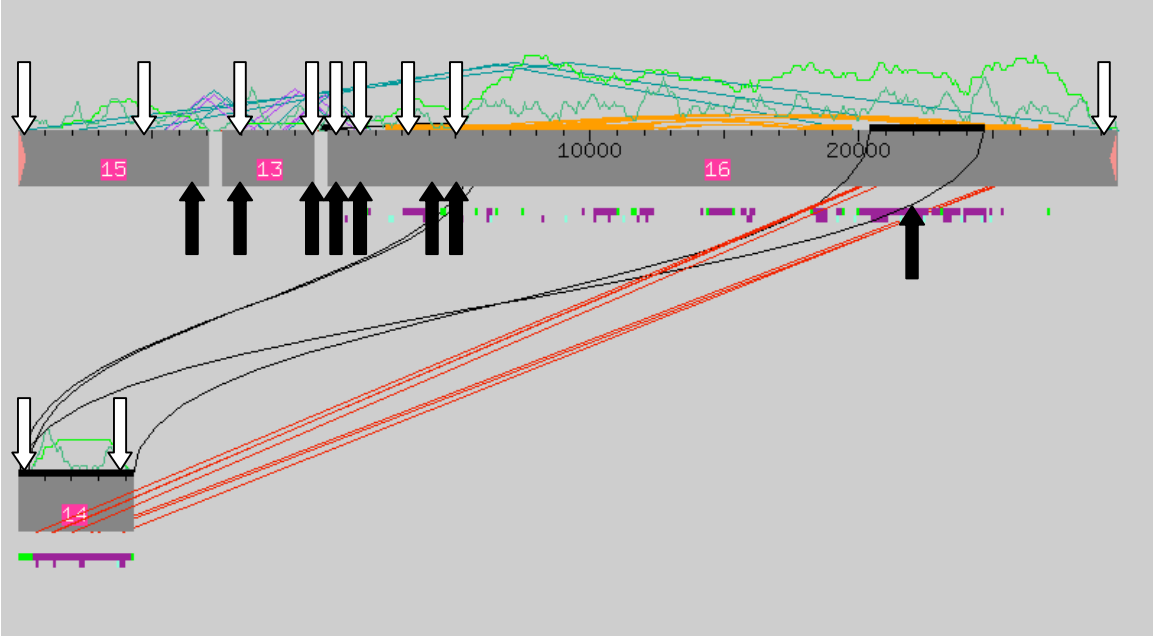

Figure 5: Comparison of my reads (black) with autofinish's reads (white).

Autofinish called reads (depicted in Figure 5) that were mostly redundant with my reads, attempting to close gaps and correct low quality consensus regions of the fosmid. Autofinish uniquely called reads at the ends of the fosmid, where I didn't try to extend its sequence, and near the ends of contig 14, which match for the most part in contig 16 according to my polymorphism hypothesis. In addition, it called one read in contig 15 in a single strand/single chemistry region, which I didn't call any reads across since it was high quality. The one set of reads which I called and autofinish didn't were my reads

across the polymorphism area of contig 16, which it presumably didn't call since it regarded the two different sequences as two separate contigs.

Results of First Round of Sequencing

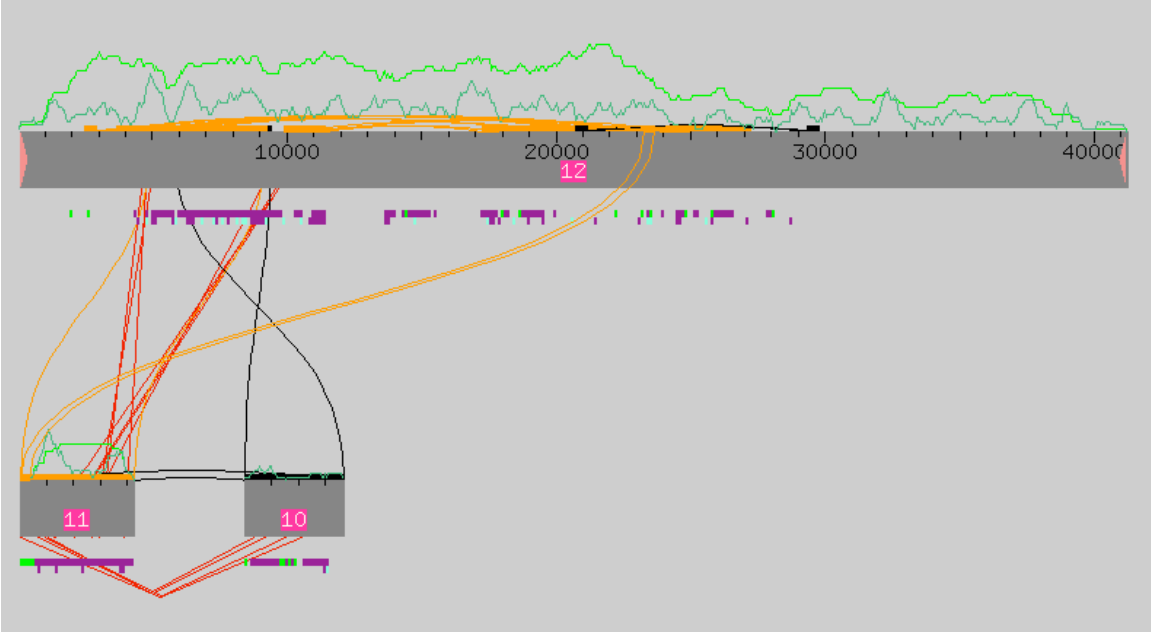

Figure 6: Assembly view after rerunning phredPhrap with new reads.

New reads were incorporated by rerunning phredPhrap. The new assembly (Figure 6) joined the gap between 15 and 13 and that between 13 and 16, but split the putative polymorphism-containing contig into two. Comparison of the three almost-identical contig sequences (from contigs 11, 10, and 12) revealed only two types of sequences at each discrepant base, so there was no evidence indicating that I should discard my polymorphism hypothesis.

Searching for the new reads in the assembly revealed that most of the reads provided useful data. Exceptions included oligos 6 and 7, which both produced significant low quality and double sequence and left behind a low quality region. Oligo 9 didn't provide any useful sequence and in fact was assembled into the wrong part of the read, but oligo 8, its partner, yielded enough sequence to fix the low quality area which oligo 9 was meant to fix. Oligos 12 and 13, my "litmus test" oligos, were low quality at the putative polymorphism area, which was in the 5' low quality areas of those reads. However, they all matched one copy of the polymorphism, lending some credence to this region being a polymorphism and not a misassembly of a repeat.

Second Round of Sequencing

Problem addressed	Direction of sequencing	Oligo sequence	Oligo location	Chemistry	Oligo name	Special
Low quality 28157-28261	→	Aatcatatacatacggggaatattt	27710-27734	All	DGA10F16.23	
	←	Ggctggttacatccgaga	28478-28495	All	DGA10F16.22	
Low quality 33373	→	cggagttgccaatcgaa	33133-33149	All	DGA10F16.14	
	←	ttgttttggttttcttttcg	33620-33639	All	DGA10F16.15	
1 strand/1 chem 25290-25365	←	ggcttatttcttatctgatcttgt	25718-25741	4:1	DGA10F16.24	No good primer available from left
1 strand/1 chem 26840-26985	→	ttaatcatttattttcgatttcatt	26623-26647	4:1	DGA10F16.25	
	←	cgccataacattaacattcactg	27368-27390	4:1	DGA10F16.26	
1 strand/1 chem 39597-40290	→	gagtgagcaagagagagatagc	39516-39537	4:1	DGA10F16.27	
	←	cgatgataccgttgttaatttg	40448-40469	4:1	DGA10F16.28	
Putative polymorphism 6595-6650	→	ttgtctggaccgtagcagt	6370-6388	All	DGA10F16.29	
	←	cctggaatggacggc	6865-6879	All	DGA10F16.30	

Table 3: Oligos called for second round of sequencing.

My second round of sequencing reactions addressed the problems that still remained from the first. I targeted two low quality consensus regions: one remaining from round one, for which I chose new primers, and one leftover where the sequencing reactions run across one of the original contig gaps left a slight region of low quality sequence. In addition, I began to address single strand/single chemistry regions; for these I simply called 4:1 sequencing chemistry, as all I needed to do was confirm data I already had with a second sequencing chemistry. Finally, I sequenced my putative polymorphic "litmus test" sequence again, moving the primers farther from that region to ensure high quality sequence in that area.

Results of Second Round of Sequencing

The assembly after running phredPhrap with the second round of sequencing reads appeared no different from the previous one, and so is not depicted here. Most reads worked; several of the single strand/single chemistry regions acquired the needed depth in their data, and one of the two low quality areas was made high quality. However, oligos 22, 23, 25, and 26 failed to correct the problems to which they were assigned, and thus a region of low quality consensus, as well as a nearby region of single strand/single chemistry, remained.

The reads called over the putative polymorphism all matched the heavily padded population of sequences (see Figure 4 above), and were now high quality. This indicated that the fosmid, which I was given contained only one copy of the sequence in this area, as would be true if it was a polymorphism rather than a repeat. As already stated, the digest was not missing a 4 kb segment, and all the discrepant reads in this area had only two populations of sequences, indicating the sequence was from each of the two copies of the dot chromosome. This data led me to conclude that the original 4 kb outlier is indeed either a polymorphic area which belongs in the same area as its counterpart in the main contig, or a repeat elsewhere in the chromosome which was mistakenly placed into this fosmid's set of reads. I combined these contigs and changed the consensus in the "litmus test" region to match the reads I had called (Figure 7).

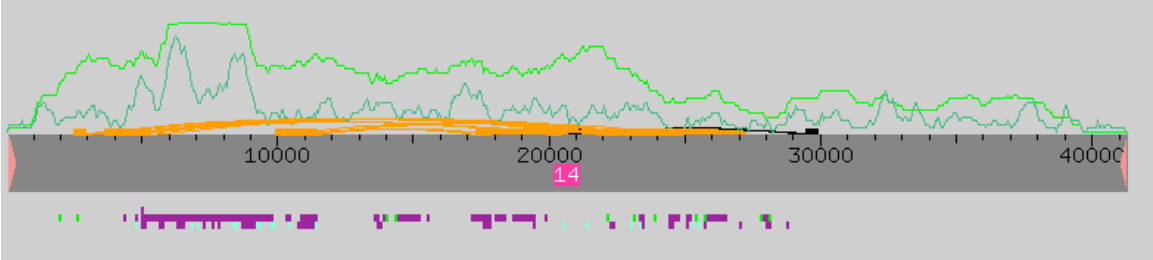

Figure 7: The assembly after two rounds of reactions and incorporation of the polymorphism.

While analyzing the digest of this assembly, I found that while the polymorphic region still matched the digest very well, all restriction digests had an *in silico* band which was about 1.2 kb lower than the equivalent real band. In every digest, this region contained sequence between 26286 and 29938, with HindIII providing the best-defined numbers (Figure 8).

Real Frag Size	In Silico Size	Position
11555	11599	Contig14 (1983-13582)
10145	10088	part vector/part insert Contig14 (1-1983)
6683	6610	Contig14 (19185-25795)
[illegible]	5603	Contig14 (13582-19185)
4832		
4395	4436	Contig14 (33242-37678)
	3652	Contig14 (26286-29938)
2938	[illegible]	Contig14 (37678-40604)
	954	Contig14 (29938-30892)
	761	part vector/part insert Contig14 (40604-41331)
	671	Contig14 (30892-31563)
	580	Contig14 (32436-33016)
	566	Contig14 (31563-32129)
	491	Contig14 (25795-26286)
	307	Contig14 (32129-32436)
	226	Contig14 (33016-33242)

Figure 8: HindIII digest of assembly after sequencing round 2.

The only anomalous sequence in this area was the persistent low quality sequence at 28157-28261, so I searched the reads in this area for evidence of a misassembly. Inspecting the traces of the reads, I discovered that virtually all the reads near the low

quality area began reporting double sequences (Figure 9) and thus quickly lost any semblance of useful data. In addition, I found that many of the reads contained repeated sequences offset by a few bases, explaining the difficulty in unambiguously sequencing the area.

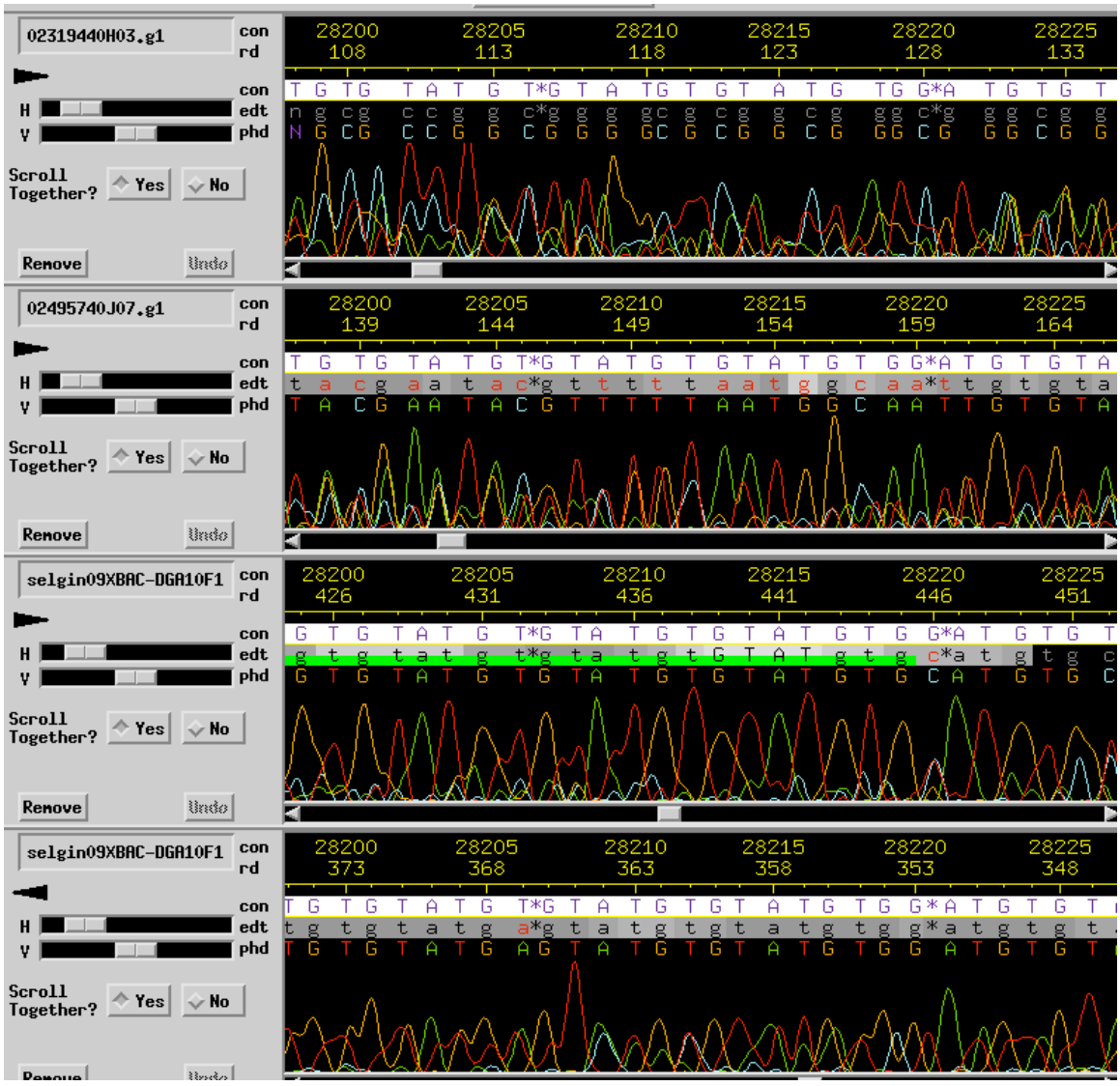

Figure 9: Double sequences near low quality consensus.

With this information, it seemed that there was sufficient evidence to justify a tear at the low quality region (Figure 10). The digests and lack of high quality sequence in that area suggested that there were 1.2 kb of sequence missing in the space between the tear, thus giving me some direction for the next set of sequencing reactions.

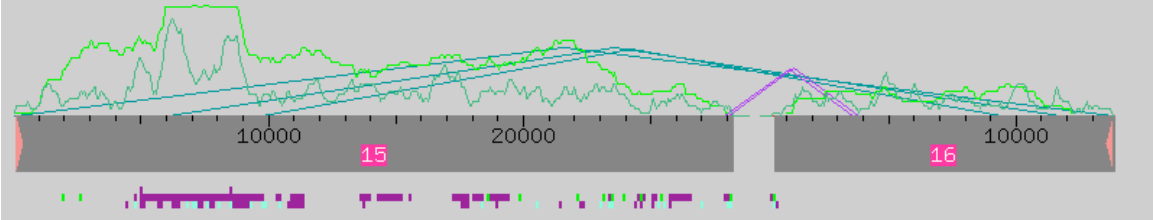

Figure 10: Assembly after tearing at the low quality region.

Third Round of Sequencing

Problem addressed	Direction of sequencing	Oligo sequence	Oligo location	Chemistry	Oligo name	Special
15-16 gap	→	Cgagagagcgaacgagc	15:27742-27758	All	DGA10F16.16	
	←	Ggctggttacatccgaga	16:1075-1092	All	DGA10F16.17	Relaxed parameters
PCR across 15-16 gap	→	Cctgcgctgcattattta	33133-33149	All	DGA10F16.14	Sequenced with PCR primers
PCR across 15-16 gap	←	Ttcaaataatttatgaccataaataaag	33620-33639	All	DGA10F16.15	Sequenced with PCR primers

Table 4: Oligos called for third round of sequencing.

Once again, different primers were chosen to try to sequence across the gap between contigs 15 and 16. In addition, the region was amplified by PCR and sequencing reactions were run on the product with the same primers used for PCR. The rationale behind calling PCR was twofold. First, a PCR product, especially if it used primers far enough away from the ambiguous sequence, might yield better sequencing data than direct fosmid sequencing. In addition, the size of the PCR product would give an indication of the size of the missing region, and confirm or refute the hypothesis that this region is what was causing the discrepancy in the digests.

Results of Third Round of Sequencing

None of the sequencing reactions worked effectively. Oligos 16 and 17 yielded virtually no high quality sequence, and the PCR sequencing reactions, though yielding some good sequence, broke down upon reaching the problem areas. However, the PCR product was revealed to be ~ 2kb in size, thus confirming my hypothesis that the low quality area in fact represents sequences over 1 kb apart. As no further reactions could be called, the gap between the two contigs was left in the assembly.

Final Work

Analyzing the final assembly consisted mostly of addressing high quality discrepancies, as many were reported by Consed. Upon analysis, the vast majority of these high quality discrepancies are in the region of 4687-9889, which is the highly polymorphic region originally reported as contig 14 in the first assembly. Several of them are discrepancies as described initially, with two populations of reads present. All of these are consistent between reads. When a single read spans two base positions of discrepancies, its bases are either both discrepant or both consistent, as would occur if a sequencing reaction

were done from one copy or another of the dot chromosome. Many of the other discrepancies only appear in one read, and thus seem most likely to be growth differences, single clones that mutated one base. However, the great number of these which occur, as well as the polymorphic nature of the area, makes it possible that some of these are actually polymorphisms only reported in one read, or reads mistakenly included in this contig which belong elsewhere in the dot chromosome. A final category of high quality discrepancies was untrustworthy bases, either near the beginning of a read or in a low quality area of the read.

Other issues were also addressed. Some smaller contigs (1-3 kb), which had not been incorporated previously because of discrepant bases also had to be force-joined, with the discrepant bases properly labeled as polymorphisms. In addition, a few unaligned high quality sequences which match no other area in the assembly were found at the ends of reads; these seem likely to be vector sequences, as they all occur at one end or another of a read (before or after some good sequence), rather than between matching sequences. No X's or N's were found in the consensus sequence. One mononucleotide run was found, but its length is confirmed by several reads, so I am confident in it. Finally, all single strand/single chemistry areas were analyzed to confirm that they had sufficiently high phred scores to be trusted. The final assembly view is depicted in Figure 11.

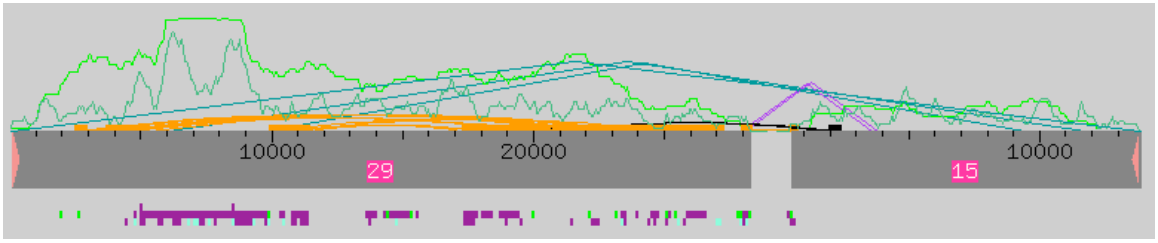

Figure 11: Final Assembly View.

Blast Analysis

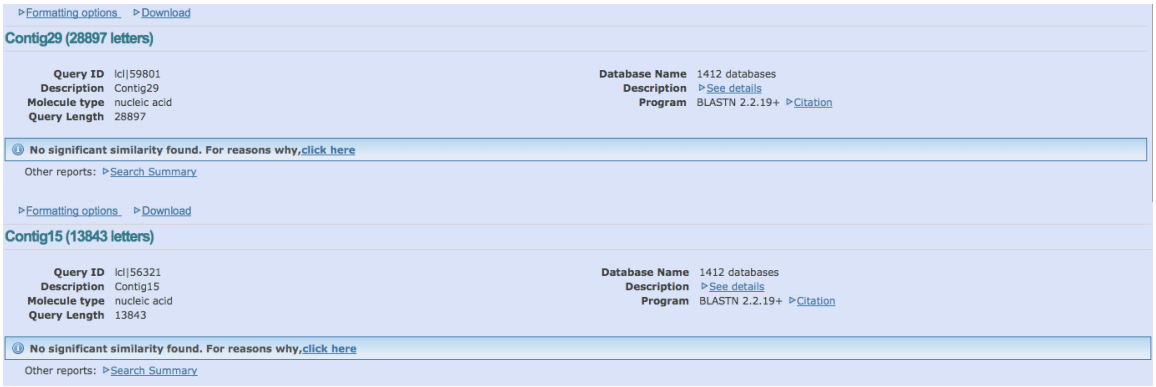

Figure 12: BLAST results for final assembly.

BLAST was used to compare my final sequence to known microbial DNA, ensuring that no microbial DNA was incorporated into the assembly. As seen in Figure 12, no significant similarity was found to any microbial sequences in the NCBI database, so I am fairly confident that my fosmid's sequence is completely *D. grimshawi* DNA.

Digest Analysis

In silico digests were run with a force join between the two contigs, which placed the PCR primers 2 kb, their known separation, apart. As seen in Figure 13, both EcoRI and HindIII yield almost perfect matches between all of the bands. The gap region no longer causes band discrepancies, thus suggesting that the PCR product is indeed representative of the distance between the two contigs. The smaller bands on the HindIII *in silico* digest which have no counterparts in the real digest seem to correspond with bands visible in the real digest image that were not recognized as such by the program.

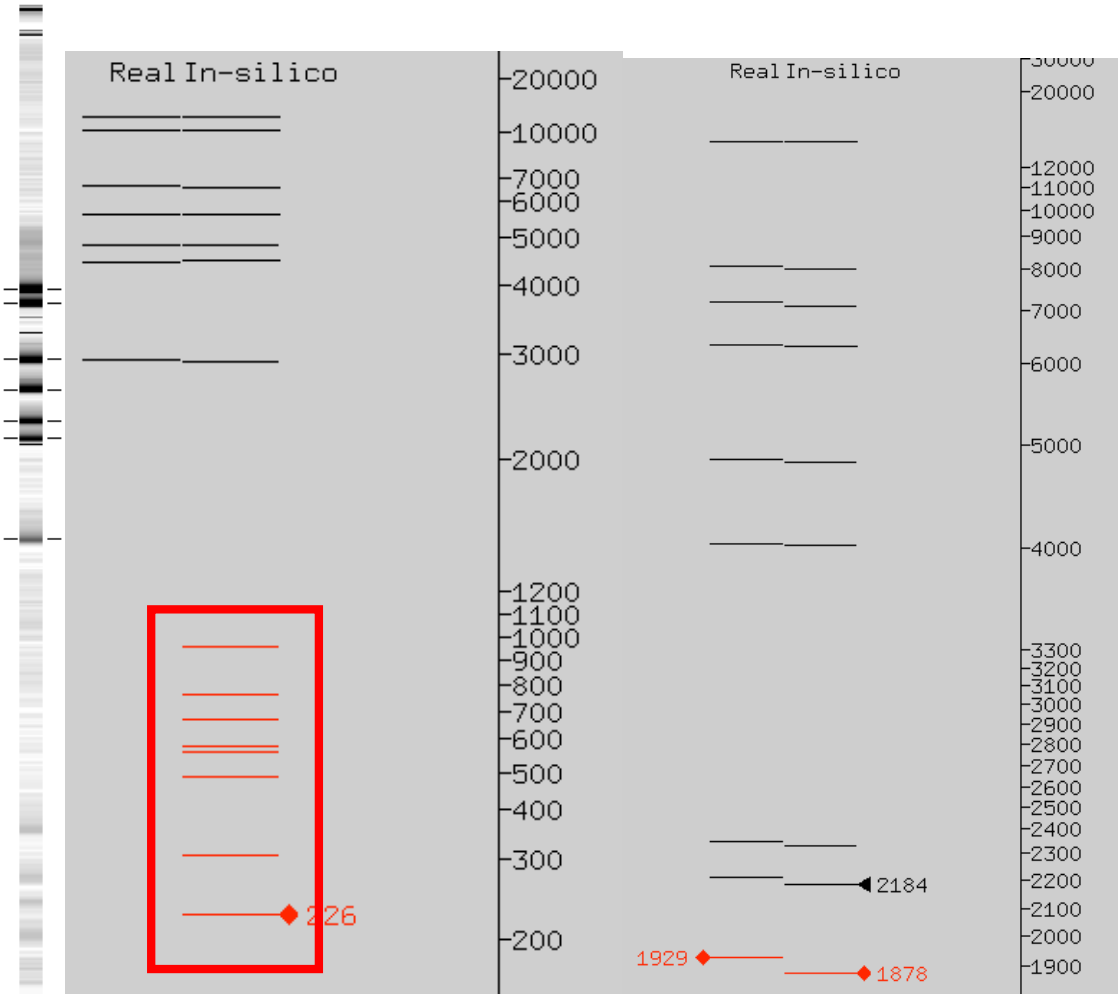

Figure 13: Real HindIII digest image (not to scale), and *in silico* digests with HindIII and EcoRI.

Future Directions

This project is not yet finished, since it still has a gap between contigs which has not yet been sequenced across. The repetitive nature of the sequences around this gap will make sequencing very difficult, but future finishers may want to clone the gap region into a vector, allowing sequencing from standard M13 (or equivalent) sequencing primers. In addition, not all polymorphisms in the contig have been unambiguously identified as belonging to the fosmid or not, so future finishers may want to sequence across those

regions to identify which chromosome copy matches the fosmid sequence. All other criteria for success in this project, such as phred 30 or above sequences, have been met, and the four initial contigs have been reduced to only two.

Special Thanks

Special thanks to all of the finishers, who were indispensable for navigating the puzzle of this fosmid, but especially to Taylor Cordonnier, whose quick work PCRing my gap region helped me narrow down its size and created the beautiful digest images seen above.