Finished (Almost) Sequence of *Drosophila littoralis*
Chromosome 4 Fosmid Clone XAAA73

Seth Bloom
Biology 4342
March 7, 2004

Summary:

I successfully sequenced *Drosophila littoralis* fosmid clone XAAA73. The first round of sequence assembly revealed three major contigs that required three rounds of additional sequencing to be joined. In addition, one low-quality region of sequence within one of the contigs was improved in further rounds of sequencing. Two subclones spanned one of the gaps, but the area required sequencing technologies other than BigDye to be joined, and sequence quality at the site remains low. The other gap may or may not have been spanned by a single subclone, but it required spanning PCR-sequencing to be joined, and sequence in the area remains of very low quality. The ends of the fosmid have been identified, and *in silico* EcoRI and SacI restriction digests of the assembly match the physical digests of the fosmid clone. However, due to low quality areas of sequence, the fosmid cannot be considered completely finished and will require at least one more round of sequencing before all low-quality areas are resolved.

Preliminary analysis:

Fosmid clone XAAA73 derived from *Drosophila littoralis* Chromosome 4 was subcloned into *E. coli* sequencing vectors and sequenced with BigDye sequencing technology at the Washington University Genome Sequencing Center. Initially one 96-well plate of sub-clones (plate uub48) was sequenced with both forward and reverse sequencing primers. I was out of town, so the sequencing reactions were performed for me by course TA Libby Lawson. I downloaded the 192 sequencing reads from the 96 subclones onto a Unix-based Apple computer and compiled it using the *phredphrap* program, then viewed the resulting compilation using the program *consed*. The assembly appeared as shown in **Figure 1** with seven major contigs of compiled sequence and a variety of spanned and unspanned gaps.

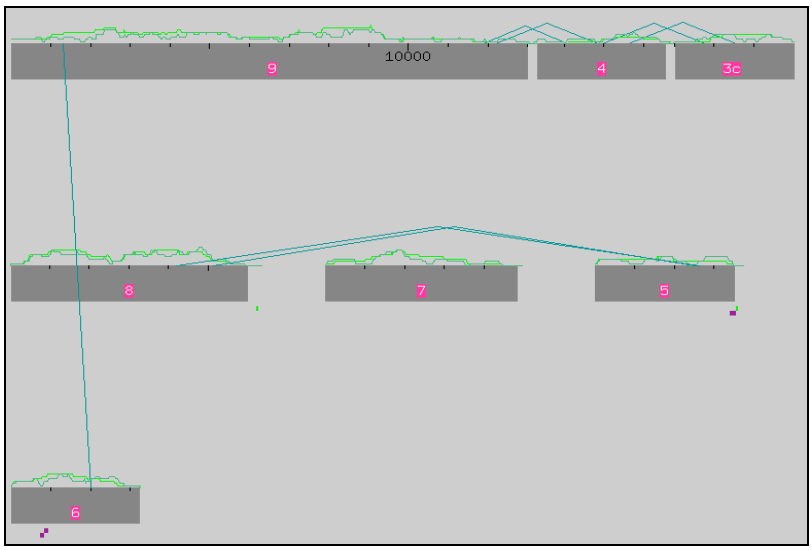

Figure 1: *Consed* assembly of preliminary *phredphrap*-compiled sequence reads from sequencing plate uub48, sequenced with both forward and reverse primers. Blue lines represent forward and reverse sequence reads from the same subclone but aligned to different contigs, suggesting that the subclones span the gap between the two contigs.

First Round Analysis:

This preliminary data was supplemented with sequence reads from two more 96-well plates of subclones from XAAA73 (plates uua76 and uua79) and I compiled these first round sequencing reads together with plate uub48 using *phredphrap* and analyzed them in *consed*. The assembly revealed three major contigs (see **Figure 2**), with the gap between contigs 2 and 3 apparently spanned by one subclone and the gap between contigs 3 and 4 apparently spanned by two subclones.

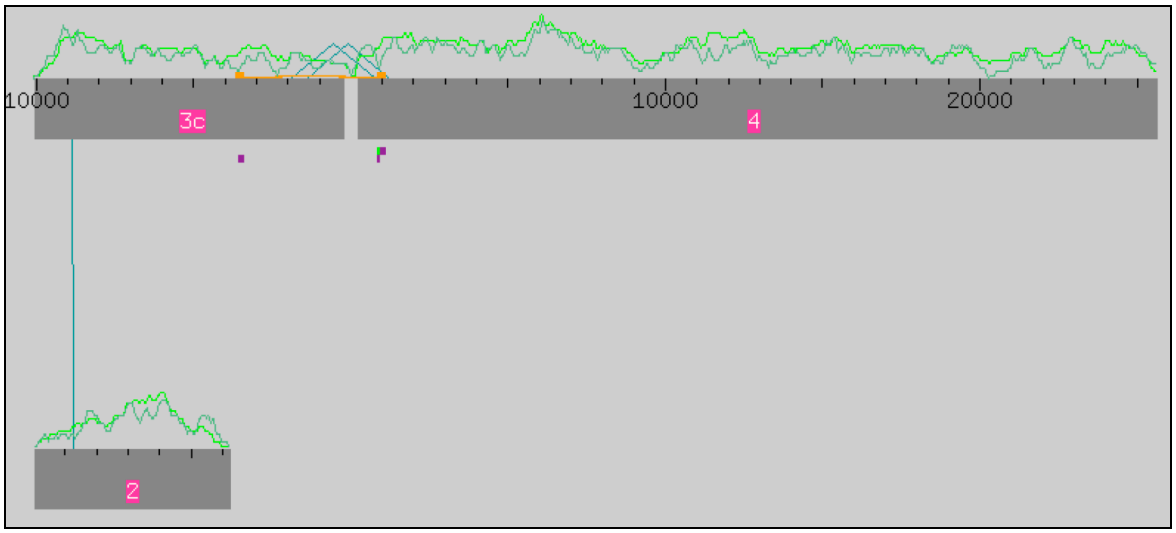

Figure 2: Assembly of first round sequence data, revealing three major contigs with both gaps apparently spanned by at least one subclone.

Analysis in *consed* revealed only one high quality discrepancy in the assembly, occurring between read uub48c03.g1 and consensus sequence in the assembly. The consensus sequence at this site had a run of 3 T bases, but clone uub48c03.g1, which otherwise aligned well, had a run of 4 T's in the same location (see **Figure 3**). However examination of the relevant chromatogram files in *consed* demonstrated the extra T base in uub48c03.g1 was likely called incorrectly by *phred*. When the chromatogram was measured manually for length by "note-card analysis" it was found to be the same length as stretches from other aligned reads with higher quality sequence at the same location that *phred* had only identified as having a stretch of three T's. Thus the extra T is likely not real.

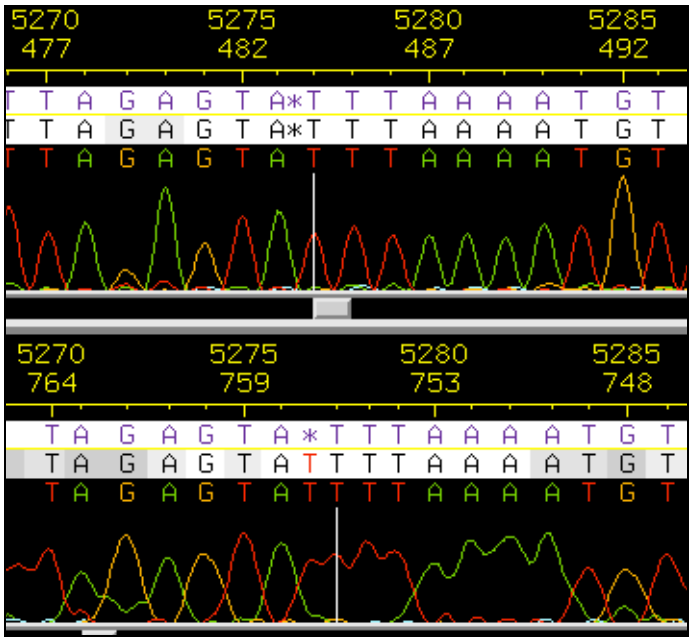

Figure 3: Alignment of read uub48c03.g1 (bottom) and another sample aligned read (top) at the site of alleged high-quality discrepancy between uub48c03.g1 and consensus. Physical measurement of the length of the run of T's demonstrates the discrepant extra T in uub48c03.g1 was likely mis-called by *phred*.

I next began analyzing sequence on contig 4. I first looked for evidence of one end of the fosmid clone. The cloning site for the fosmid vector has sequence GATC. Sequence analysis revealed that the right end of contig4 had sequence GATC as confirmed by alignment of multiple subclones, so I labeled it as one end of the fosmid clone (see **Figure 4**).

Figure 4: Sequence of right end of contig 4 from first round data, revealing what appears to be one end of the fosmid clone ending in restriction site GATC. (The X notations after the GATC indicate vector sequence, which could represent either fosmid or sequencing vector.)

Consed showed one low quality stretch of consensus sequence in contig 4 (see **Figure 5**).

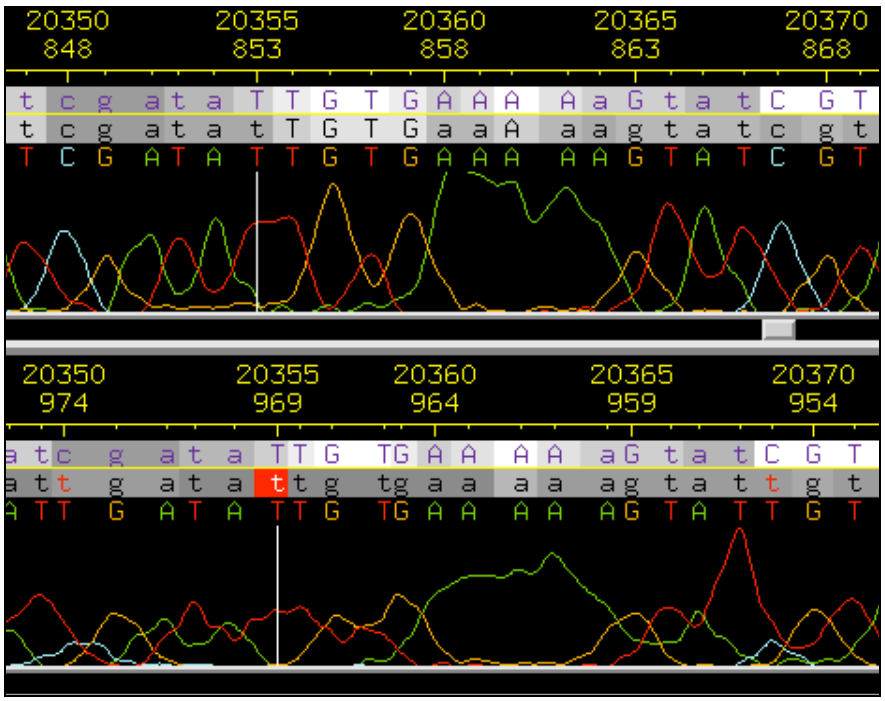

Figure 5: Low quality sequence in contig 4, consensus in region is uncertain. Ordered primer to sequence over region.

Since base quality of the two reads spanning the area appeared poor and consensus was uncertain, I ordered primer XAAA73.Oligo1 to sequence the subclone spanning the region (see ordered reads in **Table 1**, base sequences of all primers are in **Table 5** at end of report).

Table 1: Primers and reactions I ordered based on first round sequence data.

Primer Name	Position	Reason	Rxn(s). called
XAAA73.Oligo1	Contig4: 20593-20614 (C)	Low quality coverage	uub48e07_1.b1
XAAA73.Oligo2	Contig4: 198-220 (C)	Spanned gap with Contig 3	uua79h02_2.b1 uua79g04_2.b1
XAAA73.Oligo3	Contig3: 131-150 (C)	Spanned gap with Contig 4	uua79h02_3.b1 uua79g04_3.b1
XAAA73.Oligo4	Contig3: 9814-9832	Unspanned gap (with Contig 2?)	uua76c07_4.b1 uua79d07_4.b1
XAAA73.Oligo5	Contig3: 9814-9832	PCR primer for unspanned gap (with contig 2?)	XAAA73PCR5c6_5.b1
XAAA73.Oligo6	Contig2: 5971-5990	PCR primer for unspanned gap (with contig 3?)	XAAA73PCR5c6_6.b1
XAAA73.Oligo7	Contig3: 10007-10030	PCR primer for unspanned gap (with contig 2?)	XAAA73PCR5c6_7.b1
XAAA73.Oligo8	Contig2: 6145-6169	PCR primer for unspanned gap (with contig2?)	XAAA73PCR5c6_8.b1

The position of the primer with respect to the sequence assembly is shown in **Figure 6**.

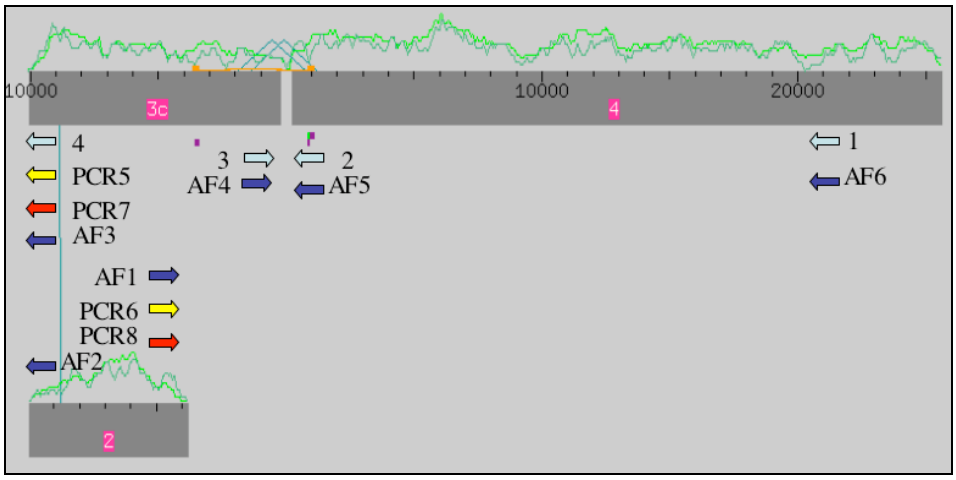

Figure 6: Sequencing and PCR oligos ordered based on first round data. Primers labeled "AF#" were ordered by AutoFinish, others were ordered by me. Primers labeled "PCR#" were used to perform PCR on the fosmid clone according to color-coded pairs, and then to sequence the PCR products produced. The other primers were used for normal sequencing on subclones. Primers are numbered in figure corresponding to the number at end of the primer name in Tables 1 and 2.

As noted earlier, there appeared to be a gap between contigs 3 and 4 spanned by two subclones. I ordered one primer for sequencing from each end of the spanned gap and ordered reads with both primers on both of the subclones spanning the gap (see **Table 1** and **Figure 6**.)

As stated earlier, there appeared to be a spanned gap between contigs 2 and 3. However closer examination of sequences from the subclone apparently spanning the gap revealed that the end of the subclone *phredphrap* had aligned in Contigs 2 was of extremely poor sequence quality and the alignment was a very bad match (see **Figure 7**), suggesting that the alignment might be incorrect and the gap between Contigs 2 and 3 might not be spanned after all.

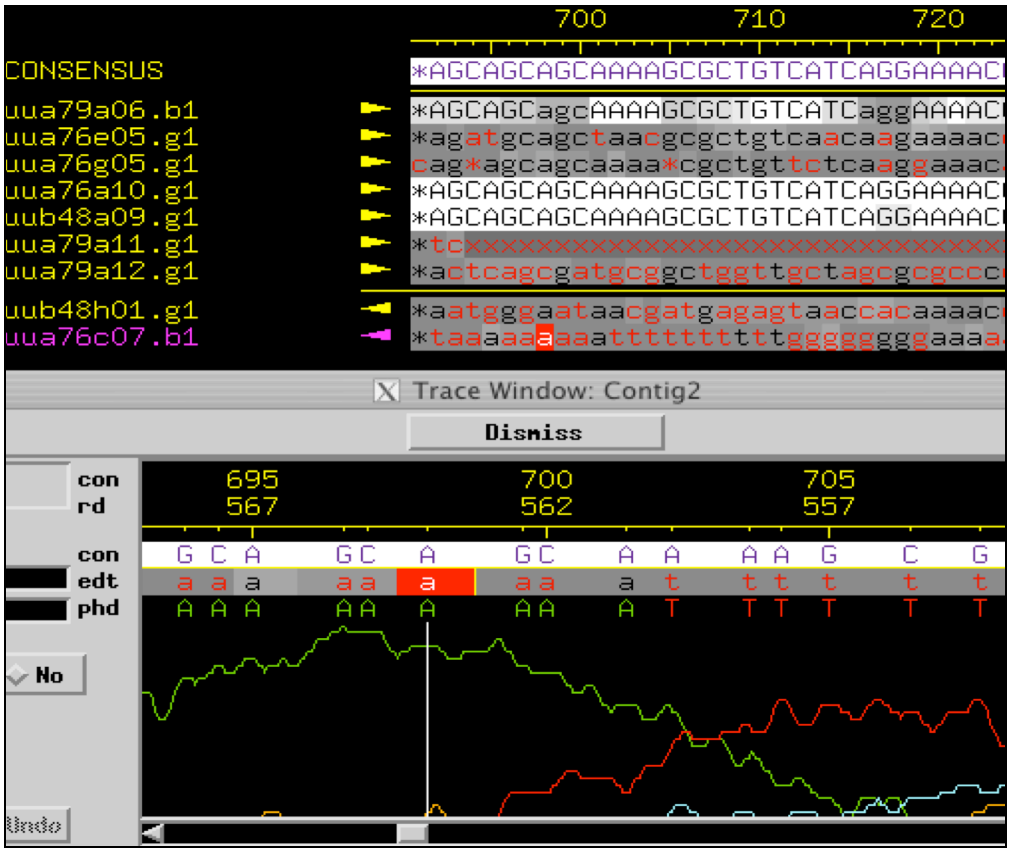

Figure 7: A typical stretch of sequence from the subclone aligned with Contig 2 that supposedly spanned the gap between Contigs 2 and 3. The sequencing read (uua76c07.b1) is highlighted in purple and the chromatogram trace is shown below. The quality of the sequence and its agreement with the consensus are both extremely poor, suggesting that the alignment might be incorrect.

The possibility that the gap between Contigs 2 and 3 might be unspanned raised the question of which end of Contig 2 actually formed gap with contig 3 and which corresponded to the end of the fosmid clone (sequence analysis indicated neither the left end of Contig 4 nor either of the ends of Contig 3 ended in a GATC sequence – data not shown). I examined both ends of Contig 2 and found that the left end, which had initially appeared to form one side of the spanned gap with Contig 3, actually ended in a GATC sequence, suggesting it corresponded to the end of the fosmid clone (see **Figure 8**).

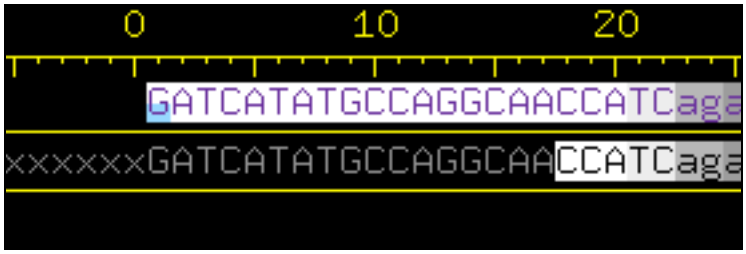

Figure 8: Left end of Contig 2, revealing GATC sequence. This suggested the left end of Contig 2 corresponded to the end of the original fosmid clone.

By contrast, the right end of contig 2 did not have a GATC sequence at the end, suggesting it did not match the end of the fosmid (see **Figure 9**).

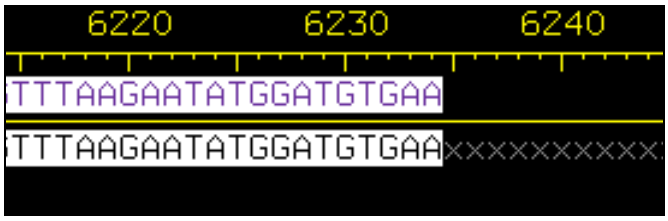

Figure 9: Right end of Contig 2, showing no GATC sequence. This suggested that the right end of Contig 2 represented one end of an unspanned gap with Contig 3, not the end of the fosmid clone.

I concluded that there was an unspanned gap reaching from the right end of Contig 2 to the left end of Contig 3, while the left end of Contig 2 represented the end of the fosmid clone. In an effort to span the gap, I ordered one primer for a sequence walk off the left end of contig 3 on two subclones that extended into the gap (see **Table 1** and **Figure 6**). I also called two sets of PCR primers spanning the gap and ordered sequencing reads that would sequence the PCR products using the primers that produced them.

In addition to the sequencing reads I ordered, AutoFinish ordered six sequencing reads and primers (see **Figure 6** and **Table 2**). The sequences AutoFinish ordered corresponded to most of my reads. Autofinish ordered sequence reads on subclones corresponding to all of the reads I ordered, but it did not order PCR reactions. Since it did not order a PCR reaction working off of the right end of Contig 2, it did order a single sequence read on a subclone walking off the right end of contig 2. It also ordered a read off the left end of contig 2, which I did not do because I considered it to be the end of the fosmid.

Table 2: Primers and reactions AutoFinish ordered based on first round sequence data.

Primer Name	Position	Reason	Matches my order?	Rxn(s). called
XAAA73.1	Contig2 6136-6157	Oligo walk off end (unspanned gap?)	No (I ordered PCR)	uua76d05_t1e1.g1
XAAA73.2	Contig2 567-551 (C)	Oligo walk off end (spanned gap?)	No (I thought end of fosmid)	uua79a06_t2e2.g1
XAAA73.3	Contig3 9859-9880	Unspanned gap (with Contig2?)	Yes	uua79d07_t3e3.g1
XAAA73.4	Contig3 240-222 (C)	Spanned gap with Contig4	Yes	uua79g04_t4e4.g1
XAAA73.5	Contig4 103-80 (C)	Spanned gap with Contig3	Yes	uua79h02_t5e5.b1
XAAA73.6	Contig4 20787-20765 (C)	Low quality coverage	Yes	uub48e07_t6e6.b1

Second Round Data:

My second round analysis had a limited amount of new data because the PCR sequences I had ordered and the sequence reads AutoFinish had ordered were not yet available. The first assembly of second round data showed essentially same three contigs as had been identified in the first round, although the arrangement on the screen was different (see **Figure 10**). The only difference was that Contigs 3 and 4 were now pointing in the opposite direction in the visual display, but it did not affect the relationship between them or Contig 2. The potentially unspanned gap between contigs 2 and 3 remained the same as noted in the first round.

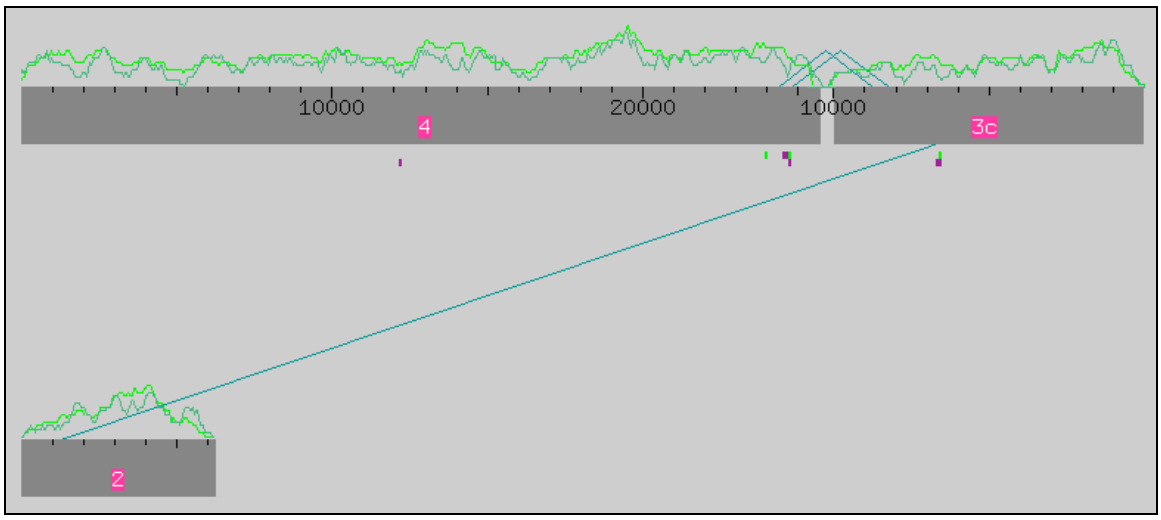

Figure 10: Initial assembly of second round sequence data. The assembly is not significantly different than the first round assembly (see Figure 2), other than a rearrangement of the contigs on the screen.

Slightly more sequence was available at spanned gap between contigs 3 and 4 because of new reads I ordered in the first round, however the short length and relatively low quality of the new reads suggested secondary structure problems in the area, implying that different sequencing chemistries might be needed to effectively span the gap. Nonetheless, on the basis of visual analysis, one of the GSC finishers and I performed a force join of Contigs 3 and 4 (see **Figure 11**).

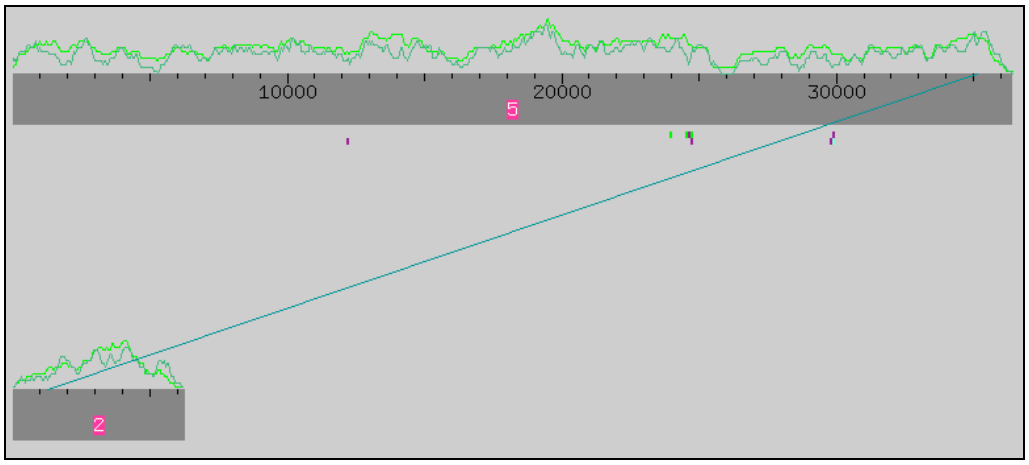

Figure 11: Assembly of second round data after force join of contigs 3 and 4 (site of join is around 26000 on new Contig 5).

The level of similarity of the sequences at the site of the force join area was still very tentative and of low quality (see **Figure 12**).

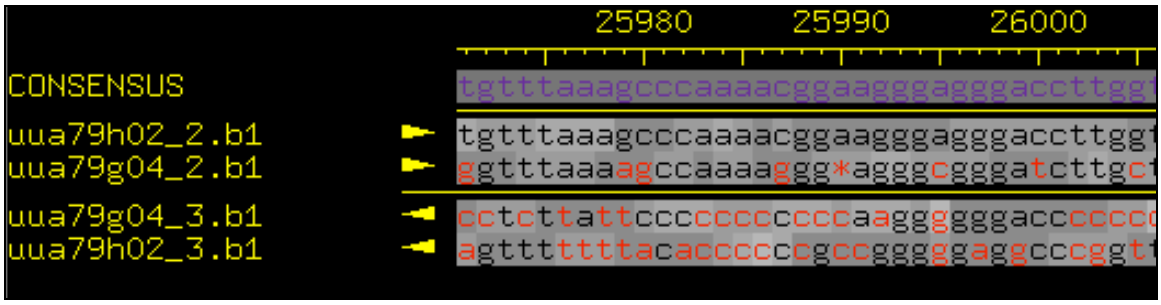

Figure 12: Sequence at site of force join between Contigs 3 and 4 to create new Contig 5. Quality of sequence and consensus is very low.

In an effort to improve the sequence at this area, I ordered sequencing reads across the area using existing primers but employing the dGTP and 4:1 sequencing chemistries instead of BigDye. I used existing vector sequencing primers and primers ordered in the first round for the second round reactions (see **Figure 13** and **Table 3**).

Table 3: Sequencing reactions ordered based on second round data using all existing primers. Vector (F) represents the forward sequencing vector for the sequencing subclone vector.

Primer Name	Position	Reason	Rxn(s). called
Vector (F)	Contig5 ~ 24360	subclone spanning force-joined gap, repeat	uua79g04.b2
Vector (F)	Contig5 ~ 24800	subclone spanning force-joined gap, repeat	uua79gh02.b2
XAAA73.Oligo2	Contig5: 25385-25407	subclones spanning force-joined gap, new chemistries	uua79g04_t2.b1 uua79g04_g2.b1 uua79h02_t2.b1 uua79h02_g2.b1
XAAA73.Oligo3	Contig5: 26430-26449 (C)	subclone spanning force-joined gap, new chemistries	uua79g04_t3.b1 uua79g04_g3.b1 uua79h02_t3.b1 uua79h02_g3.b1
Vector (F)	Contig5: ~4280	subclone spanning low quality region, repeat	uub48e07.b2
XAAA73.Oligo1	Contig5: 4991-5012	subclone spanning low quality region, new chemistries	uub48e07_t1.b1 uub48e07_g1.b1

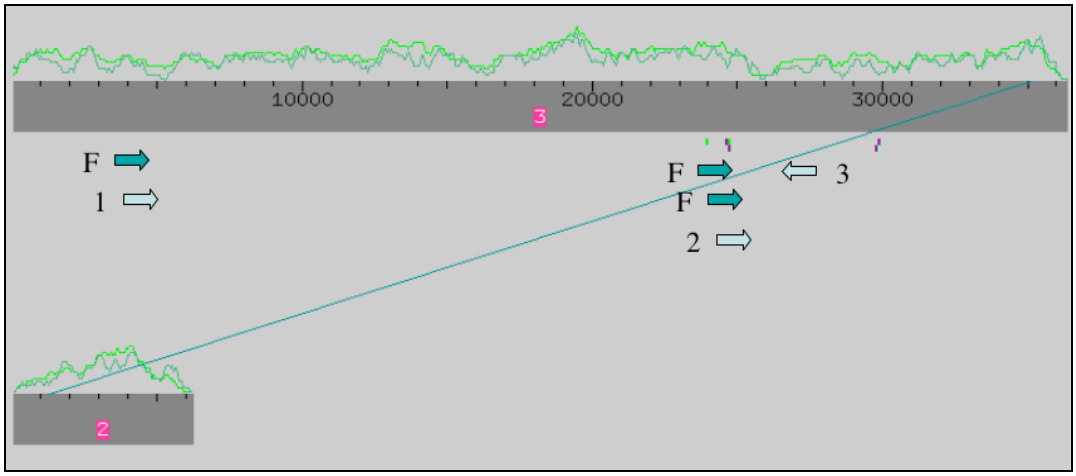

Figure 13: Sequencing reads ordered from second round data (see detailed descriptions in Table 3). Teal-colored primers labeled F represent the forward sequencing primer for the sequencing vector subclones. The light blue numbered primers correspond to primers I ordered in the first round (see Table 1).

The low quality area noted in the first round (see **Figure 5**) remains, although it was slightly improved by the new sequencing read ordered with primer XAAA73.Oligo1 (see **Figure 14**). However the short length and relatively low quality of the new read suggests secondary structure problems, implying a need for different sequencing chemistries. I ordered new reads using the vector sequencing primer and the previously ordered primer using dGTP and 4:1 sequencing chemistries (see **Table 3**).

Figure 14: Low quality region in Contig 5 noted in first round. Quality has improved slightly, but remains low.

Third Round Data:

The third round data included the Autofinish and PCR sequencing reads ordered in the first round as well as the sequencing reactions ordered in the second round. The assembly revealed two major contigs similar to what was produced via force-join in the second round (see **Figure 15**, compare with **Figure 11**).

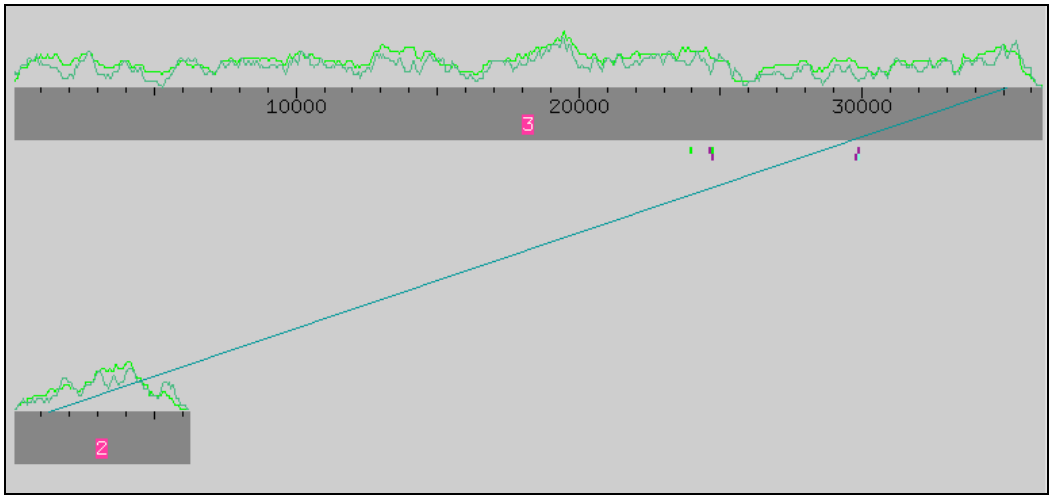

Figure 15: Initial assembly of third round sequence data. Assembly reveals two major contigs with the same potentially unspanned gap noted in first and second rounds.

Closer analysis reveals that the force join performed in the second round, although it occurred between two adjacent contigs, was not quite correct. Including the new sequencing reads generated by alternate sequencing chemistries added new sequence to

the area and more accurately joined the gap. Read length and quality of the new reads remained inferior, reinforcing the hypothesis of secondary structure interfering with sequencing, but the alignment appears convincing and the consensus is reasonably solid (see **Figure 16**).

Figure 16: Sequence of join between Contigs 3 and 4 from first and second rounds. The spanning sequence comes from new reads using sequencing chemistries other than BigDye.

Looking at Contig 2, one of the PCR reads extending off the right end worked somewhat, but the others extending off the end of Contigs 2 and 3 did not. The PCR read that produced sequence extended off the end of Contig 2 and, oddly, revealed a GATC sequence at the right end of the contig (see **Figure 17**).

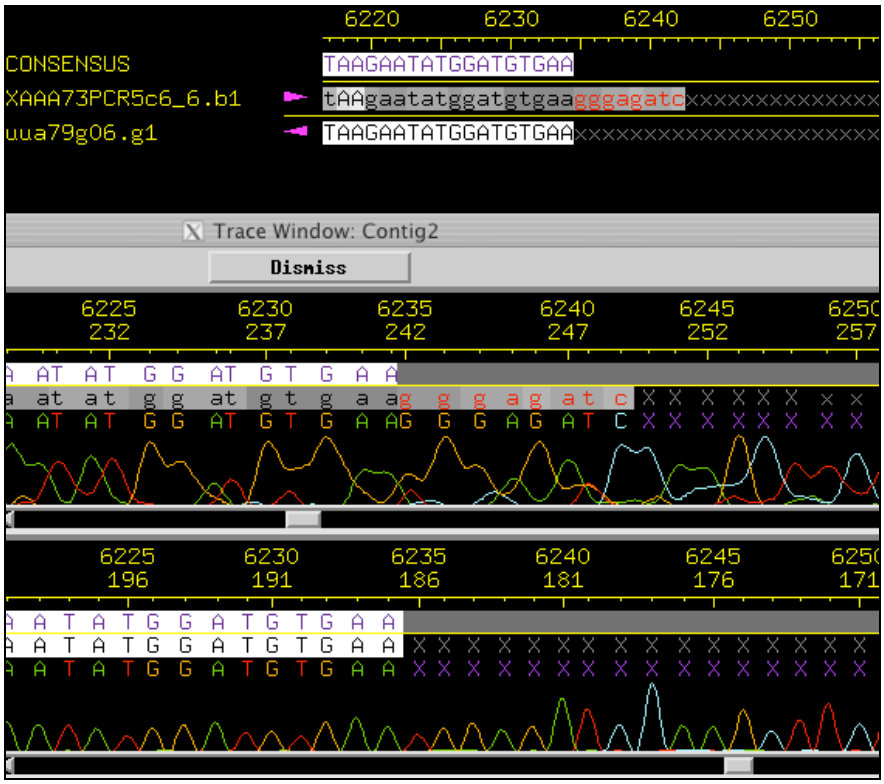

Figure 17: GATC sequence at right end of Contig 2. Sequence represents one end of fosmid clone as determined by analysis of vector sequence.

However Contig 2 had already been noted to end in GATC on the left side in the first round (see **Figure 8**), and we had assumed that this was the end of the fosmid. Closer analysis of sequence revealed that the previously identified GATC sequence on the left end of Contig 2 corresponded to a cloning insertion site into the subclone vector, whereas the new GATC sequence on the right end of the contig corresponded to the end of the fosmid because the sequence extending off the end matched fosmid vector sequence (data not shown). Thus the gap between Contigs 2 and 3 in fact stretches from the opposite end of Contig 2 than what I had previously supposed in designing PCR reactions. This raises the question of why the PCR reaction that produced this data succeeded in generating a read at all, since the primers should not have been oriented correctly to produce a product. The PCR sequence read is correct because a subclone sequence read also confirms it (see **Figure 17**). The PCR sequence read only extended out to a short distance at low quality, so perhaps it worked because of a false priming event in the PCR step.

I rearranged the contigs visually by complementing them to produce a more accurate picture in assembly view. In new arrangement, the right end of contig3 corresponds to one end of the fosmid clone and the left end of contig2 corresponds to the other end of the fosmid clone (see **Figure 18**).

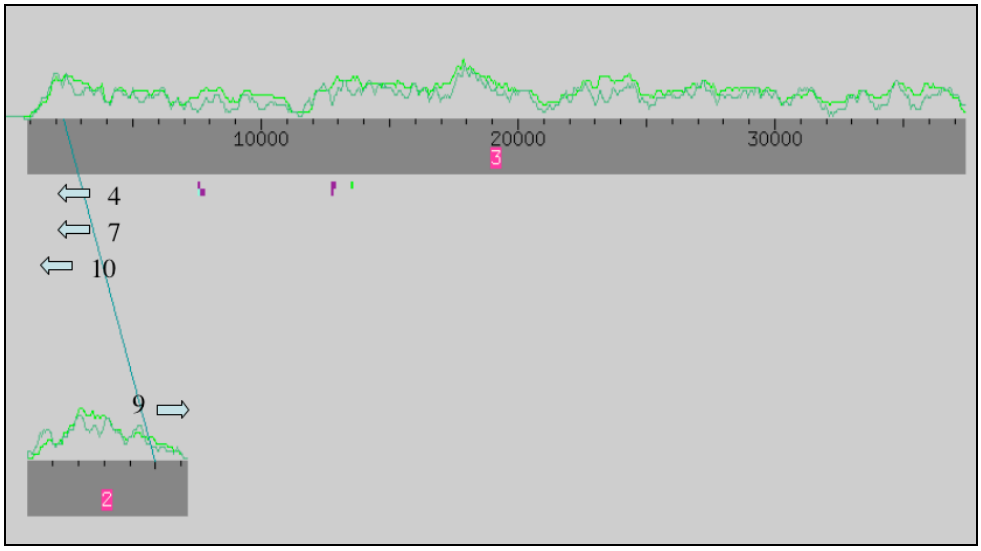

Figure 18: Rearranged assembly view of third round data with sequencing and PCR primers shown. The right end of Contig 3 corresponds to one end of the fosmid clone and the left end of Contig 2 corresponds to the other.

The gap between Contigs 2 and 3 might or might not be spanned, depending on whether the alignment of the sequence shown to be spanning the gap in assembly view is correct (see comments in **Figure 7**). In addition, a further read that had previously been unaligned was added to the left end of Contig 3 extending into the unspanned gap (data not shown). The alignment of this read into Contig 3, although apparently correct, was of low quality and needed to be confirmed with further sequencing. I designed a new PCR primer extending off new left end of Contig 3 and a new PCR primer extending off right end of Contig 2. Using those primers and previously ordered primers, I ordered two PCR reactions spanning the gap and sequenced them with all available primers and

chemistries. I also ordered reads further sequencing the subclone at end of contig3 (see **Table 4**).

Table 4: Reactions and primers ordered based on third round data. New and existing primers, * indicates newly ordered primer.

Primer Name	Position	Reason	Rxn(s). called
XAAA73.Oligo4	Contig3: 1319-1337 (C)	Sequencing low quality sequence at end of contig3 and PCR primer for spanning gap with contig2 with all available chemistries.	uua79d07_g4.b1 uua79d07_t4.b1 XAAA73PCR9c4_4.b1 XAAA73PCR9c4_g4.b1 XAAA73PCR9c4_t4.b1
XAAA73.Oligo7	Contig3: 1121-1144 (C)	Sequencing low quality sequence at end of contig3 and PCR primer for spanning gap with contig2 with all available chemistries.	uua79d07_g7.b1 uua79d07_t7.b1 XAAA73PCR9c4_7.b1 XAAA73PCR9c4_g7.b1 XAAA73PCR9c4_t7.b1
XAAA73.Oligo9*	Contig2: 6037-6057	PCR primer for spanning gap with contig3, sequenced with all available chemistries.	XAAA73PCR9c4_9.b1 XAAA73PCR9c4_g9.b1 XAAA73PCR9c4_t9.b1 XAAA73PCR9c10_9.b1 XAAA73PCR9c10_g9.b1 XAAA73PCR9c10_t9.b1
XAAA73.Oligo10*	Contig3: 211-229 (C)	PCR primer for spanning gap with contig2, sequenced with all available chemistries.	XAAA73PCR9c4_10.b1 XAAA73PCR9c4_g10.b1 XAAA73PCR9c4_t10.b1 XAAA73PCR9c10_10.b1 XAAA73PCR9c10_g10.b1 XAAA73PCR9c10_t10.b1

Fourth Round Data:

New data from the sequencing reactions ordered in the third round finally succeeded in producing one large contig (see **Figure 19**).

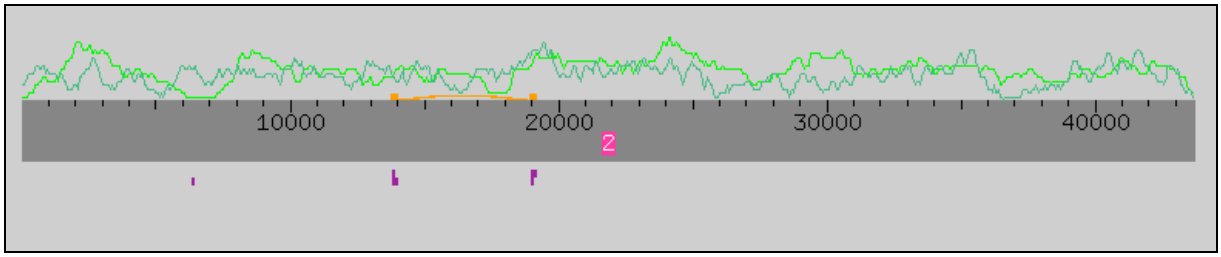

Figure 19: Initial assembly of fourth round sequence data. Assembly reveals one large contig.

The PCR sequencing reactions closed the gap between what had been Contigs 2 and 3, although quality is low at the join (see **Figure 20**).

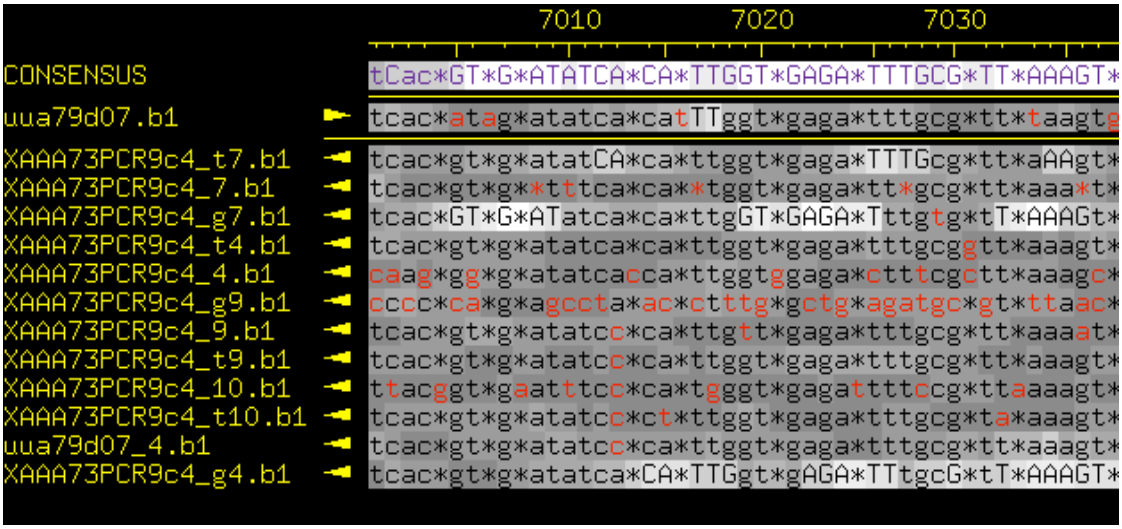

Figure 20: Gap between what had been Contigs 2 and 3 joined by PCR sequencing, but quality is low.

Examination of *in silico* EcoRI and SacI digests of the assembled contig reveal DNA fragments of the same size as noted in physical digests of the actual fosmid (see **Figures 21** and **22** respectively).

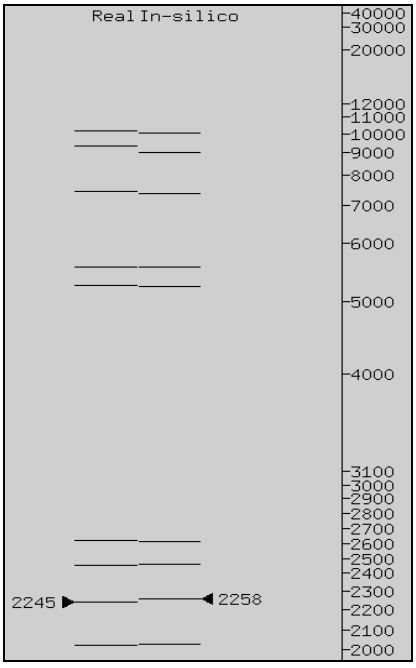

Figure 21: Physical (left) and *in silico* (right) EcoRI digests of fosmid clone. All bands appear to match in size, suggesting correct assembly of sequence.

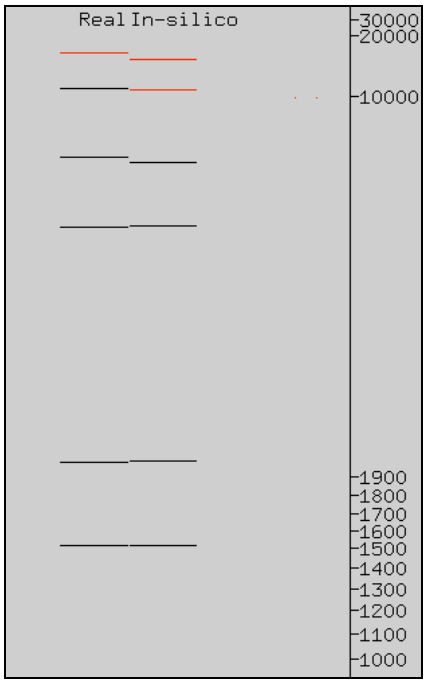

Figure 22: Physical (left) and in silico (right) SacI digests of fosmid clone. All bands appear to match in size, suggesting correct assembly of sequence.

As a final editing notation, I manually adjusted the left end of the to be consensus by removing some of the X's that were present in the initial assembly to reveal the last few bases of the cloning site in the subclone covering the region (see **Figure 23**, compare with initial sequence as called in **Figure 17**), and marked it as the end of the fosmid clone.

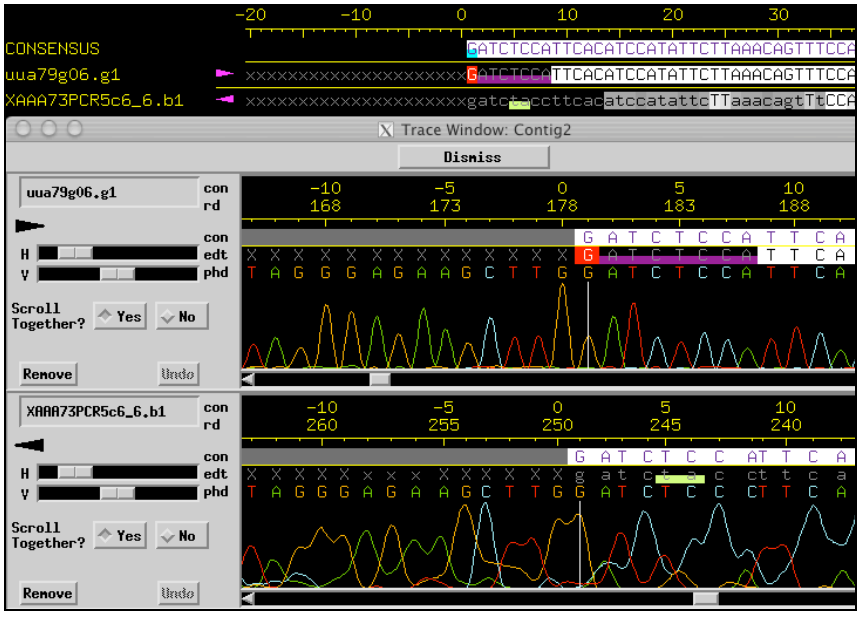

Figure 23: Manually edited consensus shwoing cloning site in fosmid vector (compare with Figure 17 showing area before editing). Tagged left end of contig as end of fosmid.

As shown in **Figure 19**, there is a tandem duplication on the contig. The 226 bp stretch of sequence from 13756-13982 and the stretch from 18934-19160 are duplicated in tandem with 100% similarity.

Final Finishing Checklist:

- The vector cloning sites are verified.
- I could not check orientation of the clone with respect to other clones without more information.
- The clone size is 43671 bps, which is the expected size.
- *In silico* digests appear to match the physical digests of the clone.
- Remaining problems and comments:
    - The contig includes numerous stretches of single strand/chemistry sequence, and some areas covered only by a single subclone sequence of high quality.
    - There is a high quality extra base in clone uua76g04.g1 at consensus pos 966, but "notecard analysis" demonstrates it to be likely spurious (data not shown).
    - The high quality discrepancy in sequencing read uub48c03.g1 discussed above (see **Figure 3**) has not been resolved, but is likely not real as noted earlier.
    - Significant stretch of sequence with low quality or quality below threshold at site of the join made by PCR sequencing (position approximately 7000 bps in contig).
    - Some low quality sequence remains in the area around 17800 bps, site of the original spanned gap between contigs 3 and 4. The precise reason for the low quality sequence is not immediately clear, but the region was only successfully sequenced with chemistries other than BigDye.
    - As noted above, the *in silico* restriction digests appear to match the physical digests.
    - Sequence analysis revealed two stretches of T's, both with length 18 bps, at positions 18892-18907 (uncomplemented) and 33610-33625 (complemented). However sequence quality of reads bordering these runs was generally high, and they did not appear to have lead to problems in base-calling by *phred* (data not shown). There were no stretches of C's 16 bases or longer.

Conclusion:

The clone is not quite finished yet. Two areas of low quality at the sites of the two joined gaps still need to be resolved by a further round of sequencing.

Table 5: Nucleotide sequences of all primers ordered for sequencing of fosmid clone XAAA73.

Primer Name	Ordered By:	Round:	Sequence
XAAA73.Oligo1	Me	First	ccactggtttaagaaatgtcac
XAAA73.Oligo2	Me	First	gatgttagccaaggaaattttat
XAAA73.Oligo3	Me	First	ctggtgatcaactttatgcc
XAAA73.Oligo4	Me	First	gcaccaaaggaattcaaga
XAAA73.Oligo5	Me	First	gcaccaaaggaattcaaga
XAAA73.Oligo6	Me	First	ggaaaagtcctgacttcgtg
XAAA73.Oligo7	Me	First	ccttttctttattttgaatttcag
XAAA73.Oligo8	Me	First	cagggataagaatataatgaacttg
XAAA73.Oligo9	Me	Third	gggtatgattgcgatttatct
XAAA73.Oligo10	Me	Third	cgattcccatggagcatac
XAAA73.1	Autofinish	First	gccattactcagggataagaat
XAAA73.2	Autofinish	First	caaagactcgagacccg
XAAA73.3	Autofinish	First	aggaataagcgtaaactcaaca
XAAA73.4	Autofinish	First	ttcattttgctcgtgtttg
XAAA73.5	Autofinish	First	gctttgtataagcaaattgaaagg
XAAA73.6	Autofinish	First	aagtttaattccaagctcctaga