Finishing Project 420-O08
Olivia Knowles
29 February 2008

Abstract:
The goal of this project is to finish fosmid 420-O08 from the dot chromosome of *Drosophila mojavensis*. This project is one of many others like it; when combined they will produce a finished version of the dot chromosome of *D. mojavensis*. I used Consed, a computer program, to analyze and finish the fosmid. Another program called PhredPhrap assembled the fosmid from a collection of sequences from 2-4 kb subclones. The problems to be resolved were issues of high quality discrepancies, single subclone regions, single chemistry regions, and two gaps. The *navigate* function and Assembly View were used to find these problem regions. At the end, all of these problems seem to have been resolved, but there was no accurate restriction digest available to confirm the accuracy of the assembly. While all bases in the consensus sequence are at or above Phred 30, confidence in the correctness of the assembly is not as high as desired because of the lack of a restriction digest.

Initial Analysis:

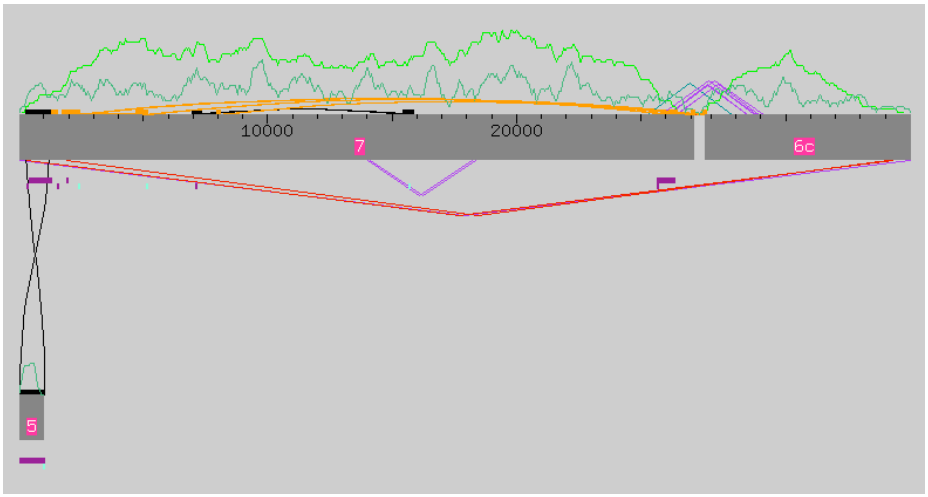

Fig. 1. Initial Assembly after running *crossmatch*

The initial assembly (Fig. 1) was composed of three significant contigs: 5, 6, and 7. The other 4 contigs present only contained a single read. There were clones that spanned the gap between 7 and 6, and Crossmatch revealed that contig5 probably belongs somewhere in contig7, as their complimentary sequences matched one another. The initial problems found by the navigation tool were mostly high quality discrepancies with a few low quality and single strand / chemistry areas (Fig. 2). The same was true for contig5 and contig6.

Contig Name	Read Name	Consensus Positions	
Contig7	(consensus)	1-4	4 bp single subclone
Contig7	(consensus)	1-142	142 bp single strand/chem
Contig7	(consensus)	1-18	base quality below threshold
Contig7	03808875O22.b1	894	high quality base disagrees with consensus
Contig7	33839411M19.g1	985	high quality base disagrees with consensus
Contig7	09332075N11.b1	6315	high quality base disagrees with consensus
Contig7	03771075F11.b1	13126	high quality base disagrees with consensus
Contig7	04091975F09.b1	15816	high quality base disagrees with consensus
Contig7	09506175P10.b1	19468	high quality base disagrees with consensus
Contig7	04169875L20.b1	20014	high quality base disagrees with consensus
Contig7	09313975B22.b1	23073	high quality base disagrees with consensus
Contig7	(consensus)	25280-25299	20 bp single strand/chem
Contig7	(consensus)	27283-27363	81 bp single strand/chem
Contig7	(consensus)	27283-27363	81 bp single subclone
Contig7	(consensus)	27342-27343	base quality below threshold
Contig7	(consensus)	27347-27363	base quality below threshold

Fig. 2 Initial problem areas on contig7

Noticing that contig6 was complimented in Assembly View, complimenting contig6 in the aligned reads window was the first change made. Now all of the contigs are running in the same direction. Next, I searched for a similar string between contigs 6 and 7 in order to create a force join to close the gap. This was not initially successful. I tried to incorporate some of the single read contigs into the assembly, hoping to receive more data in the gap region. Of the four single read contigs, only contig1 joined at the right end of contig7 (which became contig8). Having elongated the contig, I tried another search for sequence between contig8 and contig6. At that point I was able to create a force join between contig8 and 6, creating the new contig9 (Fig. 3). I was then able to tag the ends of the fosmid.

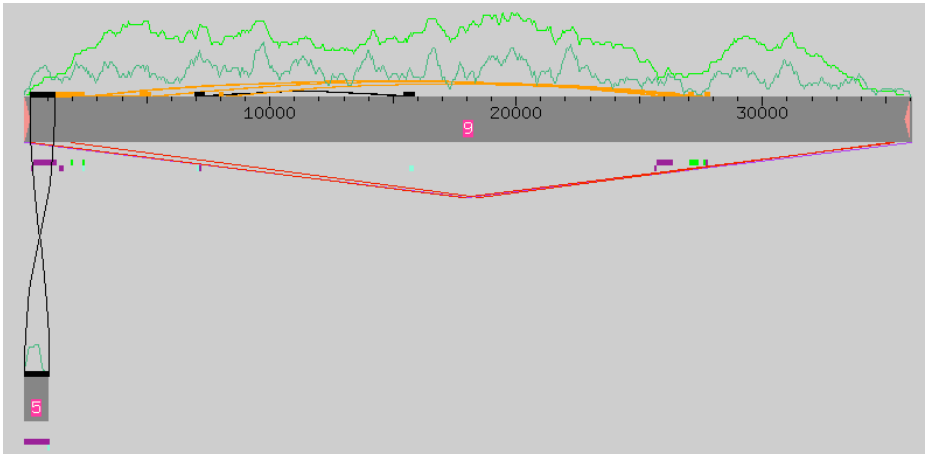

Fig. 3. Assembly view after joining the gap

Having closed the gap I began working on joining contig5 with contig9. Since Consed showed an inverted repeat in contig5 matching sequence in contig9, I complimented 5,

completed a *search for string*, and aligned 5 and 9. They aligned well with only some low quality discrepancies at the ends of the alignment, so I decided that it would be a good idea to join the two contigs to create contig10.

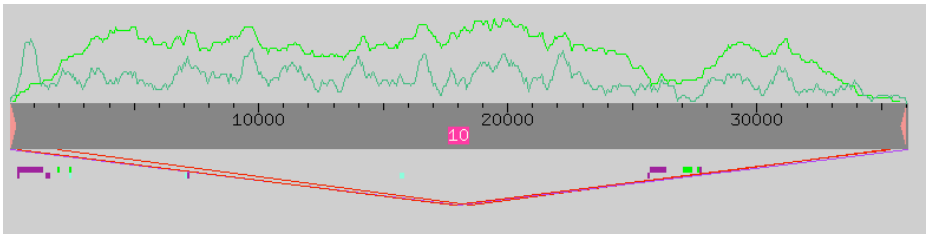

Fig. 4 Assembly view after adding the reads from contig5 to contig9

After this I attempted to add the other two single read contigs to contig10, but they were not aligning well. The reads were both low quality and would not have contributed very much to the consensus quality, so I decided that it was not necessary to incorporate them.

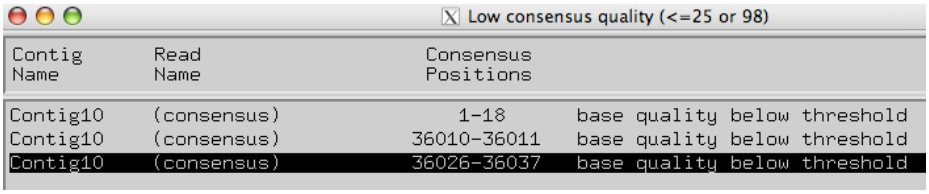

Fig. 5 The list of low consensus quality areas in contig10.

After having created the single contig, I began to resolve the low quality issues. The only low quality issues that existed were at the very beginning and very end of the fosmid. It would not be worth sequencing these areas again because other projects that are being completed simultaneously will overlap these regions, thus increasing the Phred quality.

First round ordering primers:

In the first round of ordering primers, I had already closed my gap and I did not have any low quality areas that I wanted to sequence in order to increase the phred score. I called two primers (one in each direction) in all three chemistries (BigDye, dgtp, and chem. 4:1) to confirm the gap region.

Of the 6 reactions, only one of them successfully incorporated into contig10. Nonetheless, this one successful reaction did insert over some of the gap, increasing my confidence that the force join was correctly made. The highlighted read in Fig. 7 shows the beginning of the newly incorporated read.

Standard Reactions:

Rx #	Oligo ID	Oligo Sequence	New Oligo?	Chemistry	Reaction Name
1	420-O08.9	tgatatgcgcgctgtaat	Yes	bigdye	XBAA-420-O08_9.b1
2	420-O08.9	tgatatgcgcgctgtaat	Yes	dgtp	XBAA-420-O08_g9.b1
3	420-O08.9	tgatatgcgcgctgtaat	Yes	chem41	XBAA-420-O08_t9.b1
4	420-O08.10	tgccggtagccgtaca	Yes	bigdye	XBAA-420-O08_10.b1
5	420-O08.10	tgccggtagccgtaca	Yes	dgtp	XBAA-420-O08_g10.b1
6	420-O08.10	tgccggtagccgtaca	Yes	chem41	XBAA-420-O08_t10.b1

Your order has been successfully processed on Feb-06-2008
Please keep a copy of this page for your record.

Fig. 6. First round of Primers

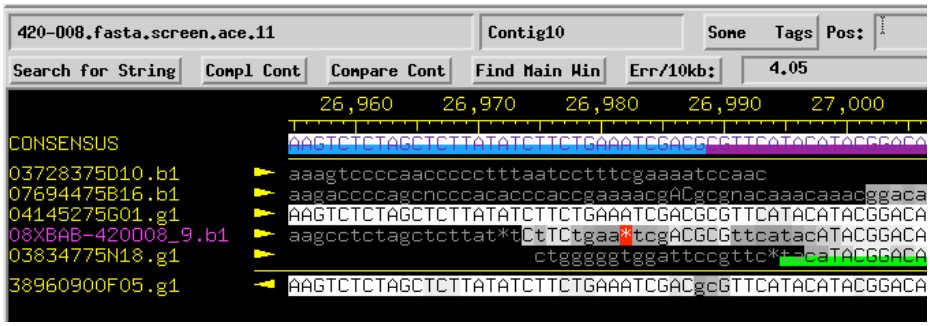

Fig. 7 Incorporated read from Round One primers.

When I ran Autofinish at this point, it decided no further sequencing reactions were necessary. I only called for two primers to check the join; these were not necessary to close the gap. Thus, Autofinish and I both made roughly the same decision. This was not a very interesting comparison, so I ran Autofinish on my first ace file to see what primers Autofinish would have called at the beginning of the analysis. Fig. 8 shows the primer list that was generated.

```
autofinish output file 420-008.080129.190019.nav

Contig5     (consensus)          1-91       Contig5     -849     91
Contig5     (consensus)        431-1009     Contig5      431   1371
Contig5     (consensus)        899-1009     Contig5      899   1839
Contig6     (consensus)          1-81       Contig6     -859     81
Contig6     (consensus)       7713-8477     Contig6     7713   8653
Contig6     (consensus)       8389-8477     Contig6     8389   9329
Contig7     (consensus)          1-88       Contig7     -852     88
Contig7     (consensus)      26635-27363    Contig7    26635  27575
Contig7     (consensus)      27275-27363    Contig7    27275  28215
```

Fig. 8 Autofinish's list of primers to call to resolve the initial assembly (file ace.1)

Autofinish would have called primers…

- at the beginning of contig7, which would not have been very helpful since it is the beginning of the fosmid
- at the beginning and end of contig7 and 6 in efforts to close the gap. This is unnecessary because I was able to force join the gap with sequence that was already available.
- at the beginning, middle and end of contig 5. Perhaps it did this because the sequence was similar to sequence in other areas and it wanted to make sure that the sequence was really correct where it was. The problem with this decision is that there was already over 8x coverage in most of contig5 and it was full of repeats (Fig. 9). I don't think that sequencing over this contig would yield anything more informative.

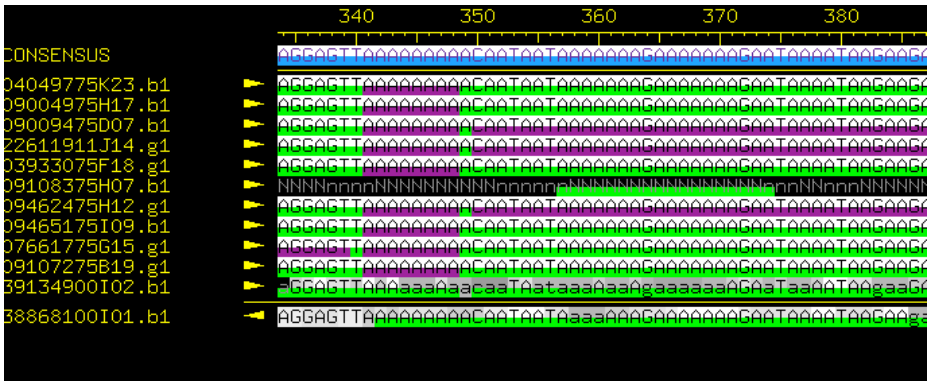

Fig. 9 Example of the repeats in contig5

At this point, since my fosmid was in one contig and my new reads had been incorporated, I decided to check my *in silico* digest against the real *EcoR*I digest. The results that I received did not give me any reliable information. The bands in the real digests added up to a 70kb piece of DNA, but we know that the fosmid is only around 40kb in size. It is possible that the enzyme didn't completely digest all of the DNA, creating more bands than there should have been. Also, the DNA sample used for the digest could have been contaminated. The *Hind*III restriction enzyme digest had the same problems as the *EcoR*I digest and the *EcoR*V and *Sac*I digests were unavailable. This further supports the hypothesis that the DNA sample was contaminated.

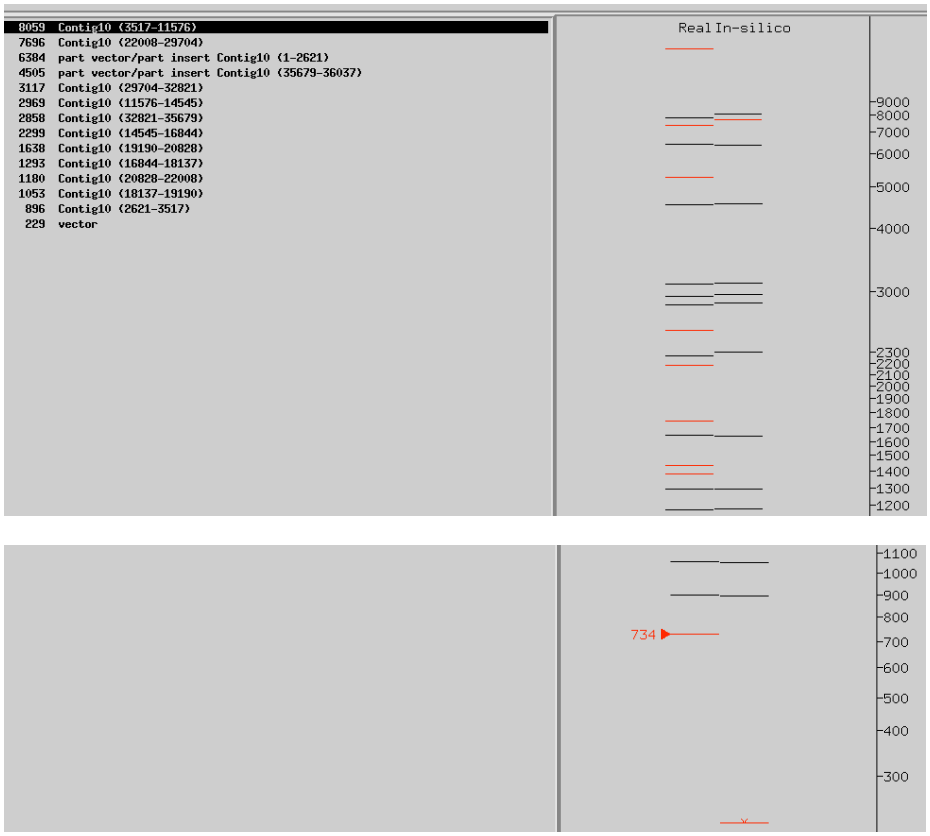

Fig. 10 Real vs. *in silico* EcoRI digest

Since I could not use the restriction digests for any information, I turned my attention to resolving the high quality discrepancies. The high quality discrepancy shown in Fig. 11 is located at the left end of where I force joined what used to be contig5 with contig9. At first I thought that this was just an area where contig9 was missing information and contig5 had extra information, so I was planning on changing the consensus. Then I saw that there was a pattern to the discrepancies. The same discrepant sequence existed on one strand where there were pads on another strand. This led me to believe that I had incorrectly positioned the force join, and that it needed to be moved over by the number of bases that were discrepant. The results are shown in Fig. 12.

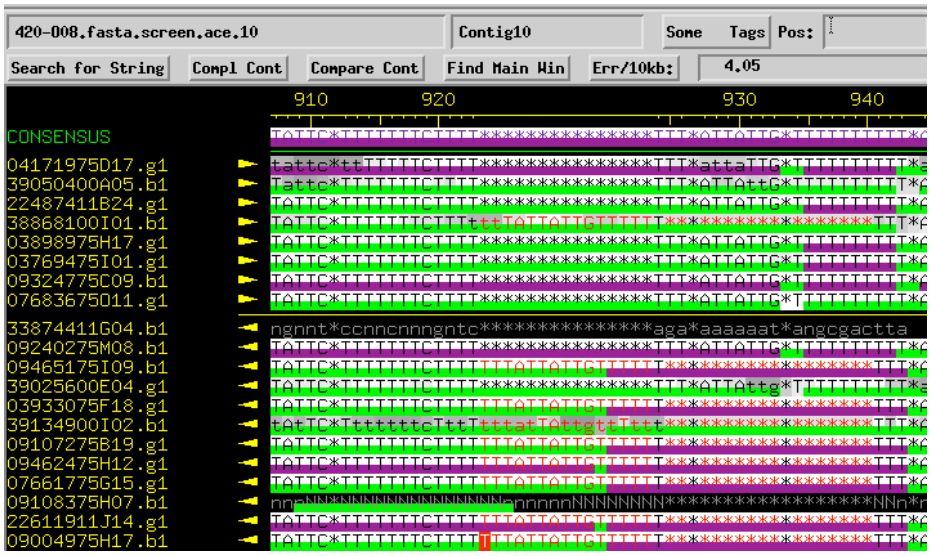

Fig. 11 Largest high quality discrepancy area

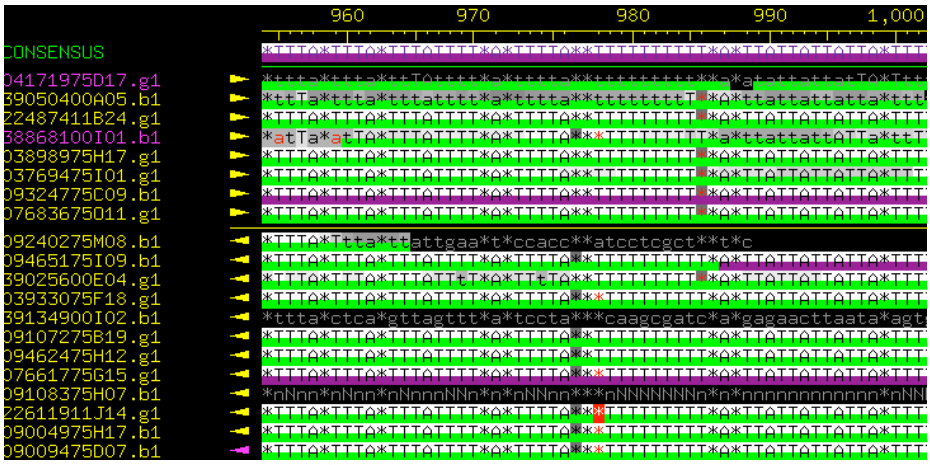

Fig. 12 Results after tearing and rejoining part of contig 10.

It seemed as though I had missed the entire sequence by one base. I tried to tear and rejoin again but I kept getting this discrepancy. I decided to take out all of these discrepant reads because I had enough other reads to still have well over phred quality 30.

After tearing out all of the reads, I considered the possibility that I might have had a polymorphism in that region. I added the reads back in and found that of the 17 high quality reads that I had, four of them had 10 Ts and the other thirteen had 9 Ts. The extra Ts were tagged as polymorphisms and the consensus was changed to reflect the 10T length instead of the 9T length.

It was easier to discover the problems that caused the other high quality discrepancies in this project. Most of them were simply miscalls by PhredPhrap. A good example of this was a high quality discrepancy between the consensus G and a T in one of the reads. This T was in an area where many other Ts had been called. The trace showed that there was a peak for T but there was a higher overlapping peak for G; thus I believe this is a miscall and the base should have been called as a G. I manually made this change. See Figs. 13 and 14.

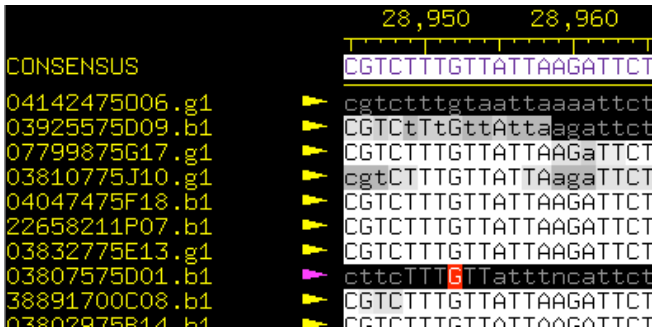

Fig. 13 Aligned Reads view of the discrepant base after the change was made.

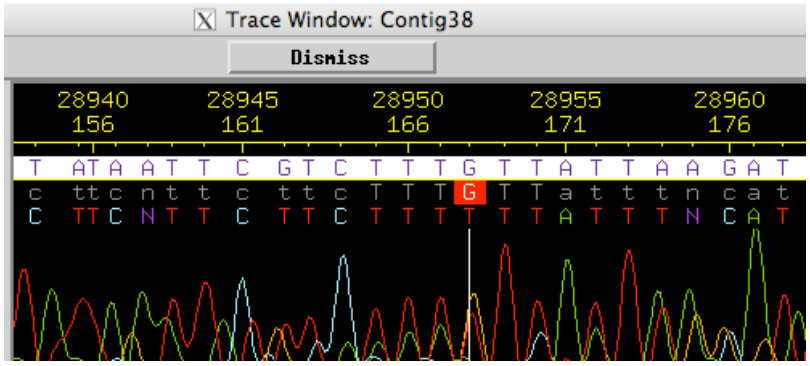

Fig. 14 The trace file showing the discrepant base.

In some cases a pad being placed in the sequence instead of calling the base caused the high quality discrepancies. Many times there was a single peak where it appears that there should have been two peaks because the single peak is much wider than a normal peak. In these cases I changed the pad to be a low quality base matching the thick peak.

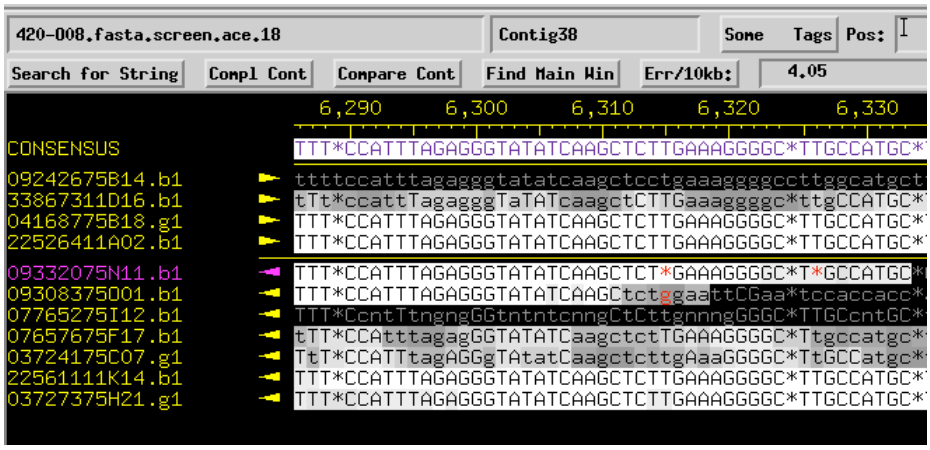

Fig. 15 Aligned reads view of the discrepancy

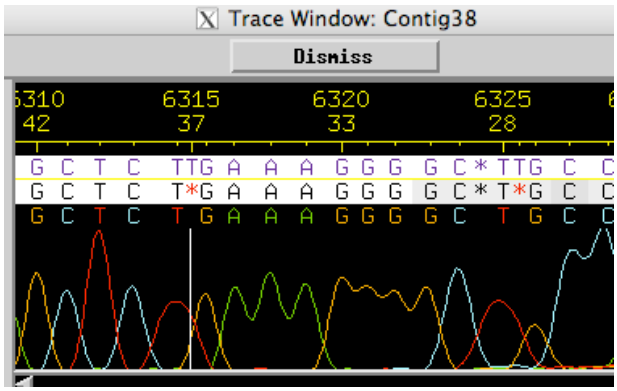

Fig. 16 Thick T peak and pad

Next I addressed single strand/chemistry/subclone regions. Many times, a single area would contain all three of these problems. These problems occurred at the very beginning and at the end of the fosmid, in the area where the force join was made in what were contig6 and 7, and two other regions of the fosmid. It was not worth calling primers for the beginning and end of the fosmid since other projects overlap that region, increasing its Phred quality. The area where the force join was made (the first primer I called for this region spanned some, but not all of the gap), despite being single stranded, had over phred 30 quality. For this reason I was not too worried about it, but I called a primer to check the region for extra conformation. Unfortunately, that primer failed to yield any data. Primers were called for the two other single sublone regions, but they also did not yield any data. Again, both of these regions have data above Phred quality 30, which is sufficient.

SCR Elgin 4/20/08 6:25 PM
Deleted: are

At this point the project has one main contig as well as fifteen other single read contigs, all less than 2kb. All but three of the latter are low quality reads, thus not worth

reincorporating. I was able to reincorporate the three reads that did have some high quality regions by using sequence match. There was a single high quality discrepancy in each case that very much resembled the other high quality discrepancies that I have already discussed. These discrepancies were easily resolved.

There are two mononucleotide runs of A's in contig13 and one mononucleotide run of T's in contig19, but none in the main contig. In addition there are no X's or N's in the consensus sequence.

The end result is the Assembly View that is shown in Fig. 17. The triangle below the assembly shows that Consed thinks there is a clone that is too long to be correct. After analyzing this situation, I found that it wasn't a problem at all. The triangle that spans the entire contig is the length of my fosmid. Consed doesn't understand that we mixed fosmid reads and subclone reads, so it has marked my entire fosmid. This is not a problem since we are expecting to have a fosmid of around 40kb.

SCR Elgin 4/20/08 6:26 PM
Deleted: on

Query= Contig57
Length=36038

No significant similarity found. For reasons why, click here.

Fig. 17 Results of the BLAST query searching for contaminating bacterial sequence.

When I had done as much as possible, I ran a blast check to ensure there is no contamination in my sequence. The feedback reflected "no significant similarity found" between my sequence and any other micro biome in the database.

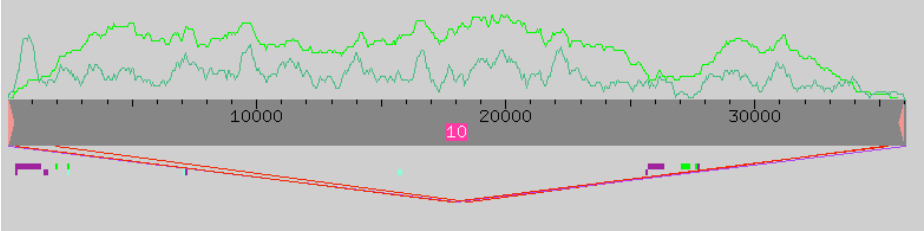

Fig. 18 Final Assembly View

In conclusion, my fosmid started out in 3 major contigs. I was able to join these into a single contig. There were no real low quality regions and all of my high quality discrepancies are now resolved. I did not need to create any assembly pieces (fake reads). I plan to continue work on this fosmid by annotating it.