Finishing *Drosophila mojavensis* Clone 430-O17

Priya Srikanth
Bio 434W, Revision 2
4.20.2007

Abstract

The small fourth chromosome of *Drosophila melanogaster,* also known as the dot chromosome, is unlike the other three chromosomes of *Drosophila* in that it contains largely heterochromatic sequence. The 1.2 Mb arm of the dot chromosome has normal gene density, but a higher frequency of repetitive sequences compared to the euchromatic arms of chromosomes 1-3. The many genes packaged in the dot chromosome's heterochromatic regions are expressed more strongly than genes in the heterochromatic regions of other chromosomes. This unusual chromosome provides an opportunity to study differences in euchromatin and heterochromatin coding and formation. To investigate the effects of sequence organization on formation of heterochromatic and euchromatic domains, Bio 4342 has used comparative genomics, looking at dot chromosome sequences from *D. melanogaster* and *D. virilis*. 2007 marks the start of analysis of a new dot chromosome – that of *D. mojavensis*. My project focuses on finishing clone 430-O17 from this chromosome.

Initial Analysis

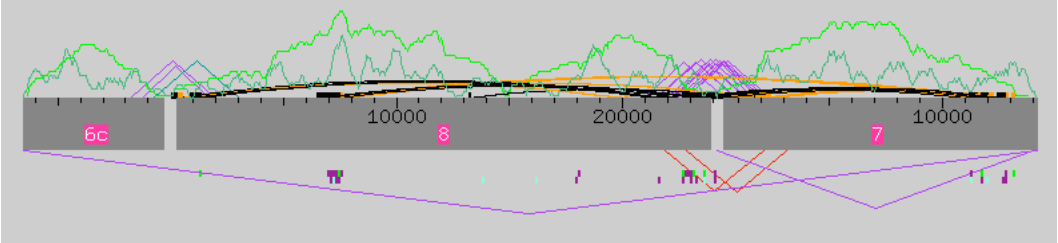


Figure 1. Initial Assembly View of 430-O17 with crossmatch.

Figure 1 shows the original Assembly View of clone 430-O17 with crossmatch showing tandem repeats in orange and complemented repeats in black.[1] The first task was to identify the ends of the fosmid. This was done by searching for the two forward and reverse sequencing reactions performed on the intact fosmid. Ends of these reactions located at the end of a contig were tagged as fosmid ends. One fosmid end was at the end of Contig 6c and the other was at the end of Contig 7. I uncomplemented Contig 6c to orient the fosmid end at the beginning of the contig.

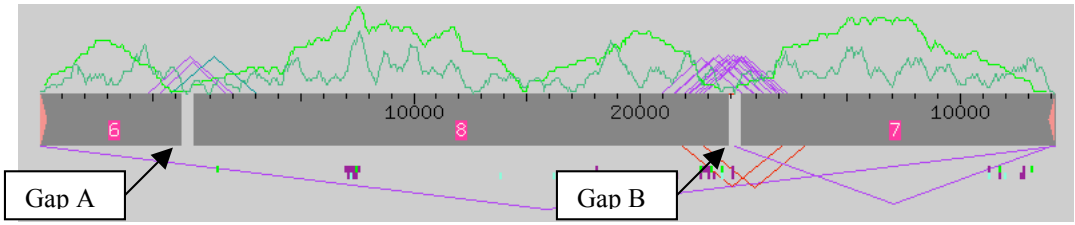


Figure 2. Assembly view showing Contig 6 and tagged end reads (pink triangles shown).

[1]Complemented repeats are sequences in the consensus sequence that match elsewhere to the complementary strand of DNA.

Two problems to resolve in the initial assembly were closing the gap between Contigs 6-8 (Gap A) and the gap between Contigs 8-7 (Gap B). Since both gaps were spanned by paired-end reads (shown in purple in Figures 1 and 2), I was confident that the three contigs were in the correct order. However, Gap B was flanked by repeats on both sides (Figure 3), and Gap A was flanked by a repeat in Contig 8 (Figure 4). This could pose a problem when ordering oligos to sequence across the gaps, as oligos in repeat sequences may not anneal at the desired location, but could instead anneal at the repeated sequence elsewhere in the fosmid. Before concluding that oligos would be needed to cover the gaps, I attempted to find similar sequences at the ends of the contigs on either side of the gaps; these would facilitate forced joins. No similar sequence regions were found that allowed alignment of the sequences flanking either Gap A or Gap B. Since no join could be forced, oligos would be needed to generate sequence data to close these gaps.

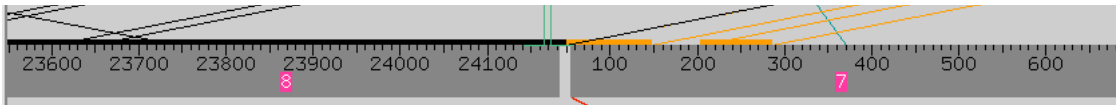


Figure 3. Gap B, with repeats at the end of Contig 8 and beginning of Contig 7 in black and orange.

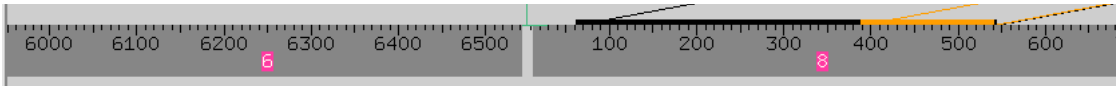


Figure 4. Gap A, with repeat at the beginning of Contig 8 in black.

There were also numerous high quality discrepancies in my fosmid. All but four of these discrepancies were located in Contig 8 (Figure 5). When examining the traces of discrepant reads, I found several high quality discrepancies that were not due to incorrect base calls. Rather, these were discrepancies between two very high quality traces that I could not resolve (A few examples can be seen in Figure 6; for more examples see Figures 6.1 and 6.2 in supplement.) These high quality discrepancies were clustered together in Contig 8 around bases 16662-1680 and 18105-18693.

Contig Name	Read Name	Consensus Positions	
Contig8	09309675A21.g1	10556	high quality base disagrees with consensus
Contig8	03698175F12.b1	10847	high quality base disagrees with consensus
Contig8	39135000L07.g1	16662	high quality base disagrees with consensus
Contig8	39135000L07.g1	16673-16677	high quality base disagrees with consensus
Contig8	39135000L07.g1	16678-16679	high quality base disagrees with consensus
Contig8	39135000L07.g1	16717	high quality base disagrees with consensus
Contig8	39135000L07.g1	16732-16741	high quality base disagrees with consensus
Contig8	39135000L07.g1	16748	high quality base disagrees with consensus
Contig8	39135000L07.g1	16759	high quality base disagrees with consensus
Contig8	39135000L07.g1	16780	high quality base disagrees with consensus
Contig8	03705375A08.b1	18086	high quality base disagrees with consensus
Contig8	09430375P18.g1	18105	high quality base disagrees with consensus
Contig8	04115275H23.g1	18105	high quality base disagrees with consensus
Contig8	07797975M05.g1	18491	high quality base disagrees with consensus
Contig8	07797975M05.g1	18521	high quality base disagrees with consensus
Contig8	07797975M05.g1	18693	high quality base disagrees with consensus
Contig8	03802575D08.b1	22054	high quality base disagrees with consensus
Contig8	07693575L23.g1	22168	high quality base disagrees with consensus

Clustered high quality discrepancies

Clustered high quality discrepancies

Figure 5. List of high quality discrepancies in Contig 8.

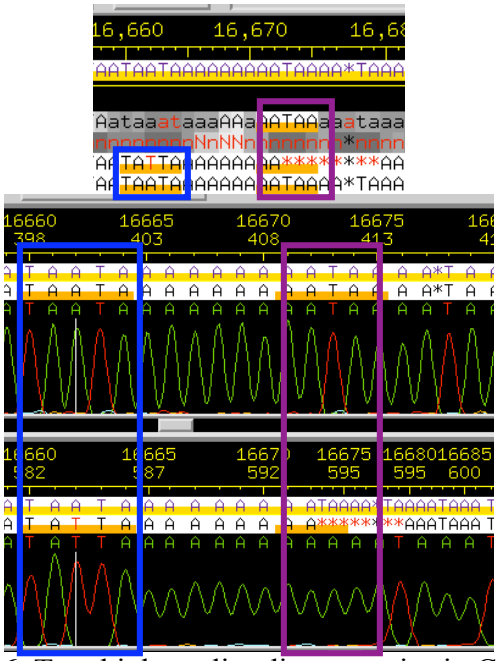


Figure 6. Two high quality discrepancies in Contig 8.

The presence of clustered high quality discrepancies seemed to indicate a misassembly. This problem would have been solved most efficiently by examining real vs. in-silico restriction enzyme digest data.[2] However, because this data was not yet available, I approached the problem from another angle. The most likely explanation for these discrepancies was that Consed had collapsed two repeats into a single region, and the high quality discrepancies were actually the differences between repeat sequences. To resolve the misassembly, I tagged seven of the discrepancies (at positions 16662, 16673, 16732, 16748, 16759, 16780, 18105) as "tell phrap not to overlap reads discrepant at this location" and ran phredPhrap. This created a new assembly with discrepant sequences in separate locations (Figure 7).

[2] "Real" refers to computer-identified bands from an image of a gel. "In-silico" refers to the band sizes that would result from a digest of the current assembly, based on location of restriction enzyme recognition sequences in the fosmid.

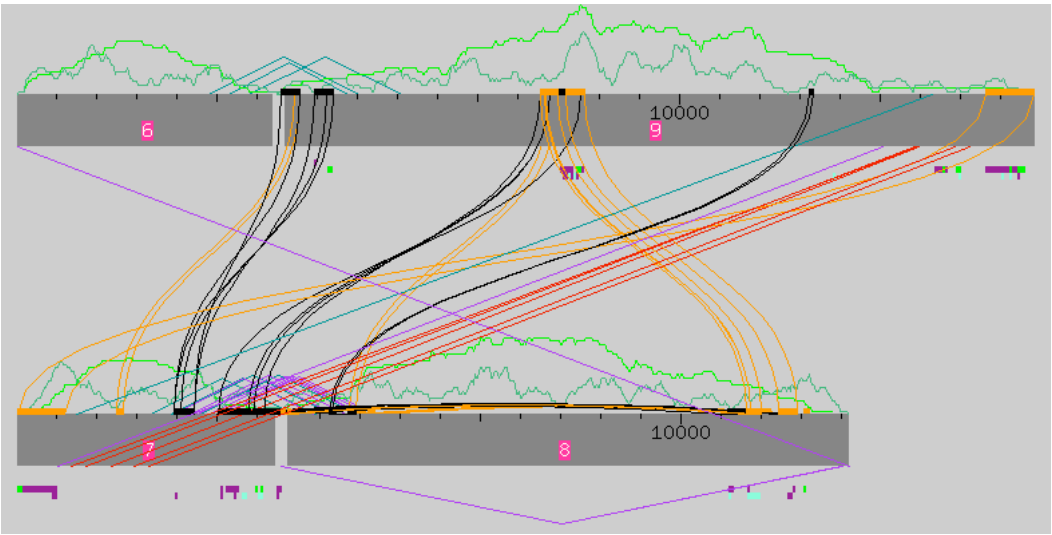


Figure 7. Assembly View after re-running phredPhrap with high quality discrepancies tagged.

The discrepant reads were split into two separate clusters of repeat sequences at the end of Contig 9 and the beginning of Contig 7, as shown by the orange tandem repeat in Figure 7 (1168 bp, 99.7% similarity). I then looked for high quality discrepancies again in Contigs 7 and 9. I found one read, 07692675B20.b1 that was aligned with the repeat in Contig 7, but contained a high quality discrepancy that matched Contig 9's repeat sequence (Figure 8). I took this read out of Contig 7 and added it to Contig 9 using the 'Search for String' and 'Compare Contig' functions to make the join, creating Contig 14 (Figure 9).

Figure 8. Read 07692675B20.b1 (highlighted in pink) disagrees with consensus sequence at bp 1096 of Contig 7 (top) and matches Contig 9's sequence (bottom) at bp 18725.

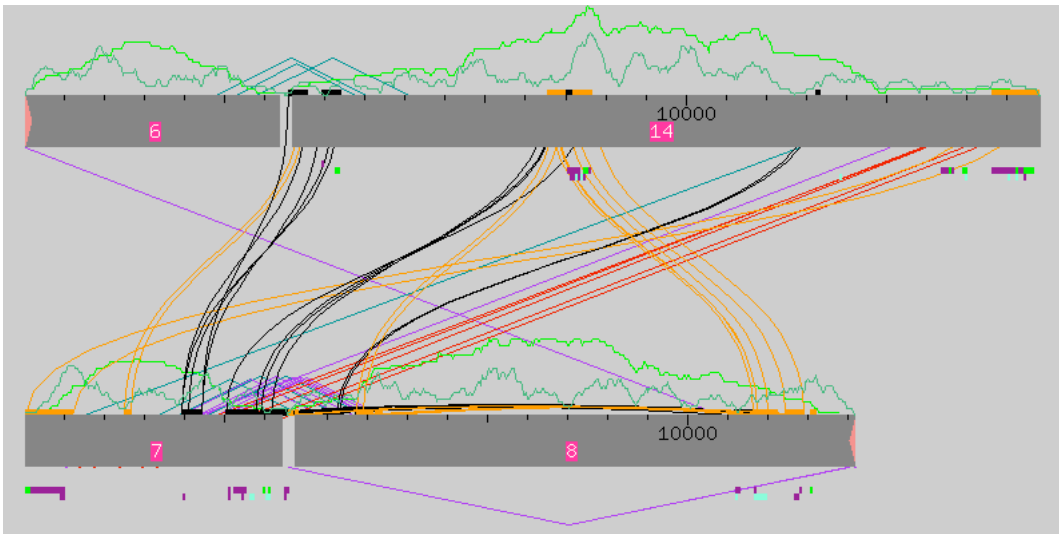


Figure 9. Assembly View with Contig 14, created from joining Contig 9 with read 07692675B20.b1.

I then continued looking for high quality discrepancies, and found that the read I tore from Contig 7, 07692675B20.b1, contained discrepancies from Contig 14's consensus sequence. At two single nucleotide discrepancies between Contig 14 and Contig 7, this read aligned with Contig 14 at one discrepancy, and Contig 7 at another discrepancy (Figure 10)! This read therefore did not align perfectly with the consensus sequence of either contig.

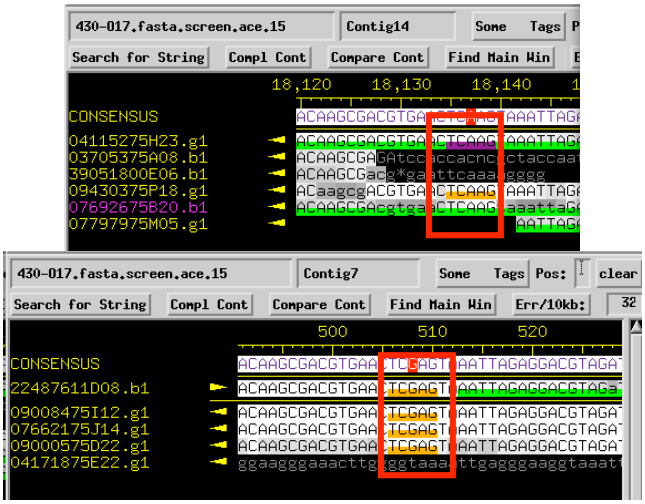


Figure 10.1 Read aligns with Contig 14 (top, at bp 18137) rather than Contig 7 (bottom, at bp 507).

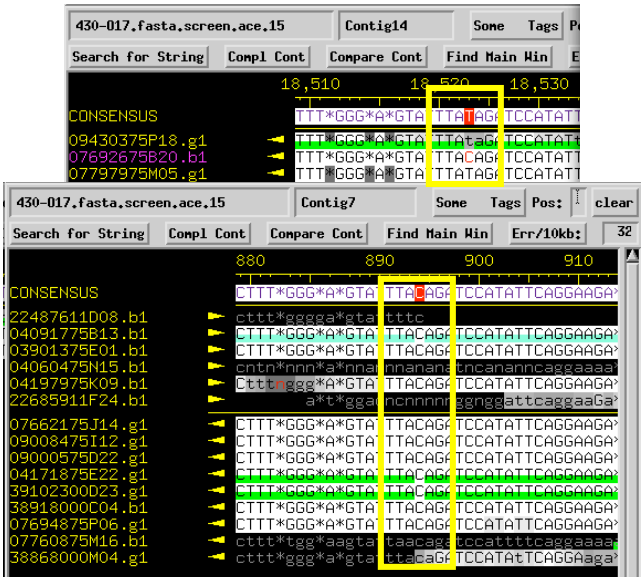


Figure 10.2 Read aligns with Contig 7 (bottom, at bp 894) rather than Contig 14 (top, at bp 18523).

Figure 10. Read 07692675B20.b1, highlighted in pink, aligns with Contig 14 at one high quality discrepancy (Figure 10.1) and aligns with Contig 7 at another (Figure 10.2).

Read 09430375P18.g1 posed a similar problem – it aligned with Contig 14 at some of the discrepancies that distinguished Contigs 14 and 7, but aligned at Contig 7 at another discrepancy. These two reads caused me to suspect that the high quality discrepancies I saw could be polymorphisms. If they were indeed separate copies of tandem repeats, there would have to be four copies of the repeat sequence (or more, if I found more reads like the two mentioned above) to eliminate high quality discrepancies within the sequence. While this could be the case, there tended to be only one or two reads that were discrepant with the consensus sequence. The hypothesis of a collapsed repeat would be better supported by the appearance of a discrepancy in a large proportion of reads. The possibility of several repeats was beginning to seem unlikely; however, I was not confident that there should be polymorphisms in the sequence because inbreeding of the source fly population usually eliminates polymorphisms. I decided to wait until I received digest information before changing my assembly.

Round One Analysis

In picking first round sequencing primers, I prioritized ordering oligos for what I perceived to be the hardest problems to resolve. This included spanning Gaps A and B and determining whether the 1.2 kb repeat in the assembly was a genuine repeat or one sequence containing polymorphisms. To order primers, I first used Consed's automated "pick primer" to present candidate oligos. I looked for oligos that were more than 70 bp and less than 200 bp away from the region that needed sequencing. Leaving a 70 bp gap between oligo and desired sequencing region makes allowance for messy sequencing chemistry and poor sequence quality at the beginning of a read, which often persists up to 50 bp into a reaction.[3] If an oligo met these requirements, I then performed an approximate Search for String looking for up to 15% sequence

[3] Depending on the problem to be solved, if there were no other primers to be found, oligos >200 bp away from the desired region were accepted. For single strand/ chemistry regions, for example, a larger distance was permitted. However, for spanning gaps, since I wanted as much novel sequence as possible, oligos >200 bp from the gap were not accepted.

similarity to see if the primer might anneal at another location in the clone.[4] Oligos with higher melting temperature (i.e. 58°C instead of 55°C) and oligos around 22 bp long were picked if there were many primers to choose from.

The mostly automated method described above succeeded in picking the 5' primer for Gap A. Picking a 3' primer for this gap proved to be difficult because the beginning of Contig 14 on the 3' end of the gap contained a repeat sequence. Consed was unable to find any primers that met the appropriate criteria. Since the repeat sequence did not begin until bp 94 of Contig 14, I hand-picked a primer from the unique sequence. To hand-pick primers I used the criteria of unique primer sequence and appropriate primer length (~22 bp).

Picking primers for Gap B was also difficult, since both sides of the gap were flanked with repetitious sequence. The end of Contig 7 (the beginning of the gap) contained a repeat sequence that extended until the beginning of the gap. The alignment of this repeated sequence was perfect until the end of Contig 7 (Figure 11) – there was no unique sequence until about 500 bp before the gap. Therefore, no 5' primers could be picked by either the automated or manual methods that did not match sequence elsewhere in the clone. At the end of the gap, the beginning of Contig 8 contained another repeated sequence. This repeat had a lower percentage similarity than the repeat at the end of Contig 7 (94.0% instead of 98.7%), and was shorter (99 bp instead of 549 bp). The lower sequence identity and length of the repeat made hand-picking a primer more plausible. The alignment of this sequence from Contig 8 with its repeat (also in Contig 8) revealed a cluster of discrepancies around position 80-95 (Figure 12). I picked a primer (oligo 3 in Table 1) containing four of these discrepancies to sequence across Gap B.

Figure 11. Alignment of repeat flanking the beginning of Gap B at the end of Contig 7 (top sequence) to Contig 8 (bottom sequence). Note the low level of discrepant bases between the reads.

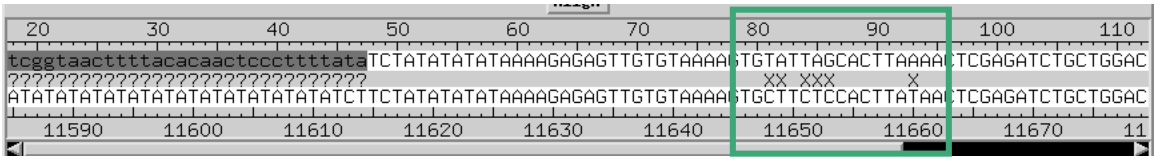


Figure 12. Alignment of repeat flanking the end of Gap B at the beginning of Contig 8 (top sequence) with the middle of Contig 8 (bottom sequence). Positions 80-95 contain a cluster of discrepancies between the repeats.

Since I did not yet have restriction enzyme digest data, I decided to try using additional sequence data to determine whether the repeat in Contigs 14 and 7 was a genuine repeat or a polymorphism. Since these sequences were very similar (99.7% identity), and the repeat was

[4] Again, depending on the problem, different percentages of sequence similarity of the oligo to other locations in the fosmid were tolerated. For example, if there were few oligos found at a suitable distance, and all had some level of similarity elsewhere in the fosmid, I tried to minimize the chance that the oligo would anneal elsewhere by minimizing similarity to other locations. When an oligo matched another location, I also tried to choose an oligo whose last base in the direction of sequencing did not match the sequence at the other location.

very long (1168 bp), I designed primers in the unique sequence at the ends of the repeats pointing toward the putative gap between Contigs 14 and 7. If data from these sequencing reactions matched the current consensus sequence, showing no polymorphisms, this could provide further evidence that the sequence was not a repeat, but contained polymorphisms from WGS sequencing. There would still be a chance that the underrepresented sequences were copies of repeats that could not be sequenced by the reactions called, since a reaction would have to extend further than 1.2 kb to hit another copy of repeated sequence. This is very unlikely, as sequencing reactions rarely extend past 600-700 bp.

Table 1. Oligos I called for the first round of reactions. Direction is indicated by → (5') or ← (3'). "Added" or "Not Added" indicates whether the resulting read was incorporated into the Consed assembly of 430-O17.

Oligo	Sequence	Dir.	Problem	Result
1	gaagctctcagcgtcaataata	→	Gap A	Added
2	gtgctttggaatgaacatacatt	←	Gap A hand picked	Added
3	agatctcgagttttaagtgctaa	←	Gap B hand picked	Added
4	cttttggcttcattaattatgg	→	1.2 kb repeat	Added
5	ttggttcaccaaaactctttcgaac	←	1.2 kb repeat hand picked	Added
6	gtaccgttcctctaattgttgat	←	repeat 'gap' hand picked	Added
7	ctcctatggaatatcatgtacga	→	single strand/ chem	Added
8	ttttcgataactcctccgata	→	single subclone	Not Added
9	tgggacagtaatttacgattacc	←	single strand/ chem	Added
10	acggatttcctgatgatttat	→	single strand/ chem	Not Added
11	tgctgctaatgttgtgttagtg	←	single strand/ chem	Added

Table 2. Oligos called by Autofinish using the original assembly.

Oligo	Sequence	Dir.	Problem
1	cagcgataggatgataataactaat	←	end of contig 6c, Gap A
2	ccagataataccataaatggaagaa	→	end of contig 6c, fosmid end
3	tgatccgcctcgtgg	←	end of contig 8, Gap B
4	tgaatacccgctaaaagcta	→	end of contig 8, fosmid end
5	ccccacccctatatccac	←	end of contig 14, Gap A
6	ccaacactaaacatgttgaaaata	→	end of contig 7, Gap B (2 matches)
7	gcgcagtcgctgctta	←	single strand/ chem
8	cgcgtcgatctcaagata	←	single strand/ chem
9	agggcgaaccccaat	→	low base quality
10	catggggcacagttagaa	→	single strand/ chem
11	gcccacgtctcacagc	→	single strand/ chem
12	ccaagctttcaccaacaa	→	? (2 matches)

One contig in original assembly

Comparison to Autofinish

Although the assembly I used for calling the first round of reactions was different than the original assembly used by Autofinish, the issues addressed by both myself and Autofinish were largely similar. Autofinish called reactions off the ends of each Contig, covering Gaps A

and B, as well as the ends of the fosmid. While running reactions off the ends of the fosmid would usually be unnecessary, I later discovered (before the third round of reactions) that in my case, I did need extra data at the ends of my fosmid because I could not identify any vector sequence. Thus, what would usually be a drawback to using Autofinish (calling extra reads when unnecessary) would have been an advantage for finishing my fosmid. Both Autofinish and I called a few reactions to cover single subclone areas and single strand/ chemistry regions.

Autofinish did not call any reads around the 1.2 kb repeat, as there was no repeat present in the original assembly. The program did not take any action to resolve the clustered high quality discrepancies found in the original assembly. Two of the oligos recommended by Autofinish were unacceptable because they matched exactly to two different locations in the original assembly. One of these oligos (#12 in Table 2) did not have any obvious function – I could not identify any problem regions around the oligo in either of the two locations it matched. This demonstrated the limitations of Autofinish – it can resolve issues such as gaps, single subclone regions, single strand/ chemistry regions, and low quality regions, but cannot address high quality discrepancies or the possibility of misassembly. The criteria for primer selection are also questionable, as two oligos picked by Autofinish matched two different locations in the clone.

Results of Round One Reactions

Luckily, all of my hand-picked oligos worked, despite limited analysis of primer suitability. Overall, my first round of reactions was very successful, with 9 of 11 resulting sequences added to my assembly. These reads spanned Gaps A and B successfully, resulting in the assembly in Figure 13. The reads sequencing the 1.2 kb repeat all agreed with the consensus sequence, and not with any of the discrepant reads. This supported the possibility of polymorphisms, but did not exclude the possibility of a tandem repeat, as discussed earlier.

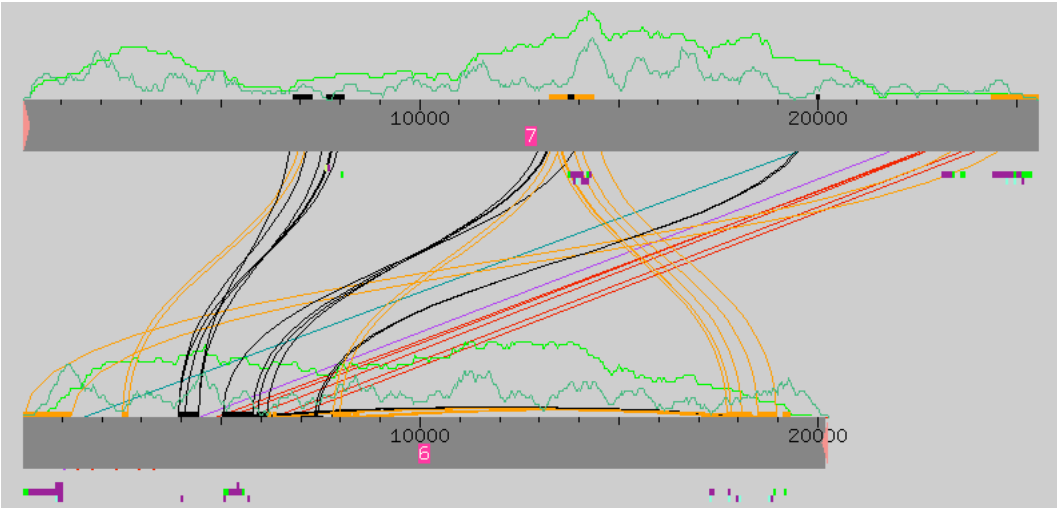


Figure 13. Assembly View after incorporation of round one reads.

Digest Analysis

Upon receiving the restriction enzyme digest data, I began analysis of the real vs. in-silico digests of my fosmid to determine whether the sequence at the end of Contig 7 and the beginning of Contig 6 (shown above in orange) was a repeat or a sequence containing polymorphisms. Four restriction digests were performed on our clones, using SacI, EcoRI,

HindIII, and EcoRV. Unfortunately, the SacI and EcoRI digests of my clone were unsuccessful, so I could use only the HindIII and EcoRV digests for my analysis.

To resolve the question of repeated sequence versus polymorphisms, I looked at these in-silico restriction digests using the assembly shown above, with two copies of the repeat sequence in separate contigs, compared to in-silico digest data using an assembly with only one copy of this sequence in one contig.[5] The HindIII digests (Figure 14) show that when Contigs 6 and 7 are joined, 3 of the mismatched bands shown in red are eliminated from the in-silico digest (the gap-spanning band of ~1200 bp, and the two bands of ~8000 and 9000 bp)[6]. Two doublets remain in the in-silico digest that are absent from the real digest (red bands in Figure 14.2). To check the accuracy of the computer-called bands, I looked at the gel image (Figure 14.3). The bands corresponding to the in-silico doublets appear thicker than the surrounding bands in the digest, making it probable that these bands are doublets that were mistakenly called as single bands by the program. If so, the in-silico HindIII digest would perfectly match the real digest.[7]

[5] To make the latter assembly, I used Compare Contig to align the two sequences and joined Contigs 7 and 6, overlapping the high quality discrepancies.

[6] Though Figure 14.2 does not show the digest to 10000 bp, there are no bands between 8000 bp and 10000 bp in the single contig digest.

[7] Note: digest data is only relevant between 1kb and 10kb, because of low accuracy of band size-calling outside of this range.

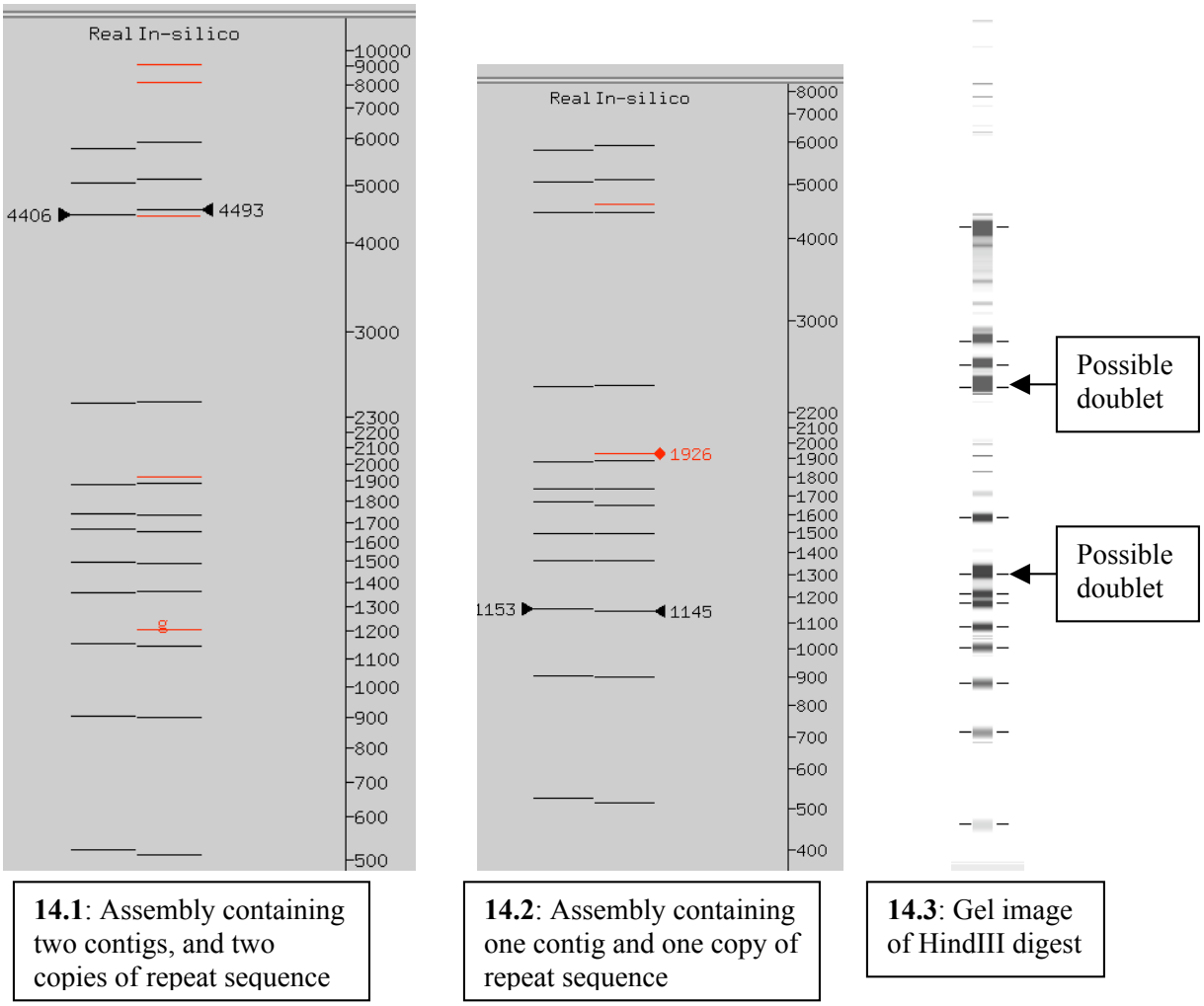


Figure 14. HindIII digests of the fosmid. Red bands in 14.1 and 14.2 indicate band sizes that do not match closely enough between real and in-silico digests to be called the same band. Bands of similar sizes that are present in both real and in-silico digests are shown in black.

The EcoRV digests reveal a similar result from joining the contigs. The digest of the two-contig assembly shows many mismatched bands (Figure 15.1) that are resolved by joining the contigs and overlapping the repeat. There is one mismatched band that remains – a doublet in the in-silico digest that is only a single band in the real digest. I looked at the gel image of the EcoRV digest to see if a single band was called where there could be a doublet. Again, the corresponding band was thicker than the surrounding bands, indicating the possibility of a doublet. Thus, the real and in-silico digests of the single contig assembly are consistent with each other, but the real and in-silico digests of the two-contig assembly are not.[8]

[8] Note the decreased number of bands in the EcoRV digest compared to the HindIII digest (14 bands for HindIII vs. 7 bands for EcoRV between 1kb-10kb). This is unusual considering that both enzymes have a 6bp recognition site (HindIII: AAGCTT, EcoRV: GATATC). This deviance from the expected frequency of cut sites is characteristic of clones containing repetitive sequence.

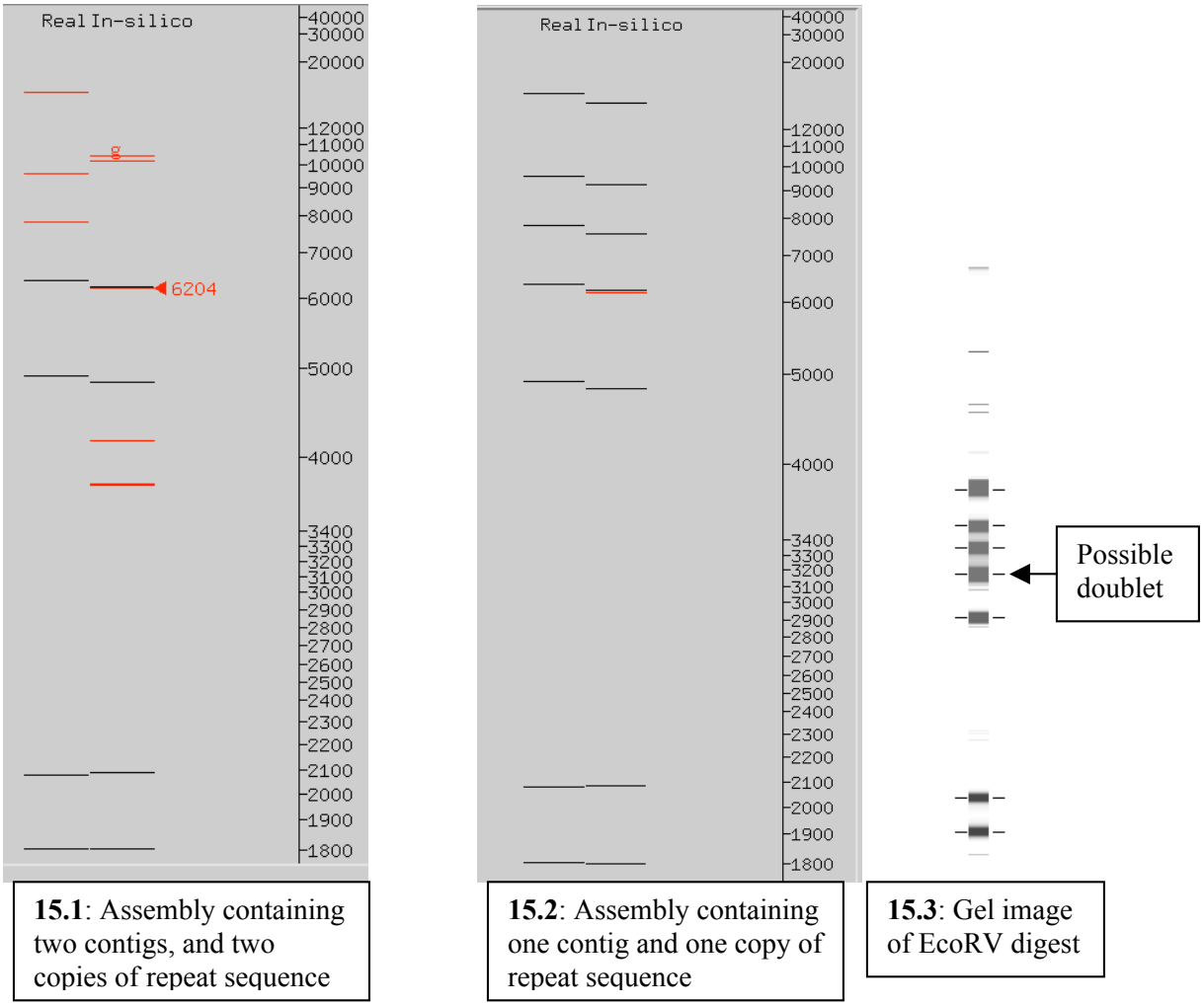


Figure 15. EcoRV digests of fosmid. There are no bands between 1kb and 1800 bp in the real or in-silico digests of either the two-contig or one-contig assemblies.

The digests provide good evidence that the repeat sequences should be collapsed into one sequence, as shown above. Collapsing these sequences could also resolve incorrect paired-end read distances (shown by red lines in Figure 13), as well as improve the low read coverage at the end of Contig 7 (shown by the dark green line above contigs in Figure 13). With these data and reasoning in mind, I aligned and joined Contigs 6 and 7, forming Contig 8. The restriction digest data indicates that the high quality discrepancies present in the assembly were not the result of a misassembly, but were most likely due to polymorphisms in the sequence. Although the fly breeding process would ideally eliminate variations in sequence, it does not guarantee that polymorphisms will not persist. Lack of recombination on the fourth chromosome increases the chance that polymorphisms will persevere even after extensive inbreeding. Assuming the high quality discrepancies to be polymorphisms, the discrepant reads were pulled out into a separate contig, Contig 14 (3 reads, 1561 bp). As predicted, joining the contigs eliminated the inconsistent paired-end read distances and improved read coverage at the join location (~25000 bp).

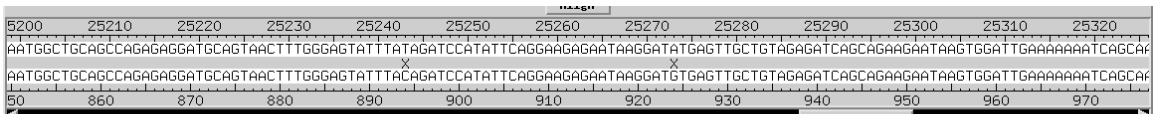


Figure 16. Alignment of end of Contig 7 (top sequence) with the beginning of Contig 6 (bottom sequence) showing two high quality discrepancies, representing polymorphisms

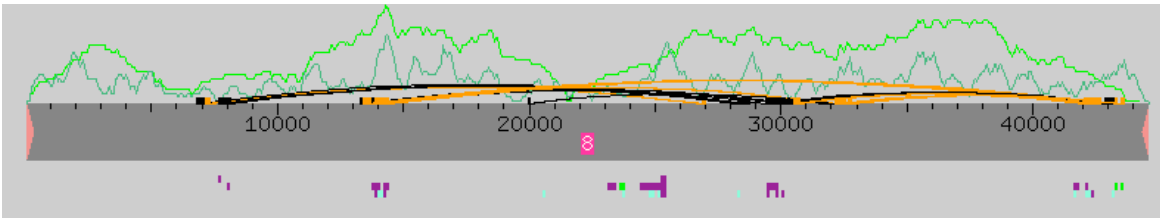
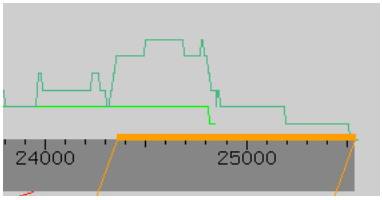
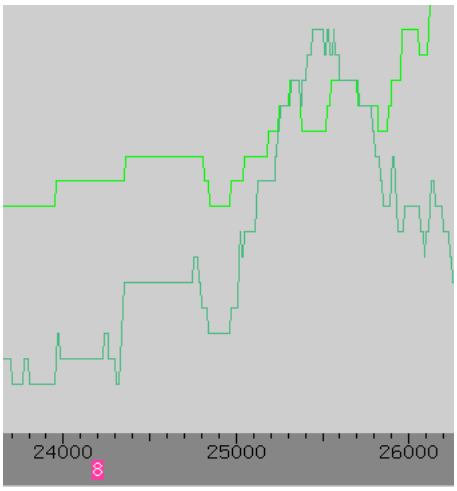


Figure 17. Assembly View after joining Contigs 6 and 7 into Contig 8.



18.1: Low read coverage at end of Contig 7 before collapsing repeat sequence



18.2: Improved read coverage in Contig 8 after collapsing repeat sequence

Figure 18. Improved read coverage after joining Contigs.

Round Two Analysis

In the midst of performing the analysis above, I had little time to prepare oligos for the second round of reactions. The oligos I did order were intended to cover single strand/ chemistry locations in the sequence (Table 3). One of these oligos (oligo 1) was used in the first round of reactions (then oligo 8), and failed. However, since there was no other suitable oligo for that region within 400 bp of the desired sequencing start, I decided to try the same oligo using 4:1 chemistry instead of Big Dye chemistry. Hopefully, the different chemistry would improve the sequencing reaction and result in a read that could be incorporated into the assembly.

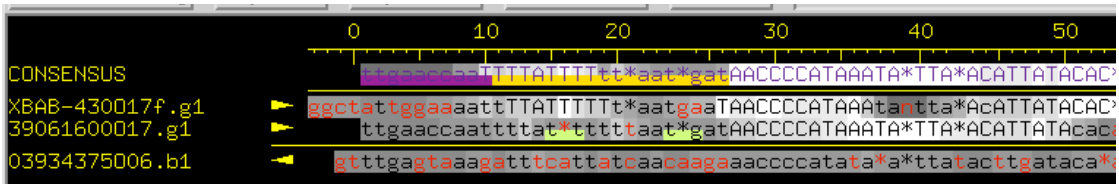
The round two reactions were not as successful as those from round one, with 3 of 6 reads incorporated into the assembly. The reaction I tried with the same oligo for the second time was successful, indicating that the change in chemistry improved the sequencing quality.

Table 3. Second round oligos.[9]

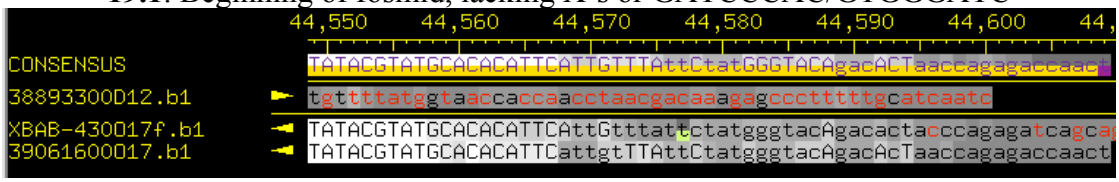
Oligo	Sequence	Dir.	Problem	Result
1	acggatttcctgatgatttat	→	(2nd try) single strand/ chem	Added
2	tgttctgaaagtttcattacga	←	single strand/ chem	Added
3	gcgacgagctaattcgatt	←	single strand/ chem	Not Added
4	cgcgtcgatctcaagata	←	single strand/ chem	Added
5	ccgttaatccgcagaatac	→	single strand/ chem	Not Added
6	ggcaaatcggagaacgta	←	single strand/ chem	Not Added

Round Three Analysis

The remaining problems in the assembly that could be fixed by sequencing were single subclone regions and single strand/ chemistry regions. Though the mouse standard we use allows such regions if all consensus sequence bases are above Phred quality 30, ideally, a finished fosmid would not have any single strand/ chemistry regions. For the last round of reactions, I ordered oligos to cover all of these regions that were still present in my clone. For regions larger than 400 bp, I ordered multiple oligos to span the region, about 300-500 bp apart. I also realized at this point that we should be able to identify vector sequence at either end of our fosmid. This would be indicated by a sequence of X's or by a GATCCCAC or GTGGGATC in the sequence at the fosmid end. As shown in Figure 19, my fosmid ends had neither X's nor the mentioned sequences, indicating that my fosmid lacked vector sequence. To verify clone ends by identifying vector sequence, I ordered oligos walking off of the beginning and end of my fosmid.



19.1: Beginning of fosmid, lacking X's or GATCCCAC/GTGGGATC



19.2: End of fosmid, lacking X's or GATCCCAC/GTGGGATC

Figure 19. Absence of identifiable vector sequence at either end of fosmid.

Table 4. Third round oligos. Oligo 5 includes Autofinish oligo 9 in Table 2, was hand-picked, and is aimed at the same region as oligo 8 in Table 1. Oligo 20 is identical to oligo 3 in Table 1, but is being used to resolve a single strand/ chemistry region rather than spanning a gap. "2nd try" indicates the second use of one oligo for the same purpose. This was only done when a suitable alternative oligo could not be found.

Oligo	Sequence	Dir.	Problem	Result
1	cccgaagatgaagtccaat	→	walk off end for vector seq.	Not Added

[9] Note: oligo 4 is the equivalent of Autofinish oligo 8 in Table 2.

2	ggatcattgcccagataata	←	walk off end for vector seq.	Not Added
3	gcgcactaaagctggctat	→	single subclone	Added
4	aaaagggagccgtcatt	→	single strand/ chem	Added
5	cccccgtgagggcgaaccccaat	→	single subclone hand-picked	Added
6	tcttcttacttgttttctttgga	←	single strand/ chem	Added
7	tctcctacacctttagagctagat	←	single strand/ chem	Added
8	gcgacgagctaattcgatt	←	(2nd try) single strand/ chem	Not Added
9	aggcatccgccattt	→	single strand/ chem	Not Added
10	gctatctcatagttgagcaagtct	→	single strand/ chem	Added
11	attacgctctaatagtttcattg	→	single subclone hand picked	Added
12	ccgtctgtccgtacacaat	→	single strand/ chem	Added
13	tgaaatgacaacaaatgttggt	→	single subclone	Added
14	tggtatgtatctccaagagcttta	→	single strand/ chem	Added
15	cgcaaaagctgatatctcc	→	single strand/ chem	Added
16	ctgcaacgttcttacttattttact	→	single strand/ chem	Added
17	cgatatgccagaaatcatataat	←	(2nd try) single strand/ chem	Added
18	tatttcctgaactttaagacctaga	←	single strand/ chem	Added
19	gggatgtgaacgactagca	←	single strand/ chem	Not Added
20	agatctcgagttttaagtgctaat	←	single strand/ chem	Added
21	ccgttaatccgcagaatac	→	(2nd try) single strand/ chem	Not Added

The third round of reactions had a good success rate (15 of 21 reads incorporated into the assembly). Notably, reactions 8 and 21, both of which were second tries using the same oligo, failed again. This indicates that these regions may be especially difficult to sequence and/or the oligos are not optimal for sequencing. If these regions still need to be covered, at the discretion of the next finisher, perhaps oligos will have to be hand-selected. Also, both sequencing reactions off of the ends of the fosmid failed (using oligos 1 and 2). Since these reactions are necessary for identification of vector sequence, they should be repeated using different oligos if possible.

Final Analysis

Figure 20 shows the Navigator of remaining issues that need to be resolved in my fosmid. The first four and last five listed problems are at the beginning and end of the clone, respectively. These problems will be solved when the ends of the fosmid are sequenced to identify vector sequence.

Low Cons Qual/High Quality Discrep (no comp/G_dropouts)/Single Stranded and Chem/Single Subclone/Unaligned H

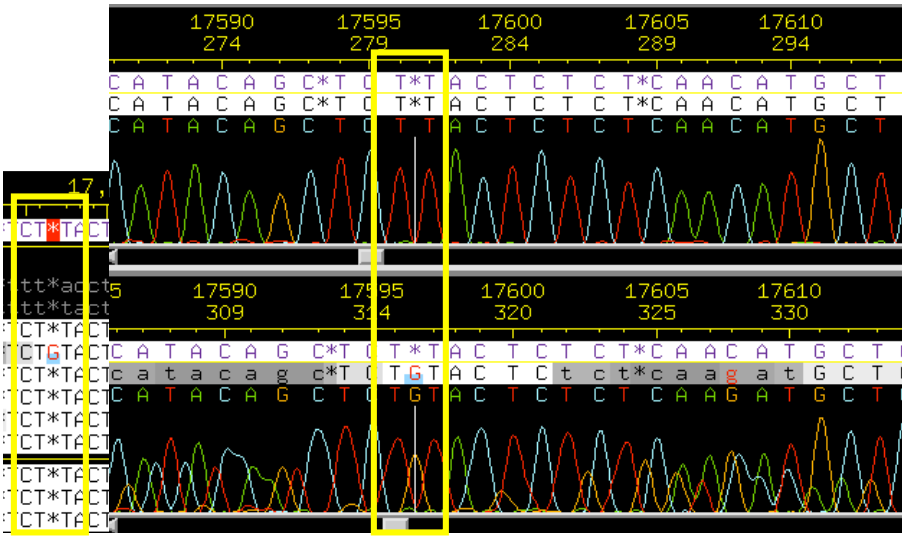
Contig Name	Read Name	Consensus Positions	
Contig8	(consensus)	1-26	28 bp single strand/chem
Contig8	(consensus)	1-7	7 bp single subclone
Contig8	(consensus)	1-9	base quality below threshold
Contig8	(consensus)	19-26	base quality below threshold
Contig8	(consensus)	8967-9111	150 bp single strand/chem
Contig8	(consensus)	9091-9092	base quality below threshold
Contig8	03698175F12.b1	17596	high quality base disagrees with consensus
Contig8	(consensus)	30299-30445	151 bp single strand/chem
Contig8	03923075C08.g1	38418	high quality base disagrees with consensus
Contig8	07684975K08.b1	42781	high quality base disagrees with consensus
Contig8	(consensus)	44499-44609	111 bp single strand/chem
Contig8	(consensus)	44576-44577	base quality below threshold
Contig8	(consensus)	44579-44581	base quality below threshold
Contig8	(consensus)	44589-44591	base quality below threshold
Contig8	(consensus)	44595-44609	base quality below threshold
Contig8	(consensus)	44604-44609	6 bp single subclone

Figure 20. Navigator showing problems remaining after all three rounds of reactions. Problems in green boxes are at the fosmid ends and will be resolved by sequencing off the ends of the fosmid.

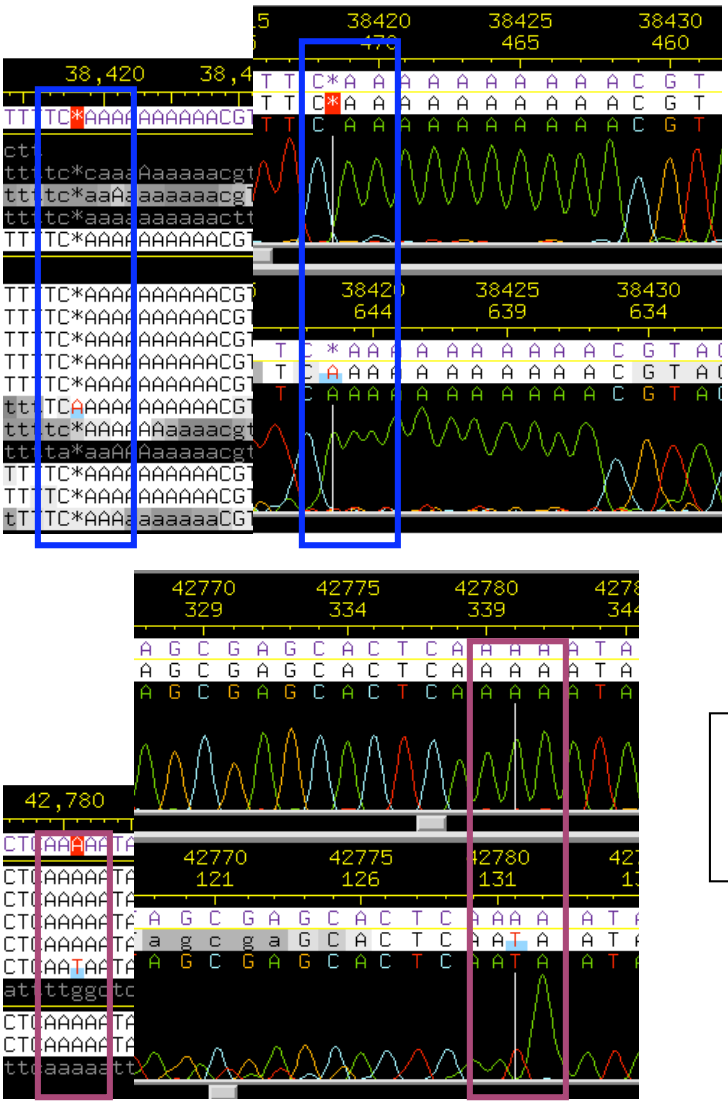
All three of the high quality discrepancies shown in Figure 20 were examined; none were genuine high quality discrepancies. These locations were marked with a comment tag (Figure 21). The read containing a discrepancy at bp 17596 (Figure 21.1) contains messy sequence trace before and after the discrepancy, as well as questionable peak spacing at the discrepancy. This caused me to conclude that the discrepancy is not high quality, and the consensus sequence is correct.

The discrepancy at bp 38418 (Figure 21.2) is in a run of 11 A's. The discrepant read contains clear peaks for each A. However, there are an overwhelming number of reads with 10 A's, so I am confident in the accuracy of the consensus sequence.

The discrepancy at bp 42781 (Figure 21.3) is in a read with a questionable trace. The base following the discrepancy is a suspiciously high peak, and the sequence before the discrepancy contains messy trace, with multiple peaks showing at one location. The questionable accuracy of the trace led me to believe that the discrepant base is not very high quality, and does not challenge the accuracy of the consensus sequence.



21.1: HQD at bp 17596. Note messy trace before and after HQD and irregular spacing at HQD.



21.2: HQD at bp 38418. Note peak clarity in A runs and high number of reads.

21.3: HQD at bp 42781. Note high A peak at bp 42782 and messy trace before discrepancy.

Figure 21. Remaining high quality discrepancies (HQDs) in clone 430-O17

I marked the 150 bp single strand/ chemistry region at bp 8967-9111 with a comment tag because, although there are only two high quality reads at this location, there are also two lower quality reads that mostly align with the consensus sequence. I tried resolving this region 3 times using two primers, but none of the reactions produced high quality sequence at this location. A possible explanation for failure of 5' sequencing reactions is the presence of an $(AT)_n$ simple sequence repeat upstream of the single strand/ chemistry region. Since 3 reactions in the same direction (→) failed, it does not seem likely that more reactions will resolve the issue. Since the traces of the high quality sequences look good in this region and the base calls are supported by low quality data, the consensus sequence appears accurate and this region should not require additional finisher attention.

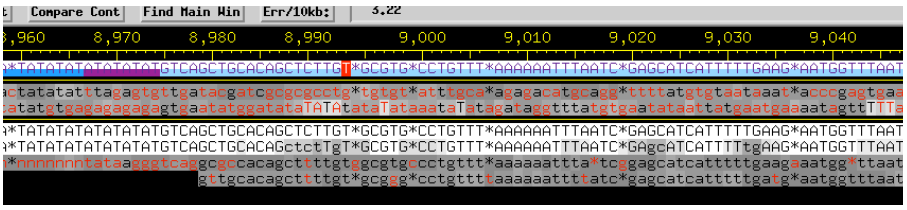


Figure 22. Section of 150 bp single strand/ chemistry region at bp 8967-9111

The 151 bp single strand/ chemistry region at bp 30299-30455 was tagged as "dataneeded" because after reviewing the area, I was not convinced of the accuracy of the consensus sequence. In much of this region, there is only one read that is high quality, and the sequence is not well-supported by low quality reads. This region used to be Gap B, and has been sequenced twice, but both resulting reads were not very high quality in this region. If possible, additional data supporting the single high quality read would increase confidence in the accuracy of the consensus sequence. However, the high quality sequence is above Phred 30, so additional sequence data is not absolutely necessary.

Figure 23. Section of 151 bp single strand/ chemistry region at bp 30299-30455.

Checklist

A BLASTN of my consensus sequence did not reveal any contigs containing vector or bacterial DNA. All BLAST results were sequences from *Drosophila* species.

Using Search for String, I found that my fosmid contains no mononucleotide runs of 15 bases or more. Search for String also confirmed that the consensus sequence does not contain any N's or X's. This confirmed that my consensus sequence did not contain any bases that could not be called by Phrap (N's) or any vector sequence (X's).

Final Assembly

Using the methods described above, I was able to eliminate most of the problems presented in the initial assembly. The final assembly meets the criteria of our finishing checklist, with the problem areas above tagged as necessary.

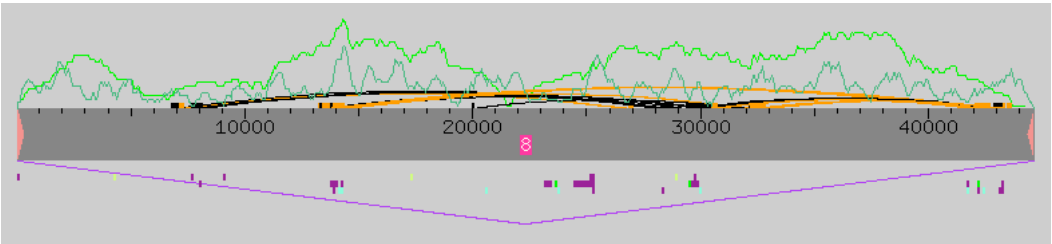


Figure 24. Final Assembly View of clone 430-O17

Supplement

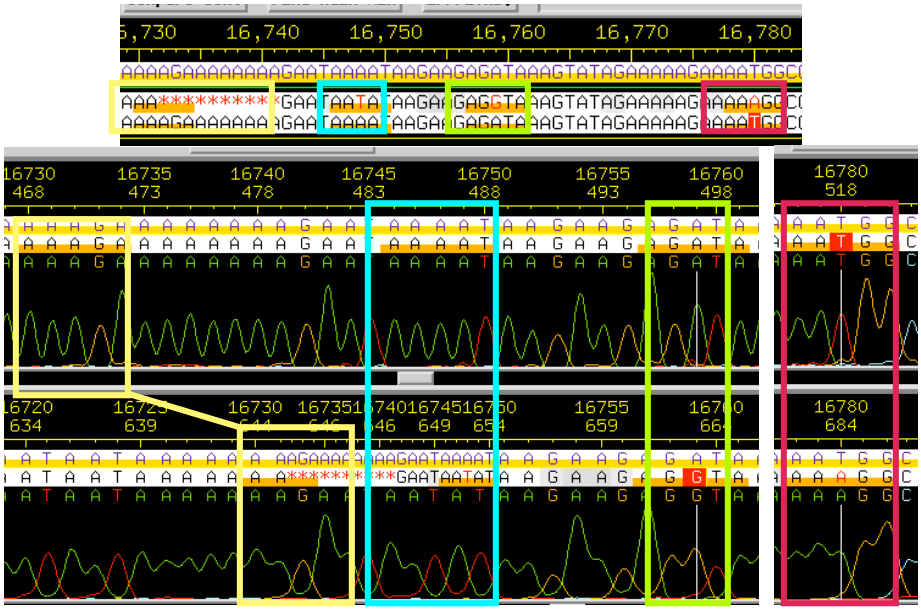


Figure 6.1 Tagged high quality discrepancies in original assembly Contig 8

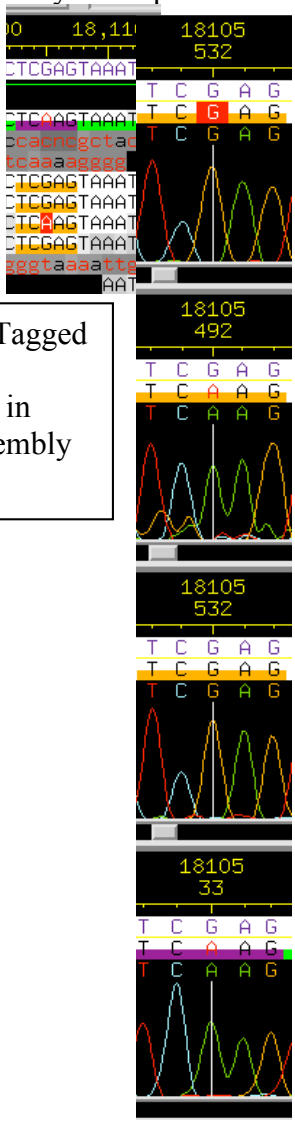


Figure 6.2 Tagged high quality discrepancy in original assembly Contig 8