**FINISHING PROJECT:**
**XBAA-16B18**
**(3rd Revision)**

Fine Song
April 11, 2006

**Abstract:**

*My project involved finishing the fosmid XBAA-16B18 from the dot chromosome of* Drosophila virilis. *It part of a larger project of comparing the dot chromosome sequences of* D. virilus *and* D. melanogaster, *with the goal of better understanding the influence of chromatin structure on DNA function. I used the computer program Consed to visualize and analyze the fosmid, which was assembled from a collection of 2-4 kb inserts from a clone library. Since I started out with one contig in my assembly, I did not have any gaps to span. The problems I focused on were low quality regions, high quality discrepancies, and single strand regions. Using the Navigate option, I ordered three rounds of additional reads and performed manual edits to resolve those problem regions. After the third round, the only remaining problems were six single strand regions that are high quality with significant depth in read coverage, giving me a degree of assurance for the quality of the fosmid assembly as a whole. I also checked fosmid integrity by going through a list of quality control measures, including restriction digests, Findid, and checking for mononucleotide runs.*

## Initial Analysis

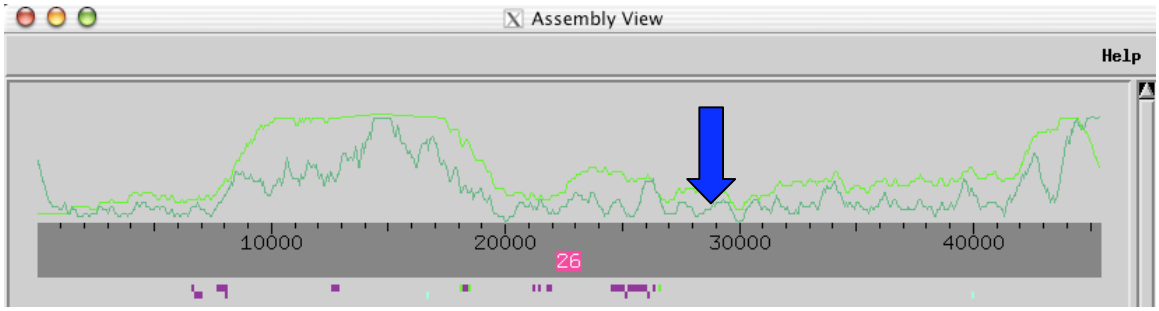


**Fig. 1. Initial Assembly View of my clone.**

This was my initial Assembly View from the production data (Fig. 1). I got the expected fosmid clone length of approximately 45 kb. Fortunately (or perhaps unfortunately in terms of learning purposes), I had no gaps, so the analysis of my clone was all contained within Contig 26. The dark green line, shown by the blue arrow and indicative of *high quality* regions, should ideally be above a certain threshold and somewhat evenly distributed. Instead, my Assembly View showed a huge peak region of *high quality* between bases 10,000 to 20,000. This was a cause of concern since it could be related to misassemblies or high repetition.

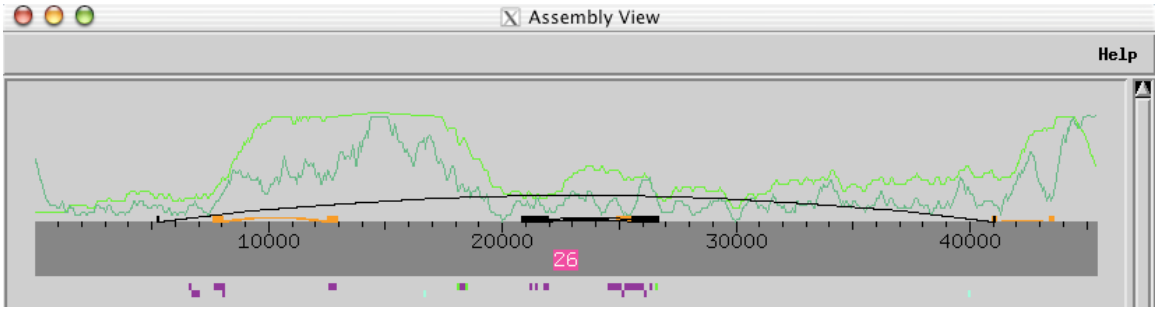


**Fig. 2. Initial Assembly View of my clone after Crossmatch.**

I investigated the possible reasons for this peak. In the initial Assembly View, there were no inconsistent forward/reverse pairs[1], which was a good sign since such pairs can indicate misassembly. I used Crossmatch (Fig. 2) to make sure there were no large repeats. The orange regions represent uncomplemented repeats[2] and the black regions show complemented repeats[3]. Very few such repeats were found through Crossmatch, suggesting that the clone did not have a high frequency of repeats that might lead to misassemblies. A possible reason for the high quality peak may simply be that the peak region was easier to clone relative to the non-peak regions.

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

| Contig Name | Read Name | Consensus Positions | |
|---|---|---|---|
| Contig26 | XBAA-aaf38f06.b1 | 1847-1969 | 123 unaligned high quality |
| Contig26 | (consensus) | 3273-3481 | 210 bp single strand/chem |
| Contig26 | (consensus) | 5292-5319 | 28 bp single strand/chem |
| Contig26 | XBAA-aaf39b02.b1 | 12006 | high quality base disagrees with consensus |
| Contig26 | XBAA-aaf40g02.b1 | 12565 | high quality base disagrees with consensus |
| Contig26 | XBAA-aaf40g02.b1 | 12741 | high quality base disagrees with consensus |
| Contig26 | XBAA-aaf49a07.g1 | 16333 | high quality base disagrees with consensus |
| Contig26 | XBAA-aaf46f03.g1 | 17597 | high quality base disagrees with consensus |
| Contig26 | (consensus) | 19559-20399 | 842 bp single strand/chem |
| Contig26 | (consensus) | 20049 | base quality below threshold |
| Contig26 | (consensus) | 20051-20060 | base quality below threshold |
| Contig26 | (consensus) | 20051-20090 | 40 bp single subclone |
| Contig26 | (consensus) | 20065 | base quality below threshold |
| Contig26 | (consensus) | 20067-20074 | base quality below threshold |
| Contig26 | (consensus) | 20082-20083 | base quality below threshold |
| Contig26 | (consensus) | 22172-22515 | 345 bp single strand/chem |
| Contig26 | (consensus) | 29861-30044 | 191 bp single strand/chem |
| Contig26 | (consensus) | 29951-29953 | base quality below threshold |
| Contig26 | (consensus) | 29970-29971 | base quality below threshold |
| Contig26 | (consensus) | 29977 | base quality below threshold |
| Contig26 | (consensus) | 29979-29983 | base quality below threshold |
| Contig26 | (consensus) | 29987-29992 | base quality below threshold |
| Contig26 | (consensus) | 29994 | base quality below threshold |
| Contig26 | (consensus) | 30000-30003 | base quality below threshold |
| Contig26 | (consensus) | 30017-30023 | base quality below threshold |
| Contig26 | (consensus) | 30027-30030 | base quality below threshold |
| Contig26 | (consensus) | 30034-30035 | base quality below threshold |
| Contig26 | (consensus) | 30037-30046 | base quality below threshold |
| Contig26 | (consensus) | 30040-30044 | 5 bp single subclone |
| Contig26 | (consensus) | 30059 | base quality below threshold |
| Contig26 | (consensus) | 30061-30063 | base quality below threshold |
| Contig26 | (consensus) | 30084-30086 | base quality below threshold |
| Contig26 | (consensus) | 30085-30184 | 100 bp single strand/chem |
| Contig26 | (consensus) | 30085-30167 | 83 bp single subclone |
| Contig26 | (consensus) | 30091-30100 | base quality below threshold |
| Contig26 | (consensus) | 30106-30110 | base quality below threshold |
| Contig26 | (consensus) | 30112 | base quality below threshold |
| Contig26 | (consensus) | 30125-30126 | base quality below threshold |
| Contig26 | (consensus) | 30147-30149 | base quality below threshold |

Go | Prev | Next | Save | Dismiss

**Fig. 3. A sample of problems from initial assembly of my clone.**

[1] Crossmatch's default algorithm filtered out 7 inconsistent forward/reverse pairs as insignificant problems.
[2] Sequences in the consensus sequence that match elsewhere on the same strand
[3] Sequences in the consensus sequence that match elsewhere to the complementary strand

A number of problems came up in the Navigate option (Fig. 3). The most prevalent problem was *low quality* regions around base 20,000 and base 30,000. Low quality regions are below the threshold Phred score of 30 and are shaded in Aligned Reads view (Fig. 4). The few *single subclone* regions that were present seemed to be tied to the *low quality* regions. These, along with *single strand/chemistry* regions, required additional reads.

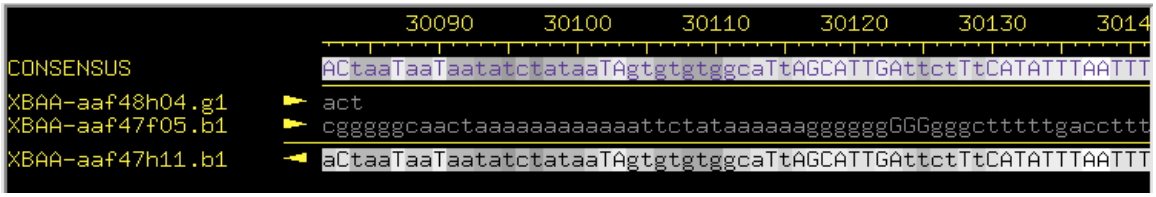


**Fig. 4. Example of a low quality consensus region in the Aligned Reads view.**

In picking oligos for the new reads (Fig. 5), I had to follow certain criteria, including picking primers at least 70 bp away from target regions that had 40-60% GC content and a melting temperature above 50 degrees. Most of the time, the melting temperature was not really an issue, but it was important to make sure that the oligo had enough GC content for good annealing.

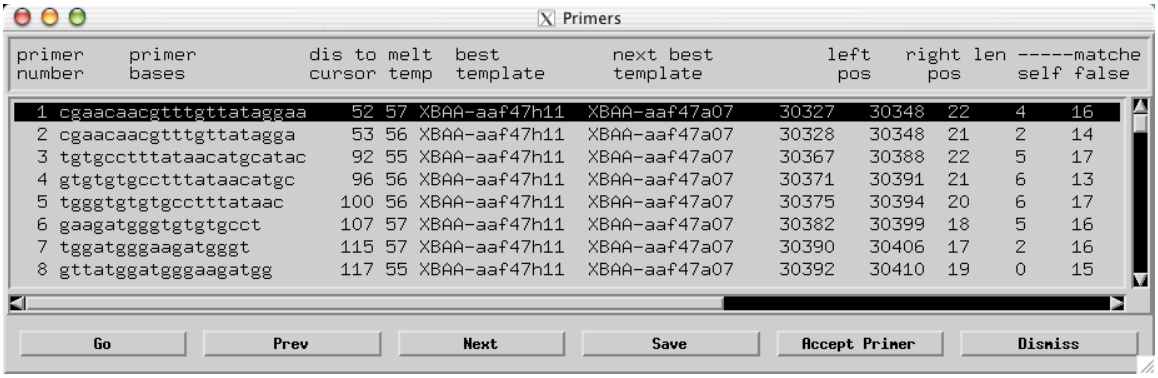


**Fig. 5. Picking primers.**

## Round 1 Analysis:

Table 1 shows the additional reads that I ordered for the first round of reactions. Note that "Success" means that the problem was solved. "Failure" means that a read was made and aligned into the assembly of Contig 26 but did not solve the problem due to *low quality* or an incomplete resolution of the problem (e.g. in the case of *single strand* regions, the additional read covered some of the region but not all). "No Rxn" means that the read was never incorporated into the assembly, likely because the reaction was incomplete or of such low quality that Consed filtered it out altogether.

**Table 1: Round 1 reactions**

| Oligo | Sequence | Dir. | Template | Chem. | Problem | Result |
|---|---|---|---|---|---|---|
| 1 | cgaacaacgtttgttataggaa | <-- | aaf47h11 | All 3 | LQ ~30000 | Success |
| 2 | cctcggatcaaatgcc | --> | aaf47f05 | BD | SS ~30000 | Success |
| 3 | cacgcatataagcatacctatgt | --> | aaf39e03 | All 3 | LQ ~20000 | Failure |
| 4 | cccggattagtccgct | <-- | aaf46a07 | BD | SS ~3400 | Failure |
| 5 | ccatgctacaagataccaagtaaat | --> | aaf41d08 | All 3 | SS ~5300 | Failure |
| 6 | gaatatgggttagaattaatctgg | <-- | aaf40a04 | BD | LQ ~20000 | No Rxn |
| 7 | tgcatttgattgaaatggg | <-- | aaf46h11 | All 3 | SS ~31850 | No Rxn |
| 8 | gagaagccagtaacttatgaatg | --> | aaf39h05 | BD | SS ~38530 | No Rxn |

BD = Big Dye; All 3 = BD + 4:1 + dGTP; LQ = Low Quality; SS = Single Strand/Chemistry

Compared to Autofinish, my list of additional reads was longer, since Autofinish only called for three additional reads (Table 2). The first read suggested by Autofinish, intended to fix the *low quality* region around base 30000, had the same template and a slightly more upstream (by ~140 bp) primer compared to my Oligo 1. The second read called by Autofinish, also intended to fix the same region, had the same template and a slightly downstream primer (by ~220 bp) compared to my Oligo 2. The third read by Autofinish was intended to fix the *low quality* region around base 20000 and had the same template and a slightly upstream (by ~300 bp) primer compared to my Oligo 3. Overall, it seemed that the main problem regions addressed by Autofinish were the *low quality* regions around base 20000 and base 30000. I would have liked to see Autofinish call more additional reads around base 20000.

**Table 2: Autofinish Suggestions**

| Suggestion | Sequence | Dir. | Template | Problem |
|---|---|---|---|---|
| 1 | cacttcttcgaagttggaaa | <-- | aaf47h11 | LQ ~30000 |
| 2 | aaactacctttcgattcgtaataat | --> | aaf47f05 | LQ ~30000 |
| 3 | gagctatatacgtgcgatcttaata | --> | aaf39e03 | LQ ~20000 |

**Interlude – High Quality Discrepancies**

There were also *high quality* discrepancies. These were addressed by going to each site individually and looking through the traces to edit the bases. I did this while I was waiting for the results of round 1, which would produce additional reads.

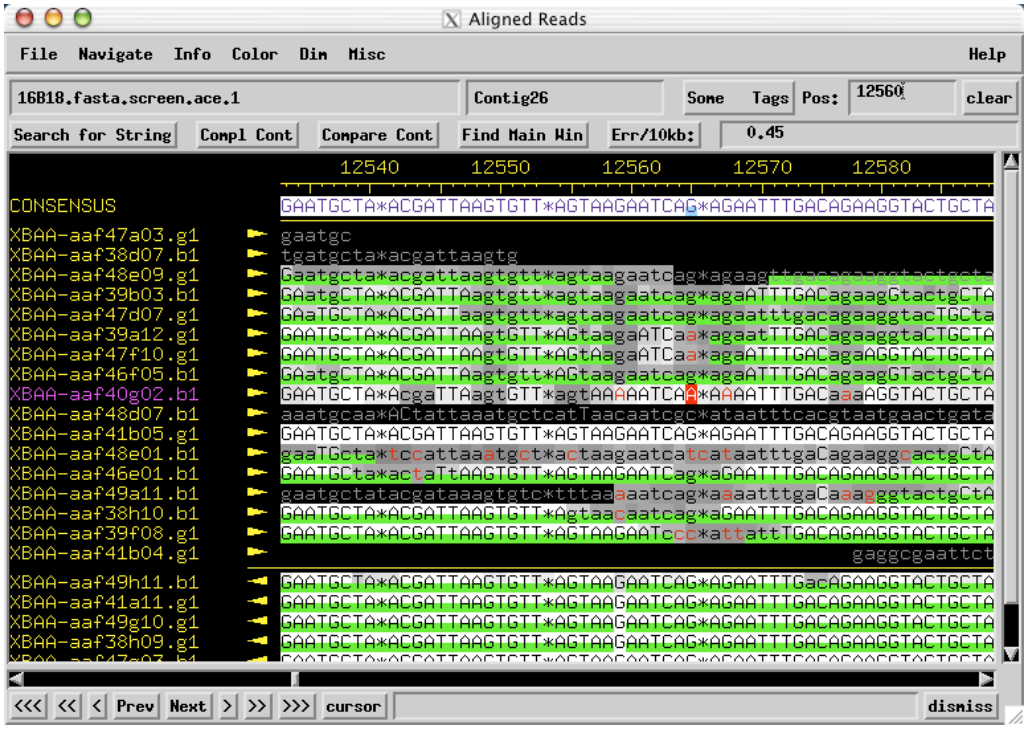


**Fig. 6. Example of high quality discrepancy.**

Fig. 6 shows that the high quality base A in XBAA-aaf40g02.b1 is discrepant with the consensus base of G in position 12565. Upon further investigation using Trace View, I found that there was a G hiding underneath a relatively low A peak, which, given the overwhelming data from other aligned reads, allow me to manually edit the A into a G to match the consensus (Fig. 7).

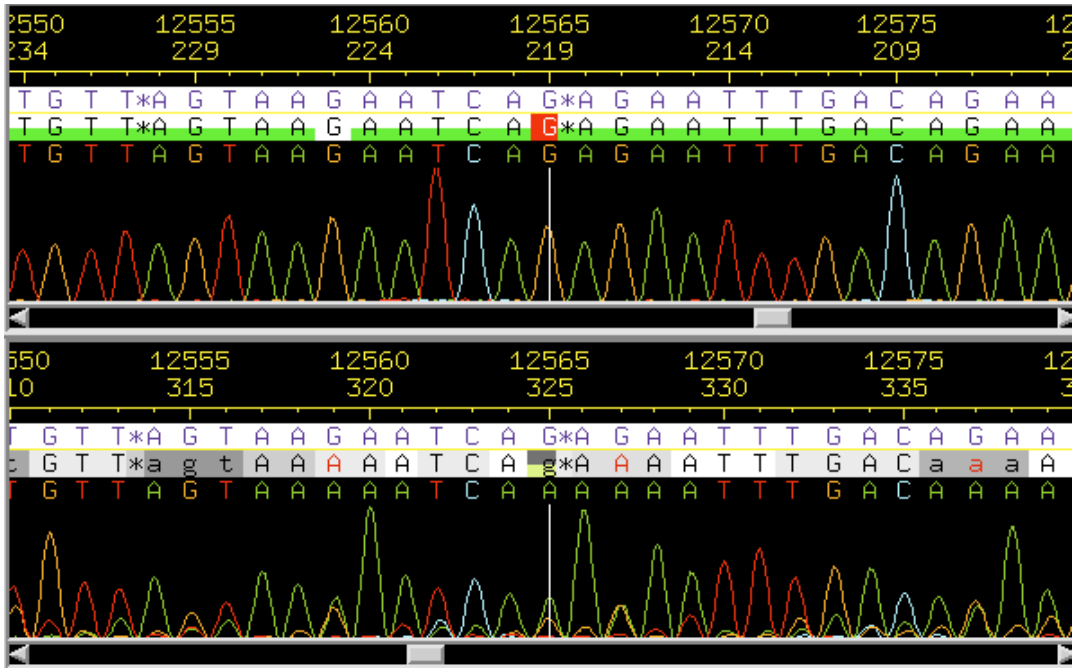


**Fig. 7. Trace view of high quality discrepancy after edit.**

I followed a similar procedure in terms of navigating through *high quality discrepancies* and manually editing them based on traces and other aligned reads. Most of the problems in this category were relatively easy to solve. For problems that could not be resolved because of discrepant bases with supporting traces, I added comment tags and edited the discrepant base into a lower case base to keep the same region from being detected as a continuing problem region by the Navigate option. Usually, comparing the discrepant base with evidence from other aligned reads is convincing enough to choose one base over the other for the consensus (Fig. 8)[4].


**Fig. 8. Example of overwhelming evidence for one base over another based on other aligned reads.**

Other problems I came across included *unaligned high quality discrepancies*. Fig. 9 shows a 123-bp *unaligned high quality discrepancy* on XBAA-aaf38f06.b1. Further downstream, there is a TATA repeat region that seems to be the primary reason for the read's alignment in this particular region of the assembly. There were *high quality* reads going both directions, so it seemed safe to take this read out and attempt to match it to another part of the consensus. In doing so, two other reads associated with this read were taken out too.

[4] One should be careful in using such logic for repeat-rich regions, but the logic works for Fig. 8 since it does not seem to be a repeat-rich region. Additionally, the discrepant A is of lower quality than the C's on other aligned reads at base 44304, giving more credibility to the C's.

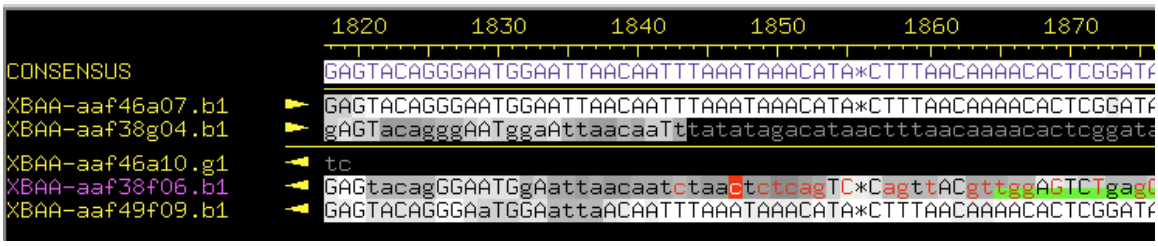


**Fig 9. Example of unaligned high quality discrepancy.**

Fig. 10 shows a segment of the newly made Contig 27. The green highlights indicate potential for the segment to match elsewhere, so I took that sequence and attempted Search for String. Unfortunately, comparing and trying to align this region of Contig 27 turned out to be fruitless, since it did not match Contig 26, the consensus sequence of interest. Even the matches with contigs other than Contig 26, like Contig 24, did not match very well (Fig. 11). I confirmed the irrelevance of Contig 27 to my consensus sequence by running BLAST and finding that Contig 27 is actually a vector sequence (Fig. 11.1).

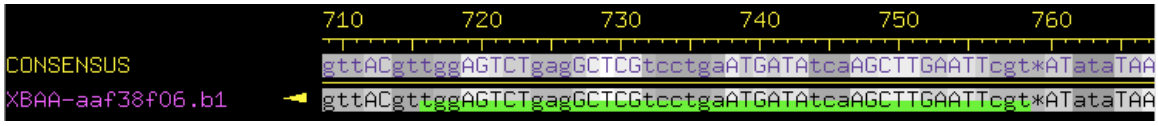


**Fig. 10. Unaligned read taken out and put into its own contig (Contig 27).**

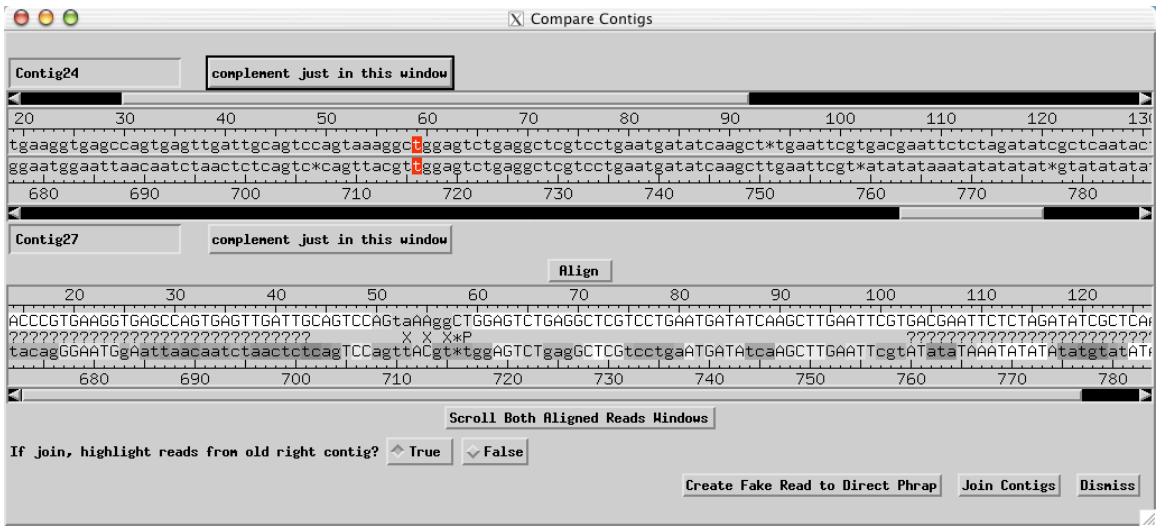


**Fig. 11. Example of failed attempt to align Contig 27.**

| Sequences producing significant alignments: | | Score (Bits) | E Value |
|---|---|---|---|
| gi\|22476919\|gb\|AF532107.1\| | Cloning vector pSMART-HCKan, complete | 105 | 2e-19 |
| gi\|22476916\|gb\|AF532106.1\| | Cloning vector pSMART-LCKan, complete | 105 | 2e-19 |
| gi\|20301816\|gb\|AY090111.1\| | Cloning vector pSMART-LC, complete se | 105 | 2e-19 |
| gi\|16973669\|gb\|AF399742.1\| | Cloning vector pSMART, complete seque | 105 | 2e-19 |
| gi\|60543985\|gb\|AY792409.1\| | Sisymbrium irio clone BAC Si51B6 i... | 69.9 | 1e-08 |
| gi\|60544058\|gb\|AY792482.1\| | Sisymbrium irio clone BAC Si52G1 i... | 60.0 | 1e-05 |
| gi\|23334946\|gb\|AC121591.4\| | Mus musculus BAC clone RP23-289H3 fro | 52.0 | 0.003 |

**Fig. 11.1. BLAST confirmation of Contig 27's irrelevance to finishing Contig 26.**

## Back to Round 1 Analysis

Upon adding new reads using PhredPhrap, some regions improved significantly. One such region was the *low quality* region around base 30000. In Fig. 12, the new reads, highlighted in purple, provided a significant boost in base quality for that region. This can be visualized in Assembly View in Figs. 13 and 14.

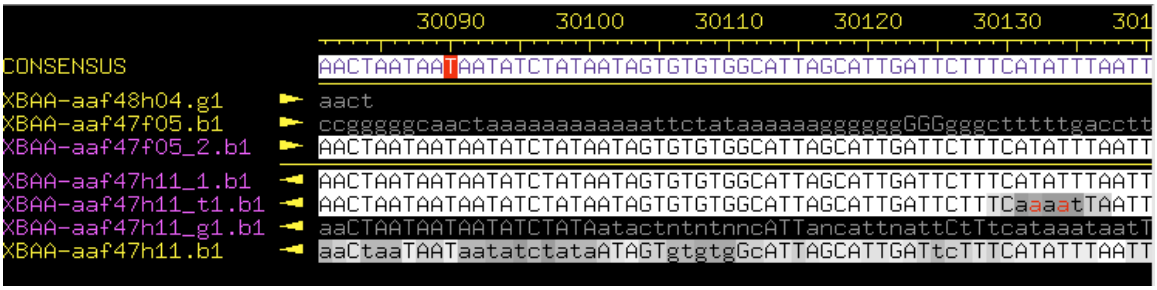


**Fig. 12. Low quality region around 30000 with added reads.**

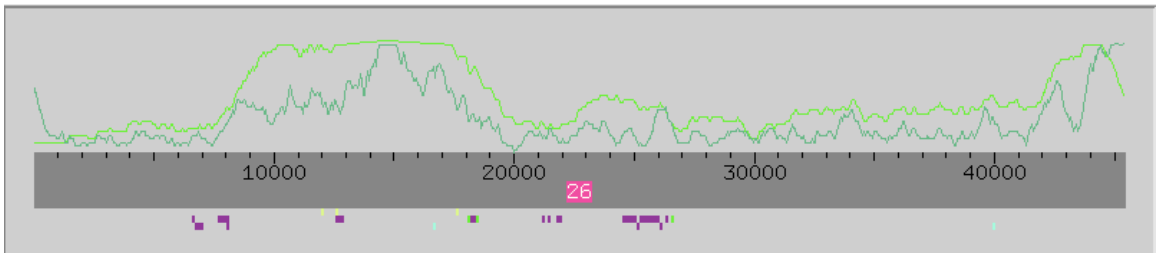


**Fig. 13. Assembly view after round 1 of new reads.**

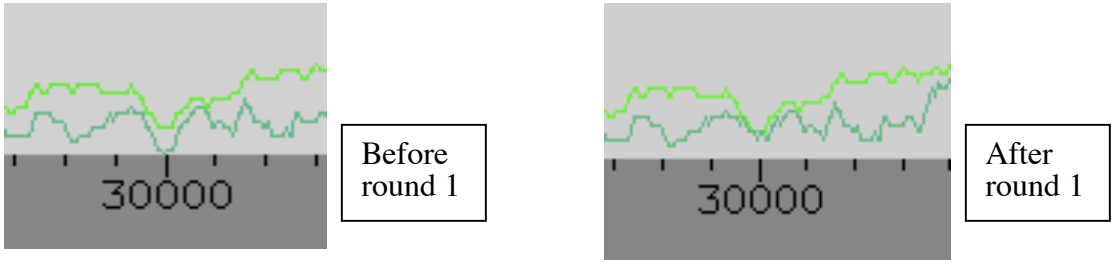


**Fig. 14. Close-up Assembly View of region around 30000.**

Unfortunately, the other *low quality* region around base 20000 did not get resolved, nor did most of the *single strand/chemistry* regions (Fig. 15). In fact, some of the reads that were called ended up creating new problems, like a XBAA-aaf39e03_3.b1, a 39 bp *unaligned high quality* segment from a BD reaction around base 20000 (Fig. 15).

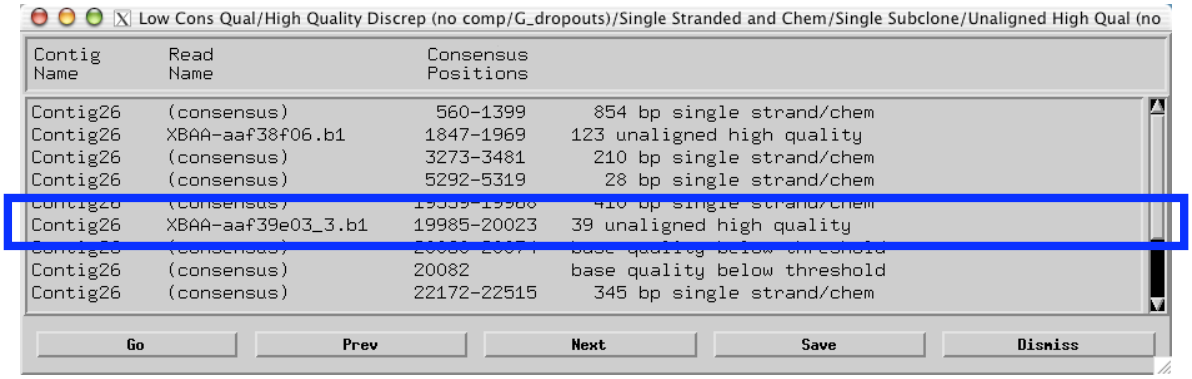


**Fig. 15. Sample of remaining problems after round 1 of additional reads and edits.**

## Round 2 Analysis

To attack the unresolved problems from round one, I called a second round of additional reads (Table 3). For the majority of the additional reads, I used different primer pairs (Oligos 9-17), most of which were targeted towards resolving the *low quality* region around 20000 and various *single strand/chemistry* regions. Three of the reads were variants from the previous round of additional reads. I switched templates for Oligo 3 and 5. For Oligo 6, I used the same template but used 4:1 chemistry instead of BD.

All this was done in an attempt to make a *high quality consensus* sequence with minimal problem regions. Typically the costs involved with these reactions would encourage me to be more judicious in the type and number of reactions used. However, since this is a well-funded, small-scale project with time constraints, it was possible to call as many reads as I did without too much reservation.

**Table 3: Round 2 reactions**

| Oligo | Sequence | Dir. | Template | Chem. | Problem | Result |
|---|---|---|---|---|---|---|
| 3 | cacgcatataagcatacctatgt | --> | aaf46h03 | 4:1 | LQ ~20000 | No Rxn |
| 5 | ccatgctacaagataccaagtaaat | --> | aaf48d06 | 4:1 | SS ~5300 | No Rxn |
| 6 | gaatatgggttagaattaatctgg | <-- | aaf40a04 | 4:1 | LQ ~20000 | Failure |
| 8 | gagaagccagtaacttatgaatg | --> | aaf39h05 | 4:1 | SS ~38530 | No Rxn |
| 9 | cttacaaatgtgtgtaaccactct | <-- | aaf40a04 | 4:1 | LQ ~20000 | Failure |
| 10 | ctttgcttggcttggaa | --> | aaf39e03 | 4:1 | LQ ~20000 | No Rxn |
| 11 | gcttatacagataaatggccttaca | --> | aaf41d09 | 4:1 | SS ~900 | No Rxn |
| 12 | ctgggcacactgttcttact | --> | aaf41d09 | 4:1 | SS ~900 | No Rxn |
| 13 | catcaggagcacttcgg | <-- | aaf46a07 | 4:1 | SS ~3340 | No Rxn |
| 14 | ccctacaactttgcagatatg | --> | aaf39e03 | 4:1 | LQ ~20000 | No Rxn |
| 15 | gcagtcgccattgctc | <-- | aaf41c02 | 4:1 | SS ~22300 | Failure |
| 16 | cacttcccaaaagttcacaa | <-- | aaf46h11 | 4:1 | SS ~31850 | No Rxn |
| 17 | ccgccattggcatatt | <-- | aaf48h05 | 4:1 | SS ~43305 | Failure |

BD = Big Dye; All 3 = BD + 4:1 + dGTP; LQ = Low Quality; SS = Single Strand/Chemistry

As you can see in Table 3, the second round of additional reactions did not have much success. Most of the read to cover *single strand* regions yielded no reaction, perhaps due to the problem region having either a bad oligo or template, e.g. something went awry in the process of producing them. Very little changed in the Assembly View with the new reads, or in the number and types of problems remaining (Figs. 16 and 17).

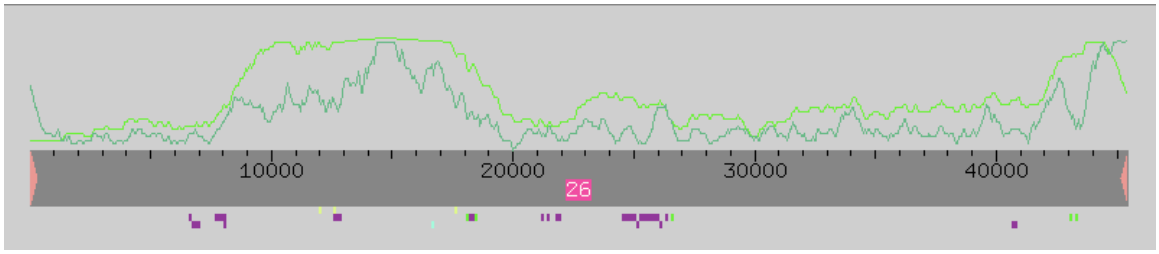


**Fig. 16. Assembly view after round 2 of new reads.**

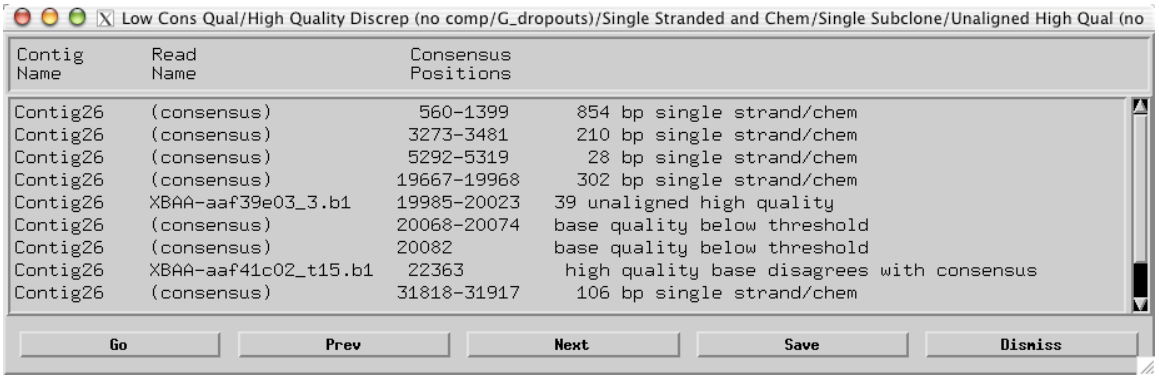


**Fig. 17. Sample of remaining problems after round 2 of additional reads and edits.**

## Round 3 Analysis

At this point, a finisher recommended not requesting any more reads for *single strand* regions. So I made one last attempt at fixing the *low quality* region around base 20000 in a third round of additional reactions, using the same oligos from previous rounds and simply changing the templates (Table 3).

**Table 3: Round 3 Reactions**

| Oligo | Sequence | Dir. | Template | Chem. | Problem | Result |
|---|---|---|---|---|---|---|
| 6 | gaatatgggttagaattaatctgg | <-- | aaf47d06 | 4:1 | LQ ~20000 | No Rxn |
| 9 | cttacaaatgtgtgtaaccactct | <-- | aaf47d06 | 4:1 | LQ ~20000 | Success |
| 10 | ctttgcttggcttggaa | --> | aaf46h03 | 4:1 | LQ ~20000 | No Rxn |
| 14 | ccctacaactttgcagatatg | --> | aaf46h03 | 4:1 | LQ ~20000 | Failure |

BD = Big Dye; All 3 = BD + 4:1 + dGTP; LQ = Low Quality; SS = Single Strand/Chemistry

After three tries, the low quality region around 20000 was finally eliminated from the list of problems (Fig. 18). The Assembly View does not show this change very well, until you zoom in and compare the initial Assembly View to the final Assembly View of that region (Figs. 29-21).

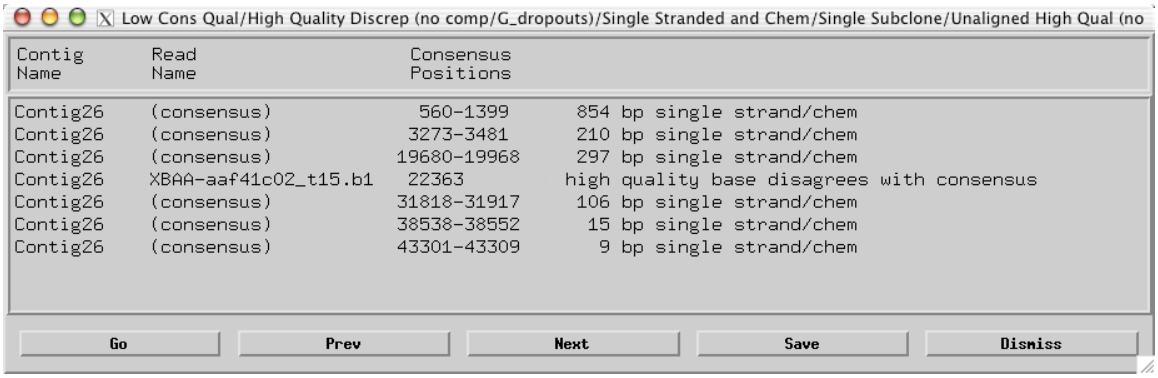


**Fig. 18. Remaining problems after round 3 of additional reads and edits.**

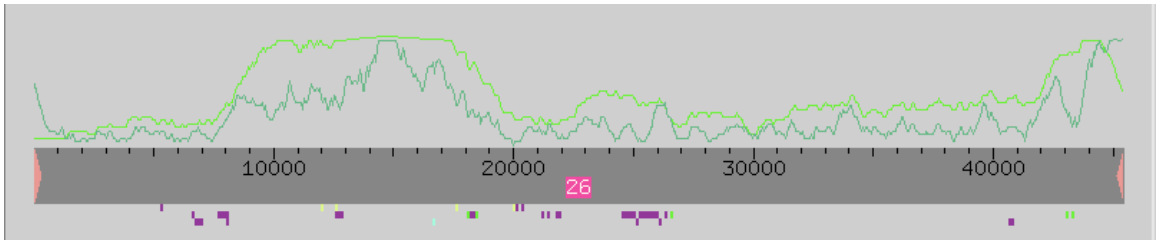


**Fig. 19. Assembly view after round 3 of new reads (final assembly).**

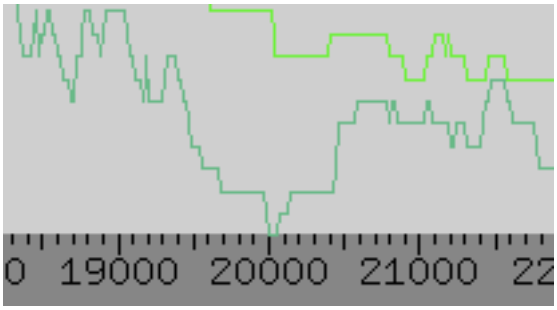


**Fig. 20. 3X zoom in Assembly View of initial data.**

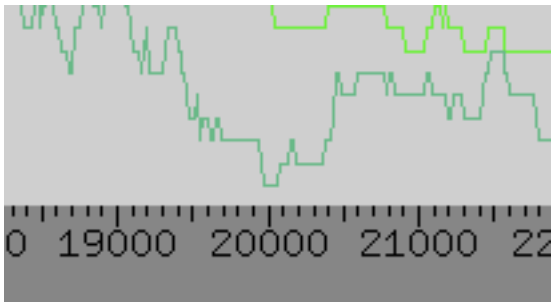


**Fig. 21. 3X zoom in Assembly View after round 3 (final).**

The *high quality* discrepancy at base 22363 was easily solved by looking at the trace of the sequence and manually editing it from a T to an A on the basis of the trace data (Fig. 22).

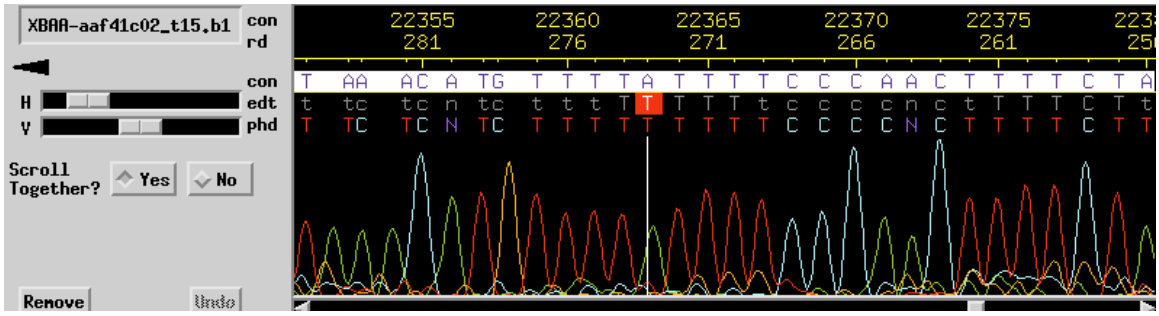


**Fig. 22. High quality discrepancy at base 22363.**

## Checking Fosmid Integrity

As part of the presubmit checklist, I looked for mononucleotide runs (> 15 A's, > 15 C's) and found only two such runs. For example, Fig. 23 shows 32 A's in a row, with significant depth in coverage. The trace showed strong A signals, giving greater reassurance that the mononucleotides are not a major concern (Fig. 24). Three of the reads were a bit ambiguous at base 12303 (blue arrow), but the depth of coverage for the base with other reads and clear traces associated with these reads more than compensated for any doubts. The other run had 16 A's in a row; this region did not show as much depth of coverage but, again, had clear traces. Fortunately, the consensus of Contig 26 did not have any X's or N's, which would have indicated vector sequence and ambiguity, respectively.

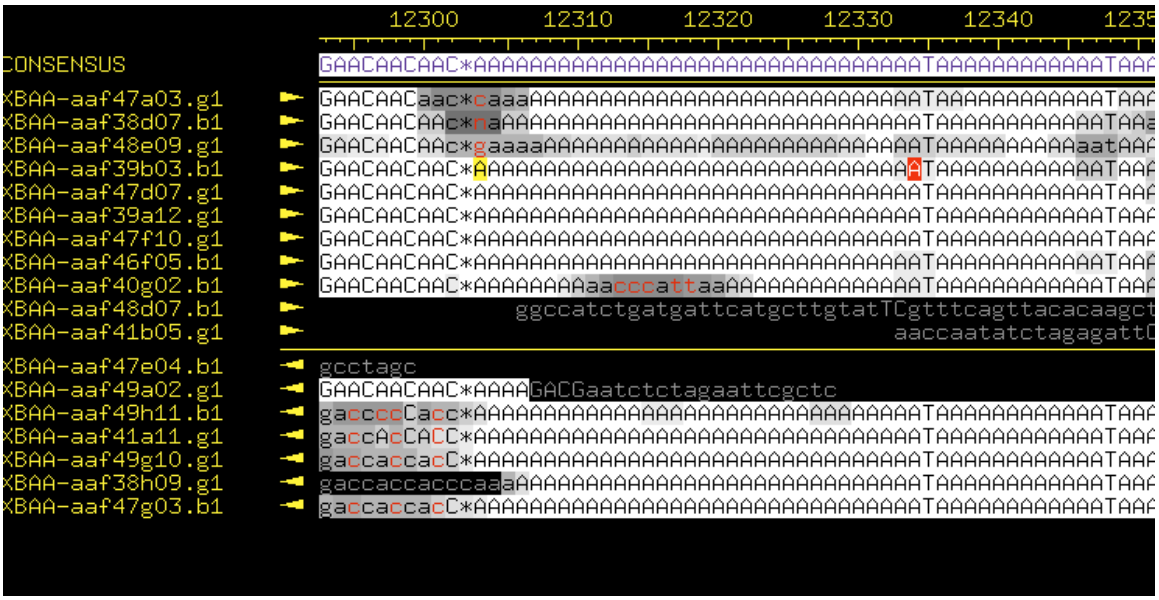


**Fig. 23. Mononucleotide run of 32 A's.**

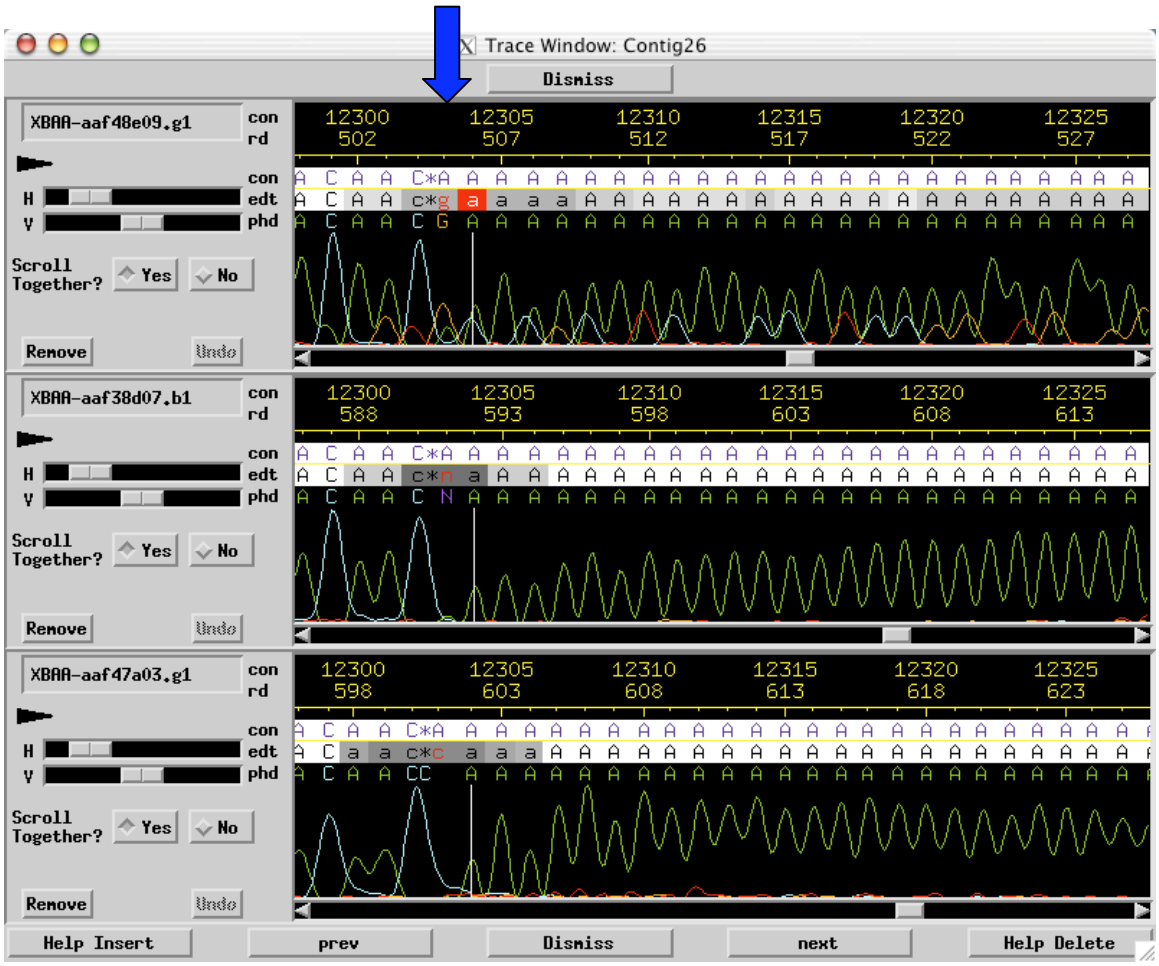


**Fig. 24. Trace view of mononucleotide run of 32 A's.**

As a further check on the quality of my finishing, I ran several *in silico* digests. Digests are a measure to ensure the correct assembly of the finishing project and can be useful indicators of misassemblies or problematic regions. The *Eco*RV digest showed that the real and *in silico* digests matched well, with a possible double band around 2100 in the *in silico* case (Fig. 25). The *Hin*dIII digest matched fairly well except for a mysterious band that was present in the real digest around 2250 but not present *in silico* (Fig. 26). Checking the gel picture of the real *Hin*dIII digest revealed that the *Hin*dIII mystery band in Display Digests could be ignored because it was lighter than the bands around it and did not fit into the gradient expected for a real band. Based on the evidence, everything seems consistent and there is little reason to suspect misassemblies or major problem regions.

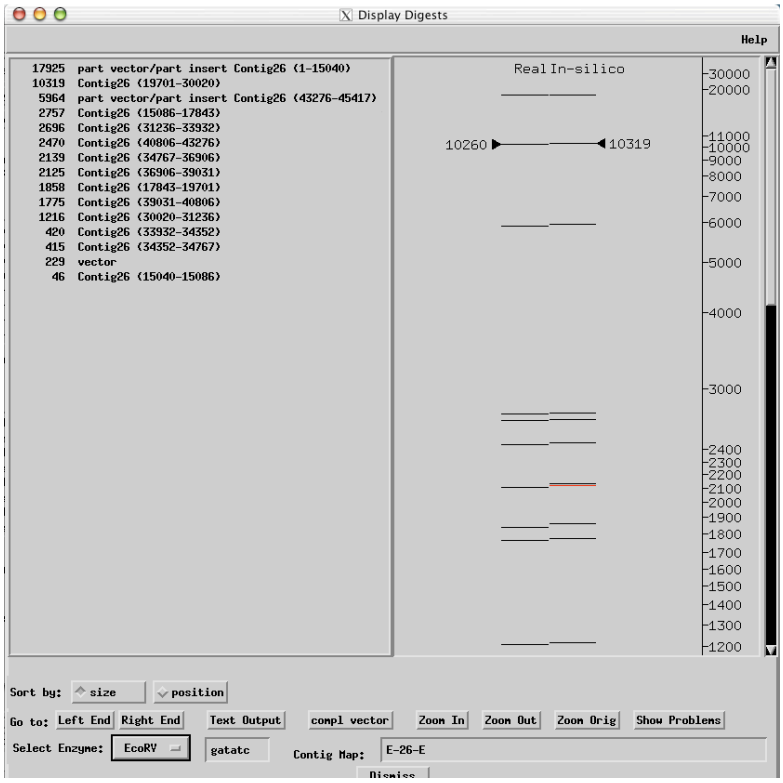


**Figure 25. EcoRV digest of my fosmid.**

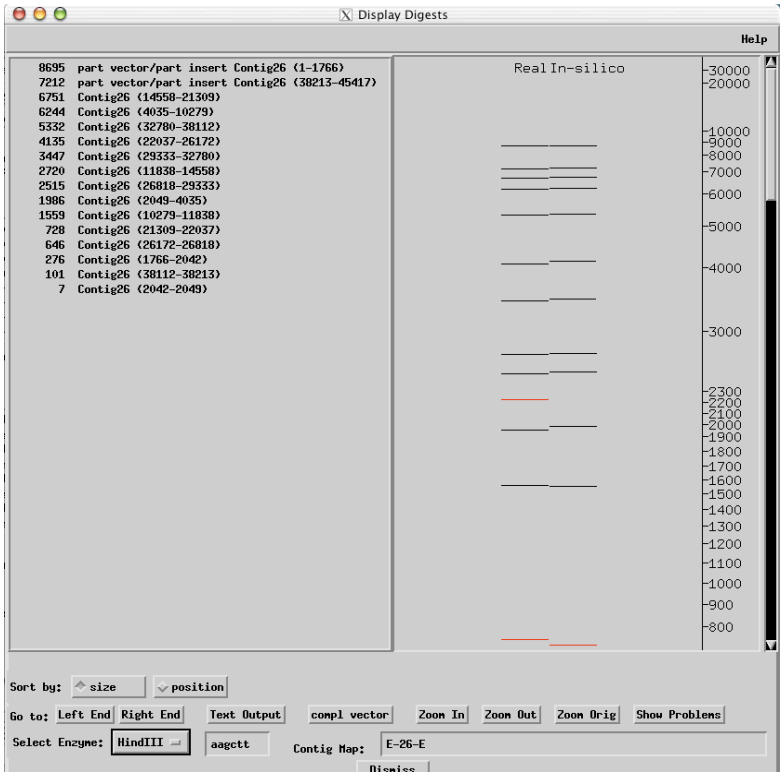


**Fig. 26. HindIII digest of my fosmid.**

Other tasks I completed for the presubmit checklist included identifying and tagging the cloning ends (GATC). I also received Findid data, which did not detect significant vector or bacterial DNA associated with my consensus, and helped confirm the integrity of the assembly. I also checked for four well-known vector sequences (GAATTCGTC-insert; GAATCGTT-insert; insert-GACGAATTC; and insert-AACGAATTC) and did not get any matches for the assembly. I did not run BLAST other than during Round 1 for a read that was pulled out (Contig 27), because overall there were no indications of significant contamination from a vector or host. I did not need to steal data or use fake reads. There were no contigs over 2 kb that were not in the assembly.

In the end, I achieved the goal of having all sequences in one contig with Phred consistently >30. The only remaining problems of my finishing project are six *single strand* regions. For each of these *single strand* regions, there are at least two *high quality* reads in one direction and some have up to five *high quality* reads in a given region, even though there is nothing in the other direction (Fig. 27). Though not optimal, the six *single strand* regions can be taken to be reliable based on the *high quality* of the reads in at least one direction.


**Fig. 27. An example of a remaining single strand region**

In conclusion, the various problems in finishing XBAA-16B18 were resolved for the most part, including the *low quality* regions and *high quality discrepancies*. Onward to annotation!