Tien Chusak
Finishing paper: XBAA-40A16
March 23, 2005

Finishing the fosmid XBAA-40A16

Drosophila virilis is a species of *Drosophila* that diverged from *D. melanogaster* approximately 60 million years ago. In recent studies, researchers have discovered that the two species exhibit some very striking differences in the structure of their chromosomes. More specifically, the fourth chromosome of *D. melanogaster* has been shown to be almost exclusively heterochromatic: condensed, replicating late in S-phase, with no meiotic recombination. The same chromosome in *D. virilis* appears to be euchromatic. The goal of our study is to sequence, finish, and annotate the fourth (also commonly called the "dot") chromosome of *D. virilis* in order to determine whether or not heterochromatic and euchromatic domains can be distinguished based on gene and sequence characteristics.

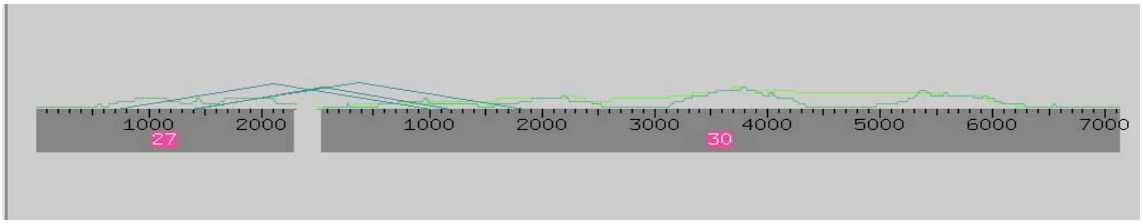


Fig. 1: The assembly of my clone at 1.64x coverage

Viewing my clone in Consed with only 1.64x coverage (Fig.1) showed only one gap that was spanned by 3 subclones. Interestingly, my assembly view shows only ~7 kb, when there should be 35-40 kb in my clone. This is probably due to Phred/Phrap Consed not having enough data to be able to bring in a lot of the reads. Upon addition of the rest of the reads (Fig.2), I saw that there were four major contigs separated by three gaps. Multiple subclones spanned all of the gaps and all of the forward/reverse read pairs were consistent.

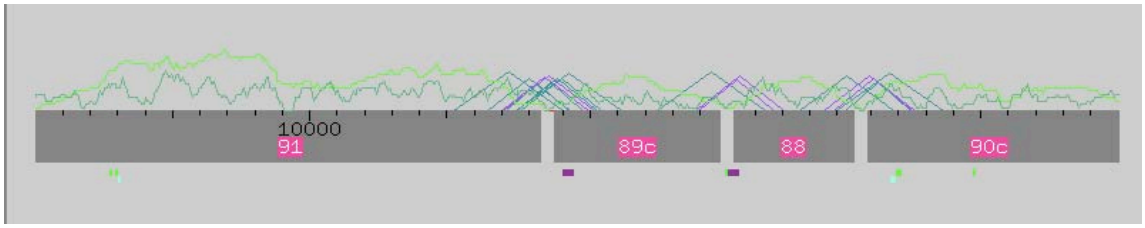


Fig. 2: Assembly with the rest of the reads added in

Closing the three gaps did not appear to be a major problem since they all had spanning subclones. In addition, there were several regions of low read depth indicated by the colored lines in Fig. 2, which often results in low consensus quality. My goal for the first round of reactions was to close all three gaps; no reactions were called in the first round to increase read depth. Table 1 summarizes the details of the reactions called. All

were performed using Big Dye, dGTP, and 4:1 chemistries. This strategy, while increasing costs, can save time in the finishing process.

Table 1: Round 1 reactions

Oligo	Sequence	Contig	Direction	Template	Goal
1	ggctcaccacacagctt	89	->	Aac50f05	Join 89-88
2	gccaagcaacaatttaaaact	88	<-	Aac50f05	Join 89-88
3	cgaggagggtaagtaccaga	91	->	Aac50f07	Join 91-89
4	tgggcagatcatggg	89	<-	Aac50f07	Join 91-89
5	ttcaagctttctagatctttgc	88	->	Aac49h10	Join 88-90
6	aaactttaagctcgcttctatatg	90	<-	Aac49h10	Join 88-90

Suggested reactions called by Autofinish were also used to compare Autofinish-assisted finishing to human-only finishing. Autofinish called reactions to increase read depth, which was something that I chose to postpone until later sessions calling reactions. Autofinish chose the oligos indicated by the arrows in the following diagram.

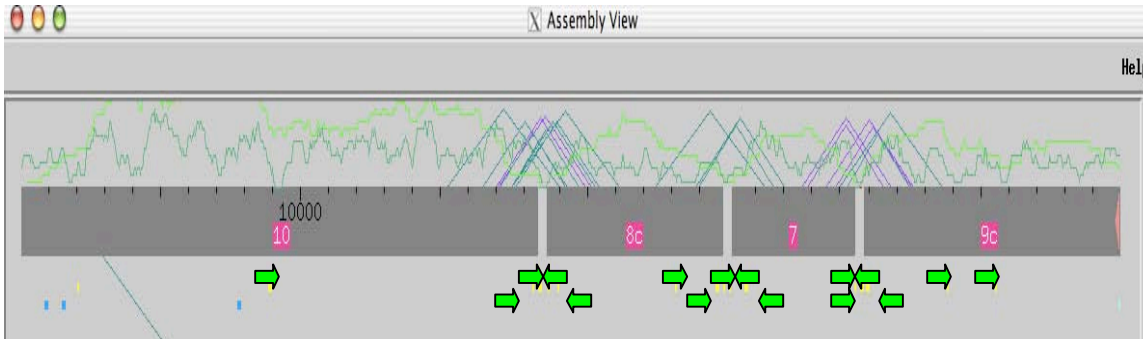


Fig. 3: Oligos chosen by Autofinish for round 1 of finishing. Arrows indicate oligos and direction.

As Fig. 3 illustrates, Autofinish called two primers on each side of the gaps while I called only one. There is a simple reason for this difference. Autofinish has assumed that there could be problems with certain regions close to the gap that caused the original reads to terminate. As such, Autofinish called oligos at varying distances from the gaps. I failed to realize that reactions fail at a fairly high rate, and I felt that using one oligo on each side of a gap was sufficient. We will return to this comparison in later rounds of calling reactions.

With the data from the first round of reactions, I was able to simply add the resulting reads to the assembly using Phred/Phrap. What I found was that the new reads closed two of the three gaps; the reactions designed to cover the third gap failed. My second round of reactions was dedicated to closing this gap.

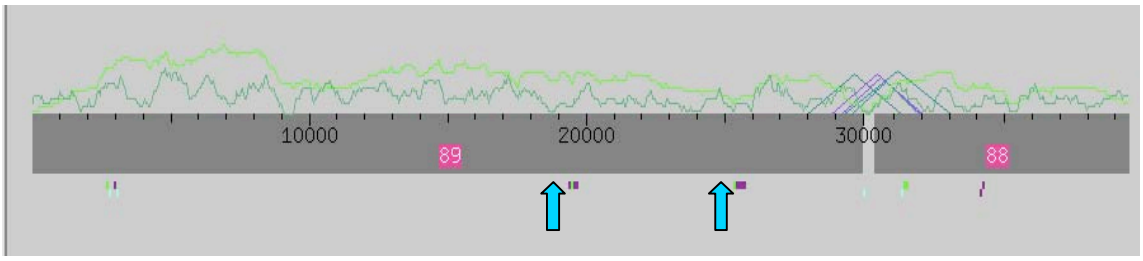


Fig. 4: Second assembly with round 1 reads added. Arrows indicate gaps that were closed.

Using the same oligos that I chose for reactions in the first round of reactions, I selected two new templates to use. Running reactions with oligos 5 and 6 yielded excellent results on templates aaa25f03 and aaa26f12. Upon addition of these reads to my assembly (Fig.4), the gap was closed. After only two rounds of reactions, my clone was in a single contig!

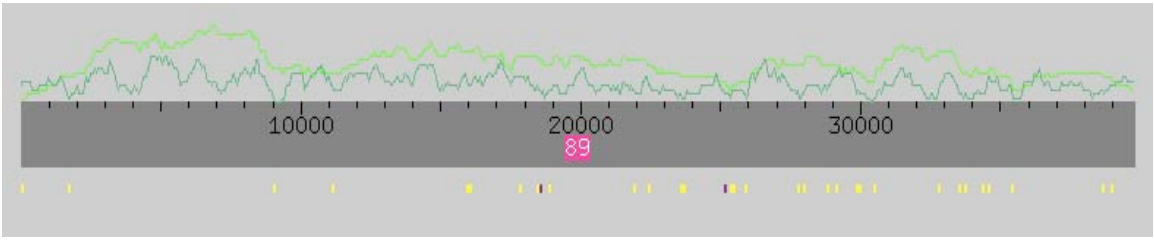


Fig. 5: Final assembly; clone is in one contig after the reads from the second round were added

For the third, fourth, and fifth rounds of reactions, regions with either low read depth or low consensus quality were my primary targets. I began encountering problems with reactions failing for various reasons, and my progress slowed tremendously. I called reactions with 26 new oligos, as well as some using old oligos on new templates in an effort to get at least a couple reactions to work. Table 2 details the reactions I called in the later rounds to finish my fosmid.

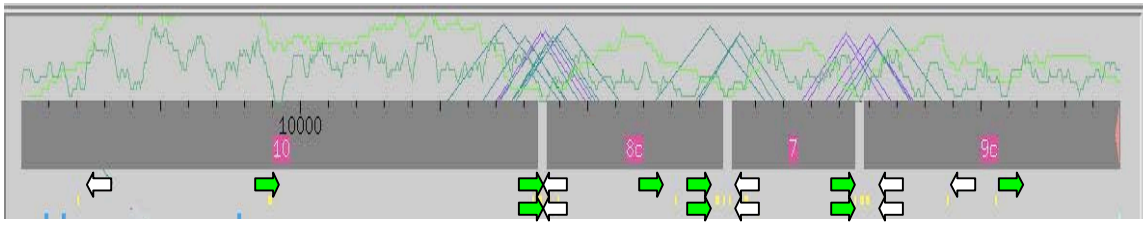
Table 2: Reactions from rounds three through five.

Oligo	Sequence	Direction	Template
7	agcgaaccgtaattcctc	->	aaa26e11_7.b2 aaa25a05_t7.b1
9	caagatgagagatcatcagtcttt	->	aac50c12_9.b2
10	accgtttaggatataagaaagtca	->	aaa25f09_10.b1
11	ttgtttgggctagaagcag	->	aaa26e01_11.b2
12	ttatgtataacaaccaactgtatgc	->	aac50f05_12.b1
13	agtggcaagcaggacc	->	aaa26e01_13.b1
14	cacaagagattataacttttgcc	->	aac50h02_14.b1 aac50h02_t14.b1
15	ggcatctggtccaaactc	<-	aaa26c08_15.b1
16	agcactgtattgattaattacacg	<-	aaa27c05_16.b1
17	cgccagtaccgtatactcg	->	aac50g07_17.b1

18	agcatgtccctcatatactacg	->	aac50g07_18.b1
19	agtgggtgaagattcatgg	->	aaa25f03_19.b1
20	agtaaagcacggtaaacgc	<-	aaa26f12_20.b1
21	gctttgctaacaaagaaatcg	->	aaa27d03_21.b1
22	tttcatccgtaaatatttaacaagt	<-	aaa25f08_22.b1
23	tgcgcacgcactagg	<-	aaa25f08_23.b1
24	ttaagcgtggcattatatcg	->	aaa26c04_24.b1
25	aacaaagggtgtcacgg	->	aaa26c04_25.b1
26	gctgatgatattaatgggaattg	->	aaa26c06_26.b1 aac49d11.b1
27	aggcaggtctgagaaaatg	<-	aaa26c06_27.b1 aac50h09_27.b1 aac50h09_t27.b1
28	tgtgcaatcgcagtaatg	<-	aac50a12_28.b1 aaa25g06_t28.b1
29	tggttgcgctcctgg	->	aaa25e03_29.b1 aaa25c03_t29.b1
30	ggatacatttggtttatgggtg	<-	aaa25e03_30.b1 aac50d07_30.b1 aac50d07_t30.b1
31	cttggtcttgttggtcc	<-	aaa25a02_31.b1 aaa26a05_31.b1 aaa26a05_t31.b1 aaa26d07_31.b1
32	gaaaataactgtctgcggg	->	aaa26d07_32.b1 aaa25a02_32.b1 aaa25a02_t32.b1

I received very little useable data from these reactions. A possible reason for such a high failure rate was that there were some organizational problems with some of our subclones, and calling reactions using these subclones resulted in nothing more than failed reads. Dye blobs also plagued my traces, and often resulted in miscalled bases if the sequence was actually continued past the blob.

A comparison with Autofinish shows that I called many more reactions, but my reason for calling each new oligo was due to a problem that had not yet been encountered when Autofinish was calling reactions.



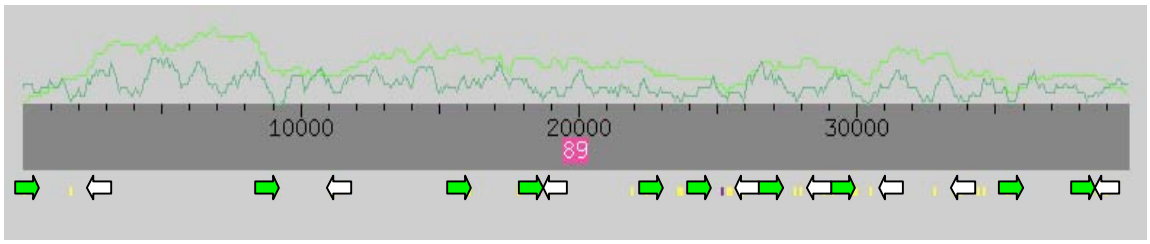


Fig. 6: Autofinish (top) vs. me (bottom). The yellow tags indicate oligos. Autofinish did not attempt to cover single strand/single chemistry regions that I did. Our reactions for improving sequence quality line up very well.

Since I had no new data from the reactions, my only option was to go through and attempt to edit regions that were of low consensus quality or had other problems. I found that Consed had incorrectly called several bases due to background "noise" in the read, and correcting these calls eliminated several high-quality discrepancies that were originally present.

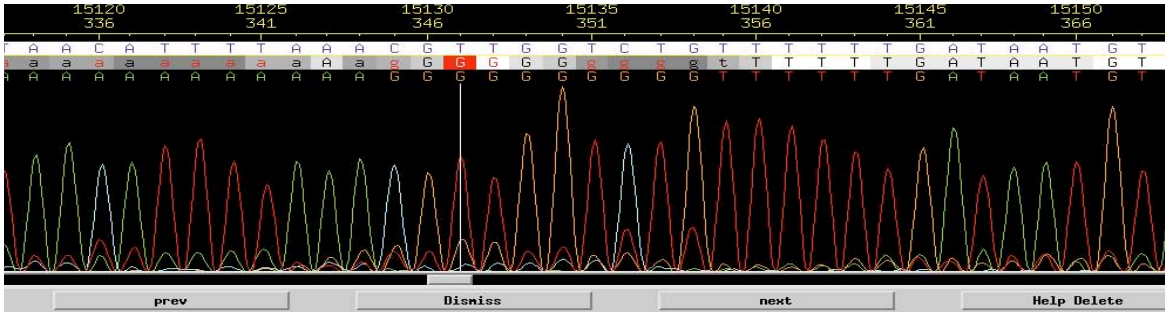


Fig. 7: Miscalled bases due to high background resulting in high quality discrepancies

I also encountered a region where Consed had misaligned two reads due to a mistake in calling the bases of one of the reads. This was causing problems downstream of the misaligned site. To correct this, I added pads to the region, and tagged the region so that the Quality Assurance group would know what I had done.

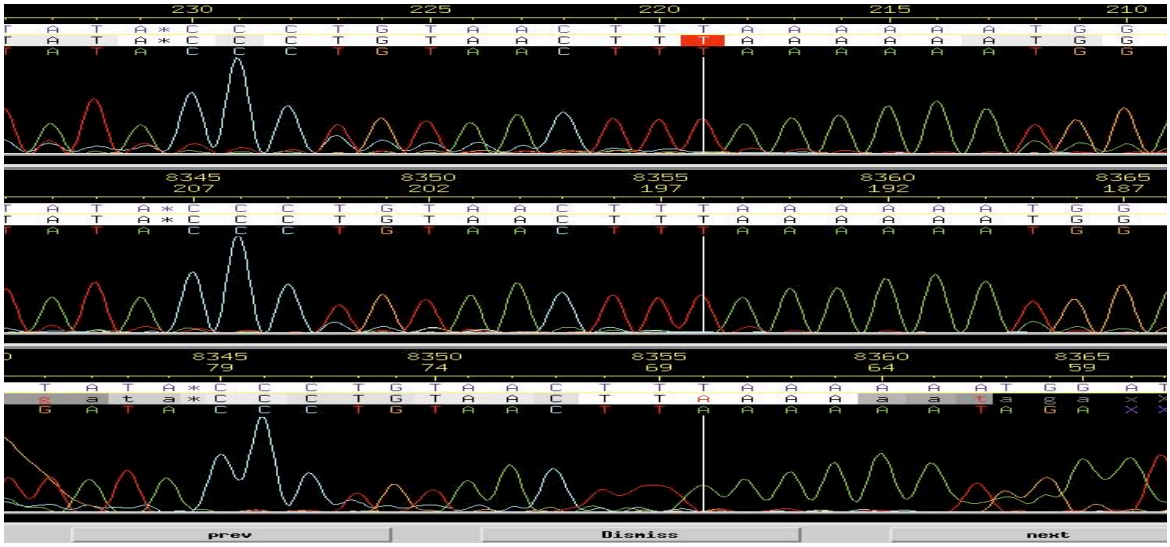


Fig. 8: An uncalled base in the bottom trace was causing a high quality discrepancy. Editing this region corrected the problem.

Despite my best efforts, some regions remained that were of such low quality that making any edits would have been counter-productive. The two remaining regions, bases 9329-9340 and 9382-9388, were tagged as low quality regions during the pre-submit process. Had I been able to run more reactions, eventually one of them would have given me enough data to finish the problematic low quality consensus and have a completely finished sequence.

Fig. 9: Regions of low consensus quality persisted throughout the finishing process. They were tagged and taken note of as I went through the pre-submit checklist.

There were also four regions that were only covered by a single subclone. The problem here lies in the possibility that the sequence data in such regions can be added in the wrong direction. I was reasonably sure that my assembly was correct since the single subclone regions were quite short, but before I was satisfied I needed to see the real and in-silico restriction digest data. The digest data confirms my assembly, and allowed me to simply tag these regions as single subclone regions with Phred scores greater than 30 and move on. In addition to these regions covered by only a single subclone, there were many more regions that were covered by only one strand or one chemistry. Fortunately, these regions all had excellent Phred scores, so I could just tag them as having Phred scores above 30 and continue finishing.

My digest data were excellent. Both the *Eco*RI and *Sac*I digests were flawless.

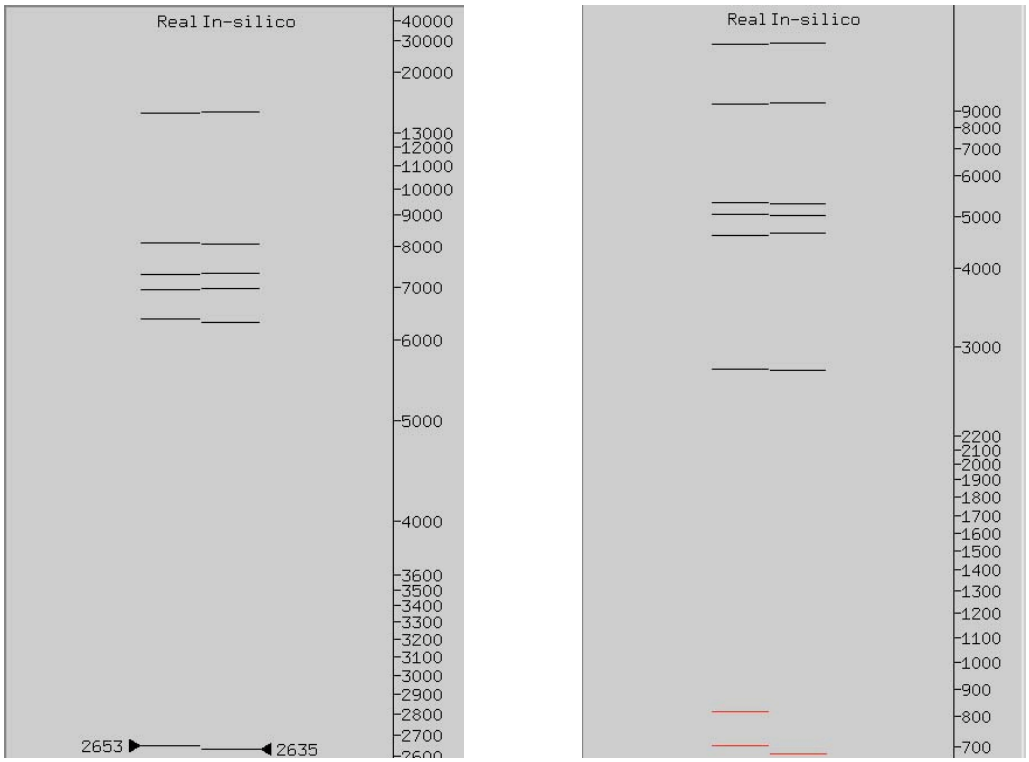
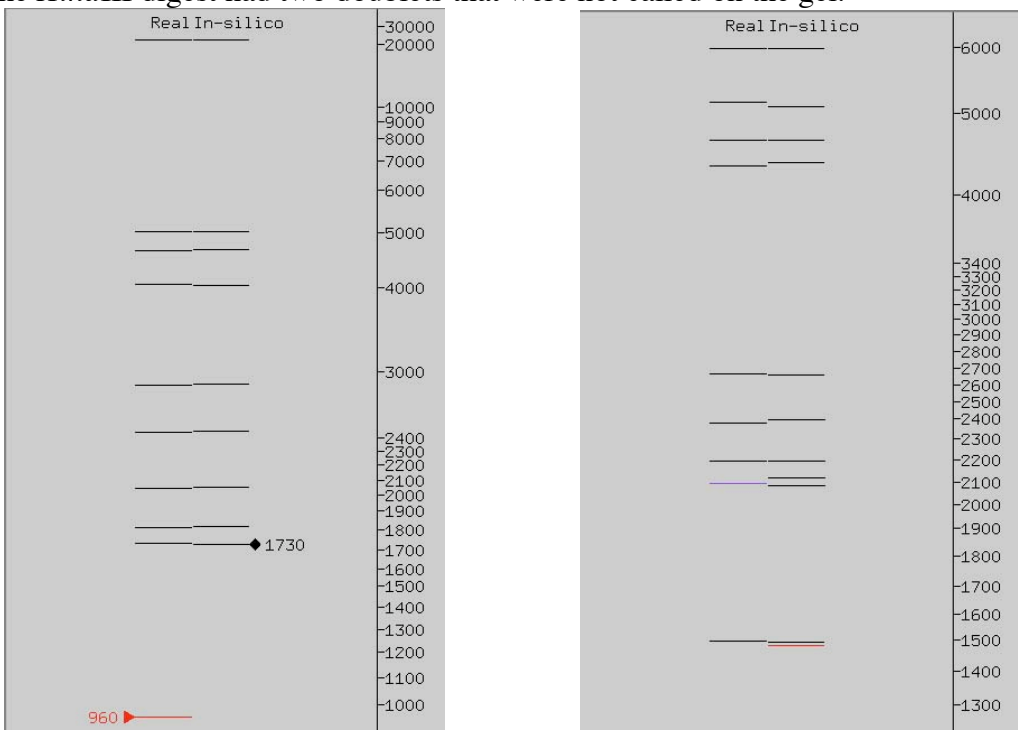


Fig. 10: *Eco*RI digest (left) and *Sac*I digest (right) show that my assembly is correct. The red bands on the *Sac*I digest are smaller than 1 kb, and can be ignored.

The *Eco*RV and *Hind*III digests had minor problems that were easily explained by viewing the actual image of the gel. The *Eco*RV digest had a miscalled band at 960 bp, and the *Hind*III digest had two doublets that were not called on the gel.



Problem	Location
Low consensus quality	9329-9340 9382-9388
Single strand/single chemistry	45-266 1563-1698 9381-9487 11014-11037 16071-16193 17510-17788 18541-18548 21615-21889 22575-22798 23617-23641 25144-25178 25411-25771 27585-27601 28012-28031 28902-28909 29252-29263 32292-32629 33633-33783 33895-34011 34207-34388 35469-35705 38684-39207
Single subclone	32292-32482 33895-34011 34285-34299 35469-35659

Fig. 11: *Eco*RV and *Hind*III digests had minor problems. Both are still good, but not as good as the *Eco*RI and *Sac*I digests.

Once I had confirmed my assembly with my digests, it was time to go through the pre-submit checklist to make sure my finished sequence was of acceptable quality. I knew that I had the two low quality regions left and several single strand/single chemistry/single subclone regions to check to make sure they were of high enough quality to accept. The following table describes the problems remaining in my clone after finishing.

Table 3: remaining problems in the assembly.

The low quality regions were tagged as such, but under normal circumstances they would have been resolved before this step in the process. The single strand/chemistry/subclone regions were all tagged, and since they all had Phred scores higher than 30, they were acceptable.

If it hadn't been for the problems I encountered with large numbers amounts of failed reads, this clone would have been very simple to finish. I had no misassemblies or contigs that had to be force-joined. After the second round of reactions I had one contig, and the rest of my time was spent in the attempt to get additional data. With help I was able to edit the majority of the low quality regions and ended up with only 18 low quality bases in my consensus. Covering the single subclone regions was a goal I was not able to attain; luckily my digest information confirmed my assembly and made the single subclone regions less of a problem than they would have been had my digests turned out poorly. In theory, this was a textbook finishing project, but it revealed some of the more practical problems even a project like this can encounter.