FINISHING *DROSOPHILA GRIMSHAWI* FOSMID CLONE DGA06H06

Jeannette Wong
Professor Elgin
Research Explorations in Genomics
02 March 2009

Abstract:

The goal of Bio4342, Research Explorations in Genomics, is to finish and annotate the fourth, or dot, chromosome of various *Drosophila* species. This paper contains my finishing work on the project DGA06H06, which contains a region of the dot chromosome from *D. grimshawi*. This project started with three contigs containing multiple high quality discrepancies, low quality consensus regions, and several repetitious regions that needed to be resolved. Analysis of data included resolving low consensus quality data, high quality discrepancies, and gap closures. In addition, digest comparisons between the finished sequence and actual plasmid were analyzed to determine the accuracy of the finished sequence. All problems were resolved except for the repetitious regions that were difficult to align properly in the fosmid. While this could have contributed to some of the digest comparison discrepancies, I was still able to finish DGA06H06.

Introduction:

Finishing of DGA06H06 will contribute to accurately distinguishing heterochromatic and euchromatic domains of the dot chromosome via the annotation process, which is based on sequence organization and gene characteristics. In addition, it has been suggested that while heterochromatin formation may be targeted by the presence of repetitious sequence elements, there are still other factors involved in the basic mechanism. While the dot chromosome appears to be mainly heterochromatic in *D. melanogaster* and mainly euchromatic in *D. virilis*, it is unclear what the make-up of the dot chromosome in *D. grimshawi* is. Finishing of the *D. grimshawi* dot chromosome will provide insight into what heterochromatin and euchromatic markers are present. Comparisons among all fully finished and annotated *Drosophila* species will provide a more accurate understanding of the mechanisms of heterochromatin and euchromatin formation. In addition, it will provide insight into how the differences in dot chromosome content among *Drosophila* species have evolved. This report thus describes the work I have accomplished in the spring of 2009 in finishing my *D. grimshawi* dot chromosome fosmid to mouse standard (single assembly, no more than one error per 1000 base pairs).

Analysis:

My analysis began with an Initial Assembly View of DGA06H06 in Consed, which displayed three separate contigs and multiple problems that needed to be resolved (Figure 1). The data quality of my fosmid was low in many regions, which was indicated by the dark green line running across the entire fosmid. The light green line represented the total number of reads at each position, which was also low in some regions. Running Crossmatch also revealed a large area of repeats in my assembly that span almost all of Contig 2 and parts of Contigs 1 and 3. More so, there were inconsistent forward/reverse pairs between and within contigs from individual subclones that were indicated by red lines. In addition, adding up all the *in-silico* fragments from the EcoRV digest revealed that the total length of my fosmid was about 35,690 base pairs.

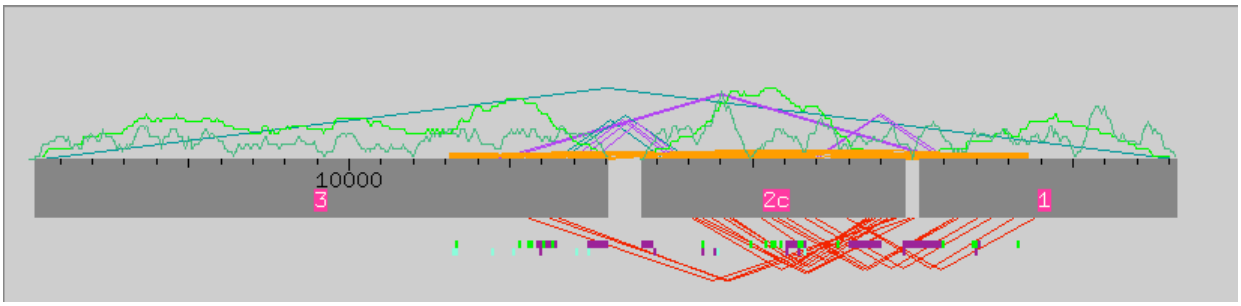


Figure 1: Assembly View of Fosmid HGA06A06

The original assembly also depicted Contig 2 in its complemented orientation, as indicated by the 'c' shown after the number. Complementation in Contig 2 means that it is in a 3'-5' orientation as opposed to the 5'-3' orientation of Contigs 1 and 3. Thus, before any contigs could be joined together, I manipulated Contig 2 into its complement in order to arrange it in the same direction as the rest of the contigs (Figure 2).

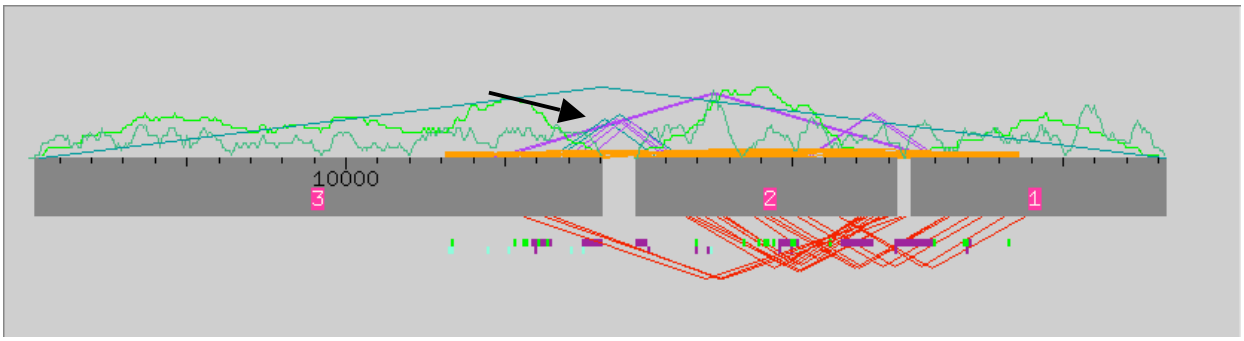


Figure 2: After complementation, forward/reverse pairs shown, and crossmatch run

First, I noticed the purple lines connecting Contig 2 and 3, which suggests that it would have been possible to close the gap between them (marked by arrow in Figure 2). However, Contig 2 mostly consisted of repetitious sequences as well as inconsistent forward/reverse pairs. In addition, there were inconsistent forward/reverse pairs between Contigs 2 and 3.

To ensure that the force join between Contig 2 and 3 was correct, I pulled out these pairs and placed them into individual contigs (Figure 3). In my Main Consed Window, the pairs pulled out were now listed as Contigs 4 through 9. I then performed a "search for string" and attempted to match the end of Contig 3 with one of the ends of Contig 2. However, this resulted in a match between sequences on Contig 3 and Contig 1, which is supported by the presence of the purple lines connecting the end of Contig 3 with the beginning of Contig 1 (marked by arrow in Figure 3). Comparisons of the paired end reads of these two contigs showed a complete match between their sequences, and a join was forced, creating Contig 10 in Assembly View (see Figure 4). While the total length of my fosmid was about 4kb smaller than the standard 40kb size, I could not find new oligos to cover the gaps seen in the Initial Assembly given the highly repetitious ends present on all three contigs. Thus, the best solution to closing these gaps was to use just the reads present in the fosmid; I was then able to manipulate these sequences to close all gaps.

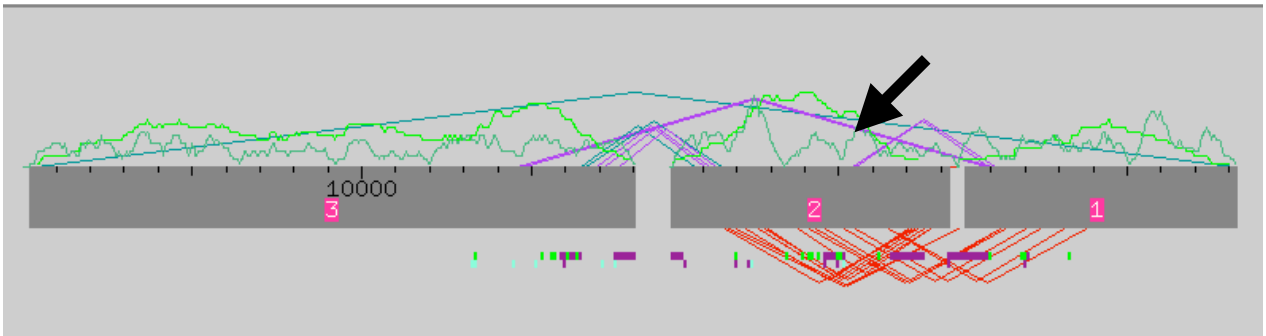


Figure 3: Inconsistent Forward/Reverse Pairs between Contig 2 and 3 removed

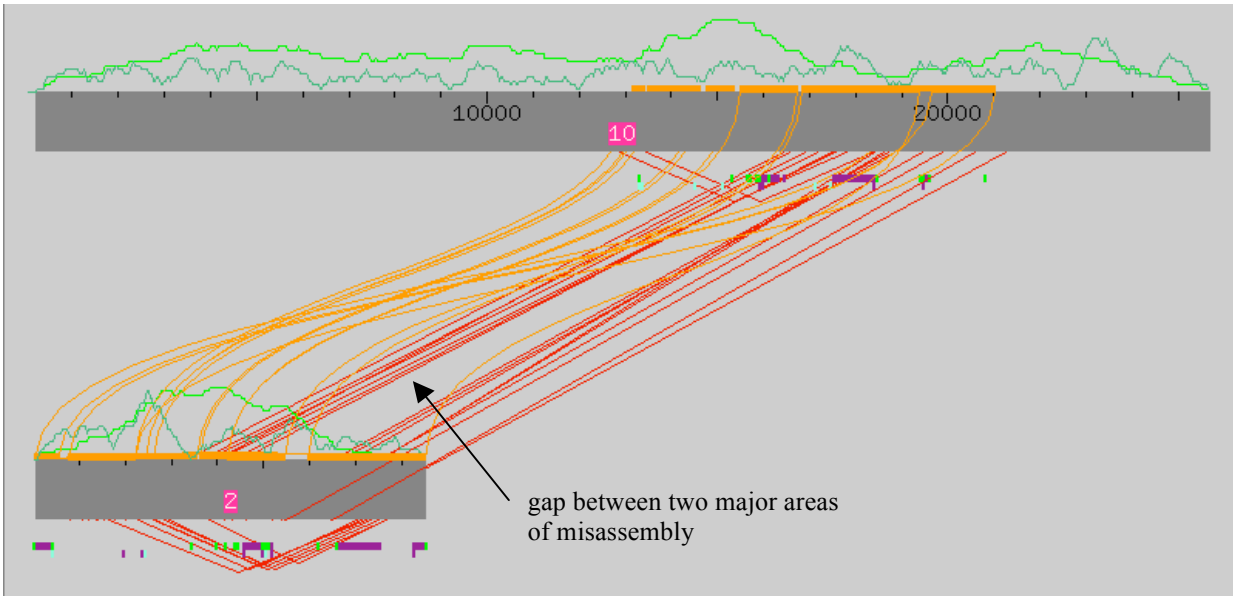


Figure 4: Force join between Contig 1 and 3, crossmatch, forward/reverse pairs

Now, one can see the major misassemblies between Contig 10 and Contig 2. As noted in Figure 4, there is a big gap between these two regions of inconsistent forward/reverse pairs. When I compared the direction of the Contig 10 and Contig 2 sequences in this gap, I noticed that all the contig 10 sequences ran the opposite direction to the corresponding contig 2 sequences (see Figure 5).

```
02308540F11.b1 <- Contig10 17874-18723 inconsistent because: too far from end of contig. lib: default 18722 from end
02308540F11.g1 -> Contig2 4140-4909 inconsistent because: too far from end of contig. lib: default 4376 from end
02354840K03.b1 <- Contig10 18059-18844 inconsistent because: too far from end of contig. lib: default 18843 from end
02354840K03.g1 -> Contig2 4305-5055 inconsistent because: too far from end of contig. lib: default 4211 from end
02569540B13.b1 -> Contig2 4611-5533 inconsistent because: too far from end of contig. lib: default 3905 from end
02569540B13.g1 <- Contig10 17907-18720 inconsistent because: too far from end of contig. lib: default 18719 from end
```

```
00618840A10.b1 -> Contig10 16913-17686 inconsistent because: too far from end of contig. lib: default 8756 from end
00618840A10.g1 <- Contig2 276-1060 inconsistent because: too far from end of contig. lib: default 1059 from end
02586140D17.b1 <- Contig2 269-1191 inconsistent because: too far from end of contig. lib: default 1190 from end
02586140D17.g1 -> Contig10 17155-17954 inconsistent because: too far from end of contig. lib: default 8514 from end
02568540L21.b1 -> Contig10 16567-17393 inconsistent because: too far from end of contig. lib: default 9102 from end
02568540L21.g1 <- Contig2 728-1548 inconsistent because: too far from end of contig. lib: default 1547 from end
```

Figure 5: Left (top box) and Right (lower box) regions of inconsistent forward/reverse pairs

This suggested that if I make a tear in between these two regions of inconsistent forward/reverse pairs, I will be able to insert Contig 2 into the middle of Contig 10. This would properly orient all reads in the same direction. To make this possible, a tear was made between 18000 and 19000 base pairs in Contig 10, which then created Contigs 11 and 12 (Figure 6).

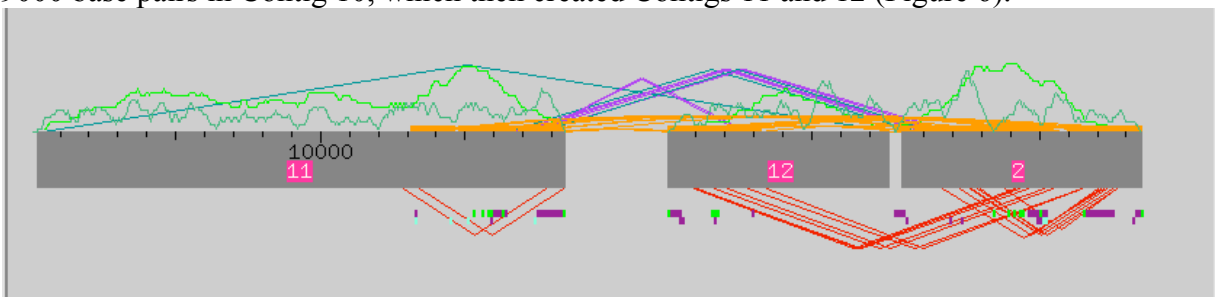


Figure 6: After tear in Contig 10

The tear made in Contig 10 created new problems that had to be resolved before all contigs could be successfully joined together. One of the new problems was that Contig 2 was not inserted between Contig 11 and 12. This is where the tear between Contig 10 was made to allow the incorporation. In addition, the only inconsistent forward/reverse pairs were now between Contig 12 and Contig 2 and within Contig 2.

Based on the forward/reverse pair information for these discrepancies, I decided to pull these reads out of their respective contigs. I hoped this would also cause Contig 2 to move into the gap between Contigs 11 and 12 with the inconsistencies removed. I then decided to experiment with the reads I had pulled out by running a mini-assembly analysis. This fortunately produced two new contigs, 14 and 15, that aligned between Contig 11 and Contig 2. The order of the contigs in Figure 7 not only justified the tear I made earlier in Contig 10 but also depicts how the mini-assembled contigs fit adjacently to the original contigs.

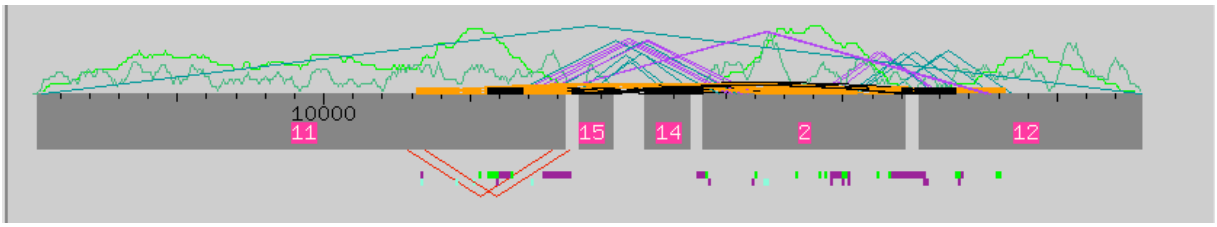


Figure 7: correct order of all contigs

From Figure 7, the black boxes and lines mean that the two repeated regions are on different strands of DNA. Despite the highly repetitious regions within my fosmid, I was still able to search for paired end reads that only matched two different contigs. Therefore, I was confident in the force joins that I utilized in the process of finishing my fosmid. First, I decided to manipulate Contigs 14 and 15 into their reverse complements before performing a "search for string" analysis in an attempt to join all five contigs. From the "search for string" analysis, I concluded that the contig order should be 11-2-15c-14c-12. There were no nucleotide discrepancies when I checked the paired end reads after alignment and before the force joins. Thus, I was able to join all five contigs into one contig (Figure 8). I was confident in these force joins and now needed to check for low quality consensus regions, high quality discrepancies, single chemistry, and single subclone regions.

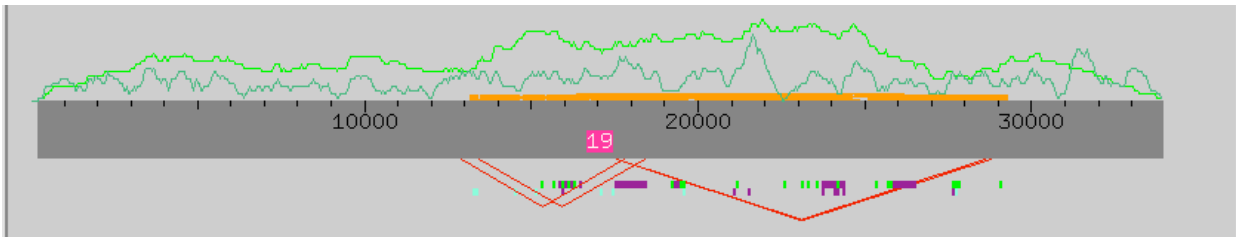


Figure 8: complete assembly of all contigs

Navigators:

I first checked Contig 19 for any high quality discrepancies. There were six high quality discrepancies. Five of the six resulted from sequencing artifacts. Compression occurred between two nucleotides, creating one peak or a broad one with two weaker peaks (indicated by arrow in Figure 9). Since these are not strong indications that warrant me to change or doubt the consensus, these discrepancies were tagged with comments explaining the sequencing altercations.

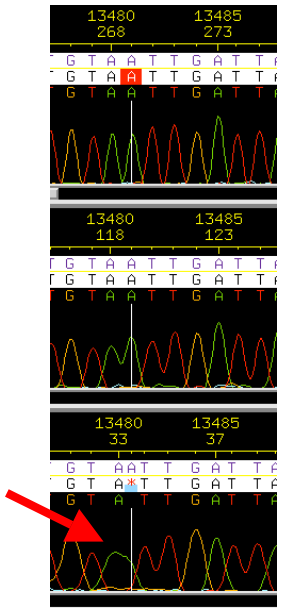


Figure 9: comparison of high quality discrepancies to other reads

However, the last high quality discrepancy showed strong peaks for both nucleotides (C and T) called at bp 17604 (Figure 10). I decided to run a "search for string" in this region with a sequence that included the T nucleotide. Since this did not match the consensus sequence, I wanted to see if another region of my highly repetitious fosmid contained the same sequence minus the one substitution of T for a C. I found this sequence in the 25,000 bp region of Contig 19. Thus, I removed these two reads from Contig 19 and realigned them in a separate contig before reincorporating them into the right place. My complete fosmid was now labeled Contig 23 after resolving this discrepancy.

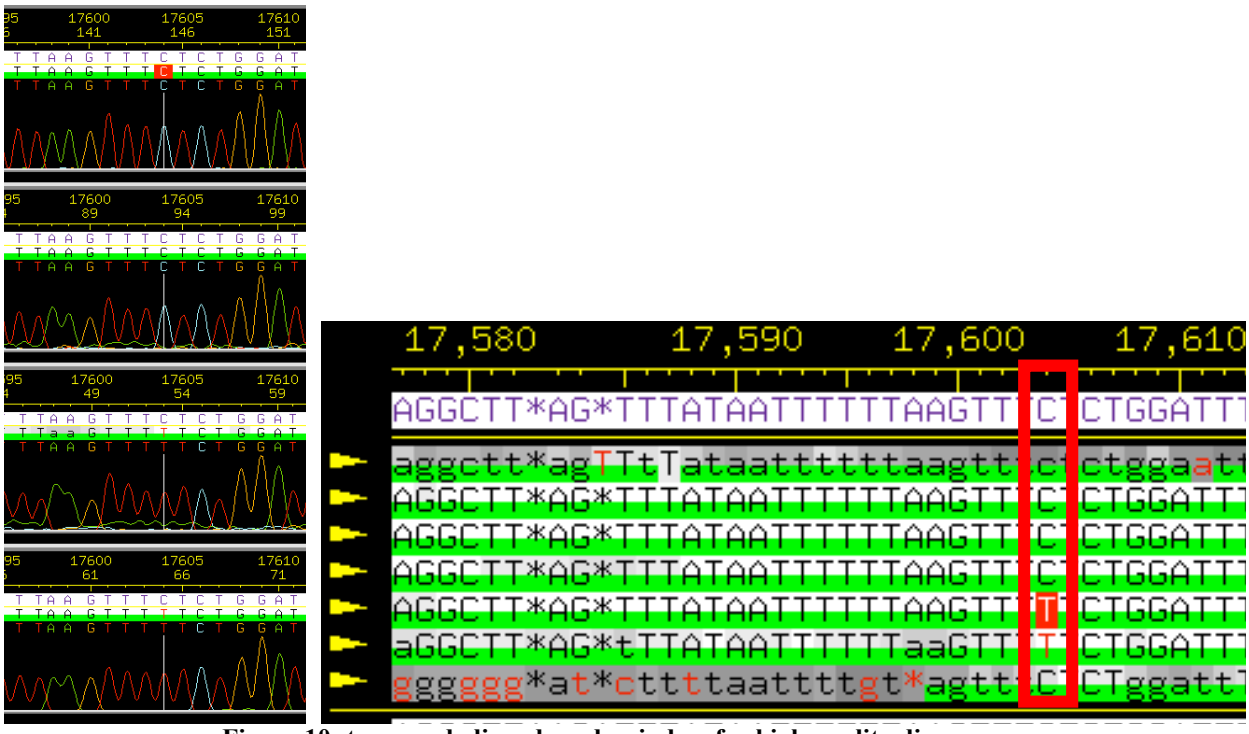


Figure 10: trace and aligned reads window for high quality discrepancy

Figure 11: purple highlighted reads are reincorporated into proper region to resolve high quality discrepancy

I also checked my fosmid for low quality consensus regions and compared these regions to those that were only covered by "one strand and only one chemistry (or none at all)." There were three major low quality consensus regions; I designed oligo primers to cover the 1-343 and 22597-22600bp regions. These were low quality areas that needed additional data to bring up the quality of these regions to above a threshold score of Phred 25. Any regions that had only one strand or chemistry and a Phred score below 30 were checked as well. However, the primers already designed for low consensus quality areas also covered these regions.

In addition, a "search for string" was done to check for mononucleotide runs. There was one sequence of 17 A's, so I designed primers to flank both ends of this region. I also checked for any single subclone regions and located one in the 9,000- 10,000 base pair region. I then created oligo primers to flank both ends of this region as well. Table 1 lists the first round of primer sequences and chemistries used to tackle these problems. I used 4:1 chemistry for all the

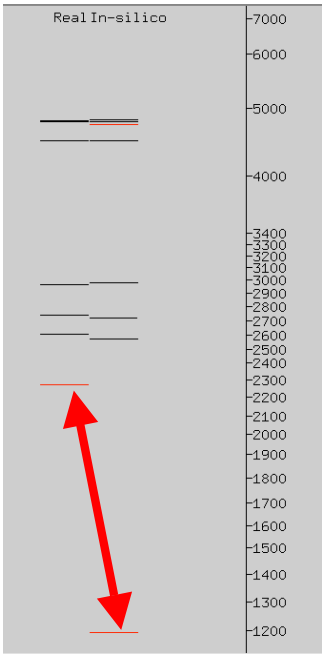
designed primers except the primers that flanked the mononucleotide run. I decided to use all three chemistries for these primers in order to make sure this area was sufficiently analyzed.

Reaction Name	Reason	Sequence	Result
selgin09XBAC-DGA06H06_t1.b1	--> LQ region bp 1-362	cataccattggagagtacgga	fail
selgin09XBAC-DGA06H06_t2.b1	--> LQ region bp 338-343	agttatgcgtattaataaagtttgc	fail
selgin09XBAC-DGA06H06_t3.b1	<-- LQ region bp 338-343	gcatagattgcctgagacaag	fail
selgin09XBAC-DGA06H06_t4.b1	--> LQ region bp 22597-22600	ggtactgagaatgttgctgatctat	success
selgin09XBAC-DGA06H06_t5.b1	<-- LQ region bp 22597-22600	catgtgcaaagttaaagcaaata	success
selgin09XBAC-DGA06H06_t6.b1	--> subclone region	ttaaaacgggttgtatcatttc	success
selgin09XBAC-DGA06H06_t7.b1	<-- subclone region	ggttgctgtcaatgcttatatta	success
selgin09XBAC-DGA06H06_8.b1	--> mononucleotide run of A's	aattcatacaaatgcatacgtaaat	success
selgin09XBAC-DGA06H06_t8.b1	--> mononucleotide run of A's	aattcatacaaatgcatacgtaaat	fail
selgin09XBAC-DGA06H06_g8.b1	--> mononucleotide run of A's	aattcatacaaatgcatacgtaaat	success
selgin09XBAC-DGA06H06_9.b1	<-- mononucleotide run of A's	aaggcatagggagcgata	fail
selgin09XBAC-DGA06H06_t9.b1	<-- mononucleotide run of A's	aaggcatagggagcgata	fail
selgin09XBAC-DGA06H06_g9.b1	<-- mononucleotide run of A's	aaggcatagggagcgata	fail

Table 1: Round 1 reactions

Round One Reaction Results:

The reads resolved all low consensus quality regions except for the right end of my fosmid. A vector insert is present from -27 to 345bp, the 5' end, in Contig 23 after the first round of reactions, so any low quality data in that region can be disregarded. A clone end tag was thus added at 345bp. The 3' end did not have a vector insert after the first round of reactions, but the low consensus quality region was almost the last 40 bp of the contig. Therefore, I decided to design another oligo primer to attempt to resolve the 3' end of Contig 23. In addition, I looked at my digests to compare the real and *in-silico* fragment sizes. From the HindIII digest, I noticed that a 1kb fragment seemed to be missing from the 4862-6056 bp region (arrow in Figure 12). Checking the digests revealed no miscalled or uncalled bands either. Thus, I decided to flank the ends of this region with oligos as well to see if I could find any discrepancies in the read. Given the high number of repetitious areas in my fosmid, it proved difficult to find digests that would support my finished contig. These were the only three oligo primers I designed for the second round, and I decided to utilize all three reactions for each primer since I wanted to make sure my results were as accurate as they could be (Table 2).




```
---------    ----------------------
  Real           In  Silico
Frag Size    Size    Position
---------    -----   --------------
    9042      8926  part vector/part insert Contig23 (33089-33908)
    4795      4805  Contig23 (6344-11149)
    4775      4764  Contig23 (23918-28682)
              4724  Contig23 (16034-20758)
    4448      4447  Contig23 (11149-15596)
    2966      2983  Contig23 (1015-3998)
    2741      2722  Contig23 (20758-23480)
    2610      2576  Contig23 (30513-33089)
    2272
              1194  Contig23 (4862-6056)
              1047  part vector/part insert Contig23 (1-1015)
               977  Contig23 (29432-30409)
               864  Contig23 (3998-4862)
               750  Contig23 (28682-29432)
               438  Contig23 (15596-16034)
               438  Contig23 (23480-23918)
               288  Contig23 (6056-6344)
               104  Contig23 (30409-30513)
```

Figure 12: HindIII digest with observed missing 1kb fragment

Reaction Name	Reason	Sequence	Result
selgin09XBAC-DGA06H06_10.b1	--> LQ region at 33000 bp	tctcaattatttatagcgctctatt	success
selgin09XBAC-DGA06H06_g10.b1	--> LQ region at 33000 bp	tctcaattatttatagcgctctatt	success
selgin09XBAC-DGA06H06_t10.b1	--> LQ region at 33000 bp	tctcaattatttatagcgctctatt	success
selgin09XBAC-DGA06H06_11.b1	--> check for missing 1kb	ccaaattctaattcttctcagttct	success
selgin09XBAC-DGA06H06_g11.b1	--> check for missing 1kb	ccaaattctaattcttctcagttct	success
selgin09XBAC-DGA06H06_t11.b1	--> check for missing 1kb	ccaaattctaattcttctcagttct	success
selgin09XBAC-DGA06H06_12.b1	<-- check for missing 1kb	catgcaaatgcagaaccta	fail
selgin09XBAC-DGA06H06_g12.b1	<-- check for missing 1kb	catgcaaatgcagaaccta	success
selgin09XBAC-DGA06H06_t12.b1	<-- check for missing 1kb	catgcaaatgcagaaccta	fail

Table 2: Round two reactions

Autofinish:

I compared the primers I had designed from the first round to those designed by Autofinish. Table 3 lists the six oligo primers created by Autofinish, but none of these primers were used in the second round of reactions that I ordered. These low-quality regions, while designed in the original contigs, corresponded to low-quality regions in Contig 23 that I created primers for. In addition, the single subclone regions on Contig 2 had Phred scores above 30, so it was not necessary to use these primers either.

oligo sequence	direction	reason
tggtaaaggtaattttcgca	<--	LQ region from 1-85bp on Contig 1
cagtcgcaggtggtgat	-->	LQ region from 8130-9066bp on Contig 1
ccatatcatattaatgtctctgtca	<--	LQ region from 17956-18892bp on Contig 3
gtttccacagaagaagaatcc	-->	single subclone from 8272-9208bp on Contig 2
ccatatcatattaatgtctctgtca	<--	LQ region from 1-754bp on Contig 3
gcgttagtgcgatgttttgt	-->	single subclone from 4704-5640bp on Contig 2

Table 3: Autofinish primers

Round Two Reaction Results:

Not far apart enough to be concerned about

The reactions from round two successfully resolved the right end of contig 23, but there were some new problems. First the read selgin09XBAC-DGA06H06_t12.b1 was added to the wrong region of the contig. It was called for the 3' end of the region where the digest comparisons suggested that there was a 1kb sequence missing in this region. Therefore, the information from this read is not useful and considered a failure. I pulled out the read from Contig 23 and placed it in its separate contig. In addition, no new data from the reads suggested there was a 1kb fragment missing from this region, so the problem was not resolved. However, the 33,894-33908bp region is resolved. Looking at Figure 13, the original reads in this region are of very poor quality compared to the new reads. They contained several overlapping, broad, and sloppy-looking peaks. I changed the consensus in this low quality region to accommodate the results from the reactions called in the second round. The high quality consensus discrepancy at 33896bp was also changed to an A because of the evidence seen in the new reads that have been added to my fosmid. In addition, the new reads allowed me to extend my consensus sequence to the beginning of the vector sequence at 33925 bp. Thus, a second clone end tag was added to Contig 23 to complete the finished sequence (Figure 14). There are two inconsistent forward/reverse pairs left in Contig 23, but I ignored them since they were not significantly far apart from one another.

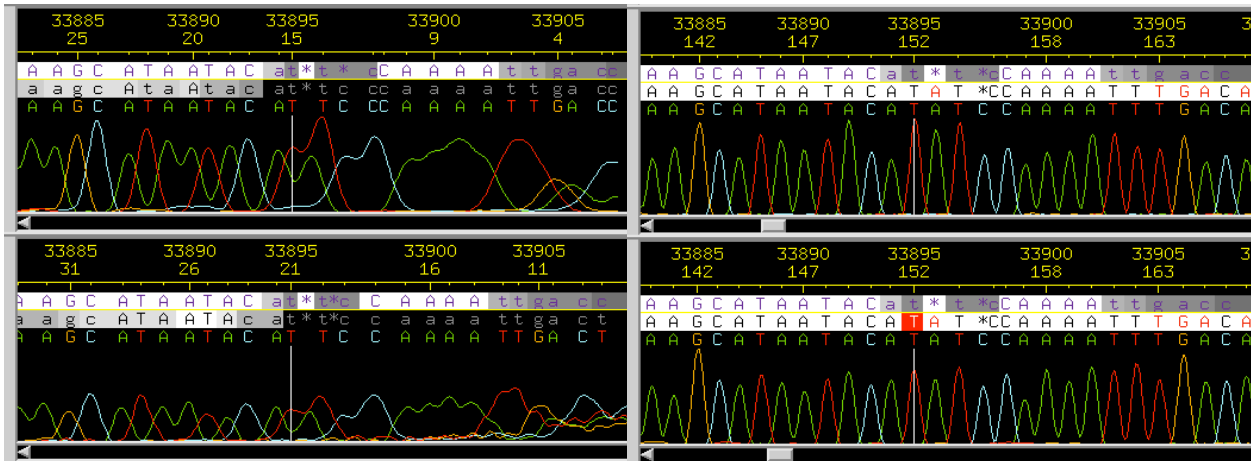


Figure 13: comparison of original reads (left) to produced reads (right) for the same region

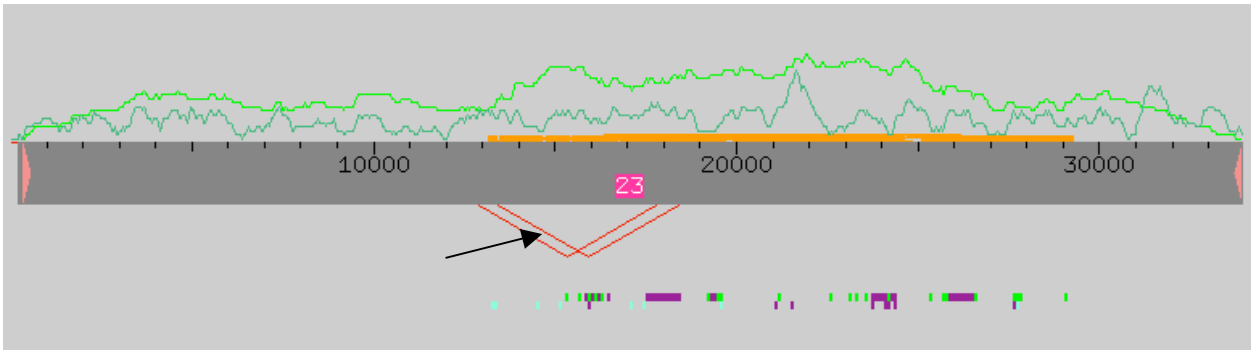
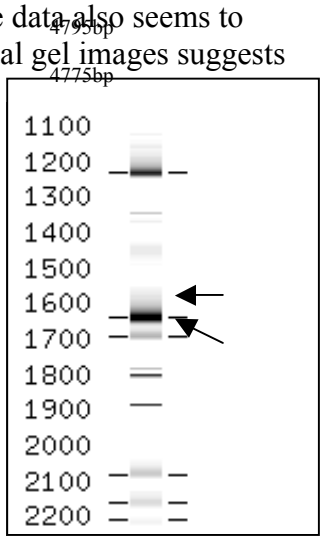


Figure 14: Finished sequence for DGH06A06

Digest Comparisons:

Of the four digests I had available to compare my contig to, only two provided some interpretable evidence to support my finished sequence (Figure 15). While the HindIII digest is not absolutely perfect with the mismatched fragments of 2272 and 1194 base pairs, the rest of the *in-silico* fragments corresponded well with the real size fragments. The data also seems to suggest that there is a doublet at the 4775 bp fragment size, but the actual gel images suggests the doublet actually occurrs at the 4795 bp fragment size. This is apparent based on the color density in this band versus the 4775 band. Thus, the HindIII digest provides sufficient evidence to support my finished fosmid after analyzing the gel image to determine where the doublet occurs in correlation to the fragment sizes. The Eco RI digest also provides some good support for my finished project (Figure 16). Since my finished fosmid is only about 34kb in length, the 40kb and 35kb real fragment sizes were not plausible results to compare my finished sequence to. Since they were much larger in length than my finished fosmid, these fragments could be disregarded. In addition, there was supposedly a doublet of 21kb fragment sizes, which is also not plausible since the sum of these two fragments would be more than 40kb as well. However, the two 10kb *in silico* fragment sizes correspond well to the 10kb real fragment sizes. These numbers are similar enough to also provide some confidence in the sequence of my finished fosmid. In addition, the 10,660bp band contains the region where I thought there was a 1kb fragment missing. It is apparent from this digest that nothing was missing from that region. The other two digests contained larger differences between the bands that could not be used to support my project.



Real Frag Size	In Silico Size	Position
40067		
35740		
21056	20459	Contig23 (13012-33471)
21056		
10735	10660	Contig23 (2352-13012)
10320	10460	part vector/part insert Contig23 (1-2352)
	469	part vector/part insert Contig23 (33471-33909)

Figure 16: EcoRI Digest

BLAST Analysis:

I ran a BLAST analysis on my finished fosmid. The BLAST program allows me to compare my finished sequence to a microbial database. If there were any commonalities between my sequence and a microbial sequence, then this would suggest that some contamination had been introduced into my fosmid. If there were no matches, then I am more confident that the sequence is correct. BLAST revealed no matches with any existing microbial genomes in the BLAST database, suggesting no contamination has occurred in my complete fosmid (Figure 17).

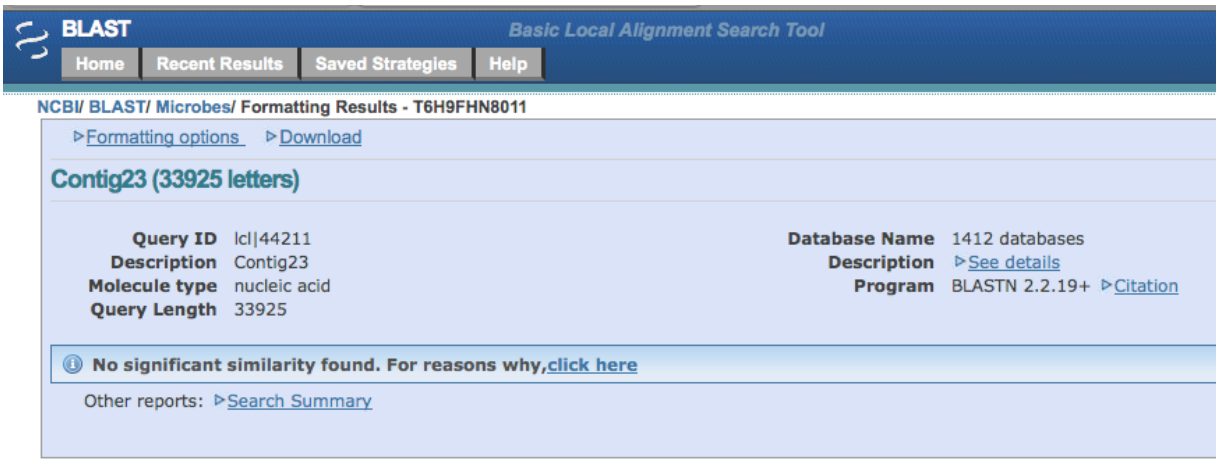


Figure 17: BLAST analysis

Conclusion:

I was able to successfully finish DGA06H06 by joining the contigs to one another as well as utilizing the forward/reverse pairs to fill in some of the other gaps. All discrepancies are resolved and the entire sequence is well above the Phred threshold score of 25. I am confident in my sequence after comparing it to the HindIII and EcoRI digests as well as running it through the BLAST program. Given more time to work with this fosmid, I would look more closely into the large tandem repeat area and attempt to discern whether or not it is possible to move some of the sequences around to make the fosmid more accurate when in comparison to the digests.