Finishing *Drosophila grimshawi* Fosmid Clone DGA23F17

Kenneth Smith
Biology 434W
Professor Elgin
February 20, 2009

Abstract:

My fosmid, DGA23F17, did not present too many challenges when I finished it's sequence to high quality. The major problems with my project mostly stemmed from large regions of low quality data. Long runs of poor read quality led to the three gaps present in my initial Assembly View. Due to the length of these regions, one primer was not enough to cover the whole gap. Having to use sequence data obtained from the previous round to call primers thereafter presented the most difficulty with this particular fosmid. By attacking each gap from both sides and tripling my expected read depth by using all three chemistries for each oligo, I was able to overcome this problem and resolve my entire fosmid to high quality.

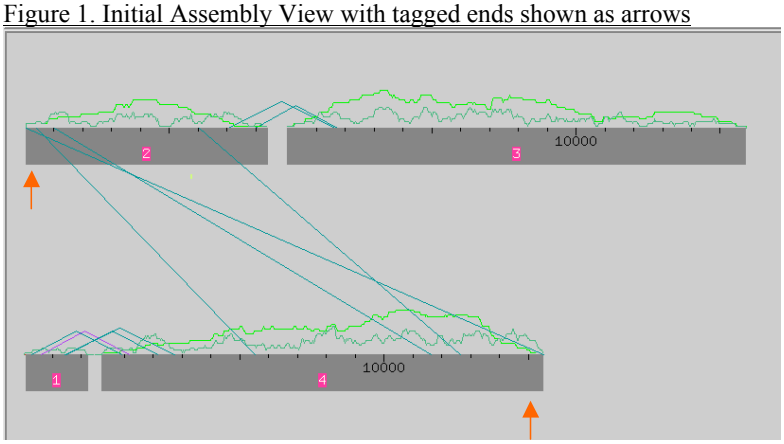
Introduction:

It is known that the small fourth chromosome of *Drosophila melanogaster* contains 80 genes. This chromosome is called a "dot chromosome" because of its small relative size in comparison to the others. It is our task in Bio4342 to finish part of the *D. grimshawi* dot chromosome and later determine how many genes are encoded in this chromosome. This species has already been sequenced but the data is raw, containing many miss-assemblies and gaps, which impede scientists as they try to draw conclusions from the data. Even though it encodes many genes, the dot chromosome contains many repetitious sequences and is comprised of mostly heterochromatin. Due to these facts, this chromosome's finished sequence is valuable to researchers studying transcriptional regulation and heterochromatin function in the cell. This project's aim is to make the existing sequence of my particular fosmid higher quality.

Initial Assembly:

Figure 1 displays the initial Assembly View of my project showing four contigs and several regions of low quality. I started by using the Cross Match program to see if there were any repeat regions. After discovering that there were none, I labeled the two fosmid ends, which were located at the beginning of contig 2 and the end of contig 4. Figure 1 clearly shows that there are several gaps in the fosmid. Therefore I tried to resolve these regions by comparing sequences from the 3' end of 2 and the 5' end of 3 and likewise for the other two gaps. The quality of the base pairs was too low on every contig end for any alignment to occur. This led me to order primers for the 5' and 3' ends of every contig to resolve those regions. I used all three available sequencing chemistries for the gap spanning reads and just the 4:1 mixture chemistry on the normal low quality regions.

Figure 1. Initial Assembly View with tagged ends shown as arrows



To double-check the reads that I ordered, I ran the Autofinish program and compared its calls with my own. As the table below clearly shows, Autofinish agreed with me on most of the calls, although I chose different oligos to achieve the same result. I believe Autofinish may have made a mistake on oligo 7b because there is no region within a 1000 base pairs in either direction that requires additional information.

Table 1. Round 1 Oligos I Called (top) versus Oligos Autofinish Called (bottom)

Oligo	Sequence	Direction	Problem	Result
1a	ttacaagtgggctctctttaat	<---	LQ bp 1-25	Success
2a	aagacatctctcactacgcatt	--->	LQ bp 4673-4788	Failure
3a	tcgttgtgacactaatatatggtaa	--->	LQ bp 8400, 2-3 gap	Failure
4a	aaatgatcctgttcaagttacac	<---	LQ @ beg. Contig3	Success
5a	tcctcatcctcactgtatctct	--->	LQ bp >15614	Failure
6a	acagctgctggttcactaat	--->	3-1 gap	Success
7a	tgttgagtgcatgtacaaattatta	<---	3-1 gap	Success
8a	gaaagaaagaaagctacactagca	--->	LQ bp >18746	Failure
9a	tttaatttgaattttccatttga	<---	3-4 gap	Failure
10a	tgatagtgcttatcgactgatatt	--->	LQ bp >15511	Failure

Oligo	Sequence	Direction	Problem
1b	gatgtggttcagatgtgttgt	<---	Same as oligo 1a
2b	cctccctcccactccc	--->	Same as oligo 2a
3b	cctccaattcgatatatatattgtg	--->	Same as oligo 3a
4b	aaggggtgggaggagtt	--->	Same as oligo 3a
5b	acatttaaataggcgtgctct	<---	Same as oligo 4a
6b	ctggccccagtatatattctt	--->	LQ @ beg. Contig3
7b	actttctaagatttgaacagaacag	--->	???
8b	caataacaataagctgttgatcct	--->	Same as oligo 5a
9b	gctgctggttcactaatacaacta	--->	LQ @ end Contig3
10b	tttattgtcgctctgattttaat	<---	Same as oligo 9a
11b	tttgatagtgcttatcgactga	--->	Same as oligo 10a

When I chose the areas that needed to have additional reads in round 1, I looked at the specific problem areas on each contig. I clustered the low quality base pairs together to consolidate the problem areas. For the example in Figure 2, I viewed 2110-2449 as one

region (designated by the orange box) and designed oligos at the beginning and end of the region to ensure proper coverage.

Figure 2. List of Problem Areas for Contig 1

Contig Name	Read Name	Consensus Positions	
Contig1	(consensus)	1-80	base quality below threshold
Contig1	(consensus)	1-2449	2454 bp single strand/chem
Contig1	(consensus)	1-284	284 bp single subclone
Contig1	(consensus)	1433	base quality below threshold
Contig1	(consensus)	2110	base quality below threshold
Contig1	(consensus)	2132-2135	base quality below threshold
Contig1	(consensus)	2174	base quality below threshold
Contig1	(consensus)	2180	base quality below threshold
Contig1	(consensus)	2182-2184	base quality below threshold
Contig1	(consensus)	2210-2449	base quality below threshold
Contig1	(consensus)	2223-2449	227 bp single subclone

To gauge the original problems, I also examined the digests and in Figure 3, two of the four digests are shown illuminating the many differences between real and *in-silico* digests. The red lines on the *in-silico* lane show the discrepant bands, representing regions where the restriction enzyme digested the DNA at a location that is different from where the digestion occurred in the real DNA. This information shows that my sequence has many discrepancies and thus needed much work.

Figure 3. Initial *Hind*III (left) and *Eco*RV (right) Restriction Digest

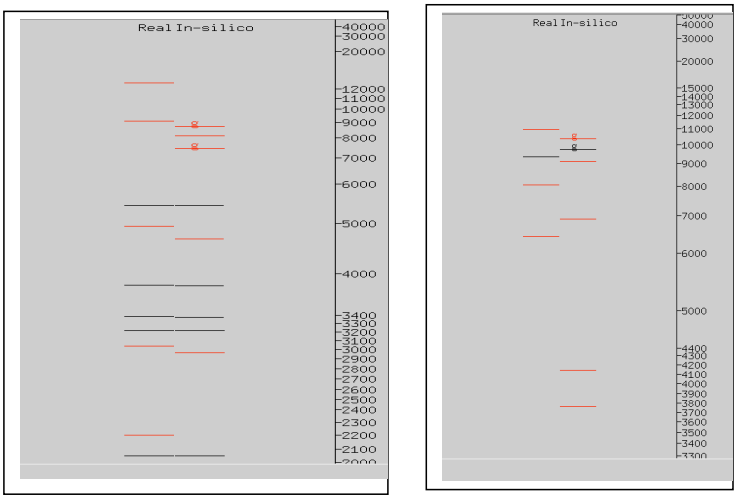
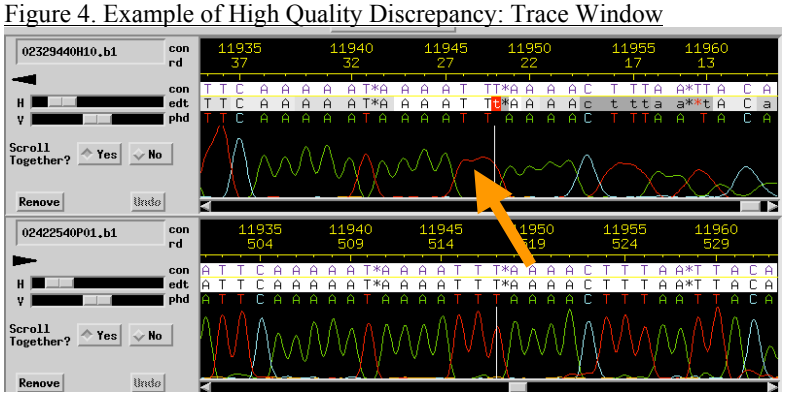
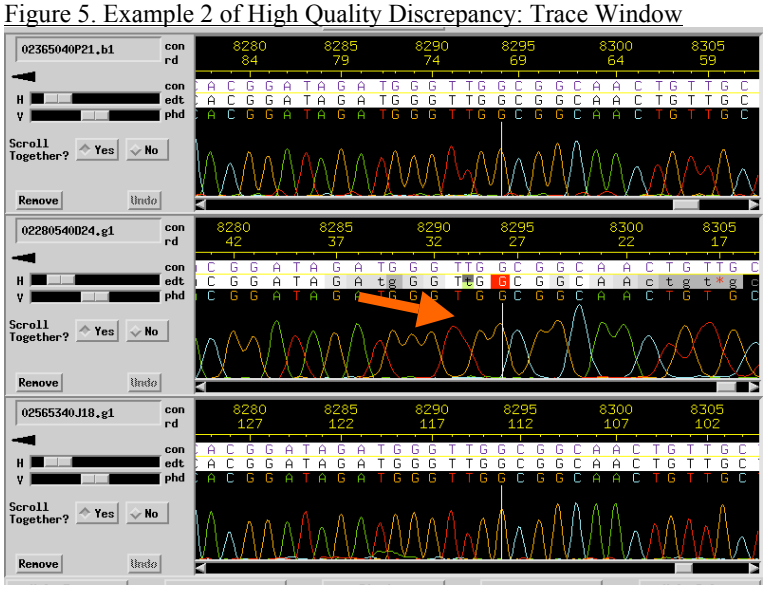


Figure 4. Example of High Quality Discrepancy: Trace Window



I then began to work on the high quality discrepancies by looking at the trace windows. A high quality discrepancy is when two reads both have high quality data, but their sequences do not match at every position. In the example above (Figure 4 read 02329440H10.b1), Consed labeled two high quality Ts followed by a pad, which is represented by a space. This disagreed with the consensus, which had three Ts. From the high quality peaks found in the bottom read (as well as several other reads at this base pair) it appears as if there should be three Ts. This discrepancy is known as a compression where the chemistry in the reaction compresses two (or more) peaks and makes them appear as one peak. I manually edited the read by changing the pad to a "T." Another example of a compression that I had to manually edit is shown in Figure 5.

Figure 5. Example 2 of High Quality Discrepancy: Trace Window

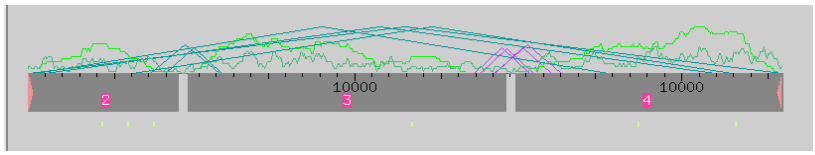


The T peak on the middle read (orange arrow) is shown as one T whereas consensus and the other two reads clearly displays two Ts. I changed the pad into a "T."

Round 1 Results:

After I received my new reads from Round 1, I reran PhredPhrap with mixed results. My new assembly view had been reduced to three contigs with the small contig 1 being incorporated into contig 3 (Figure 6). The new sequences increased the quality of the 5' end of the original contig 2 and the gap between contig 3 and contig 1, but on the majority many of the problems still remained. The failure of the new reads stemmed mainly from poor quality. Several of the new reads consisted entirely of low quality bases and others produced many discrepant bases. In addition, the digests remained relatively unchanged, indicating that most of the same problems had to be resolved in Round 2.

Figure 6. Assembly View After Round 1 Reads Incorporated



Round 2:

I ordered oligos for many of the same regions but chose different primers to obtain better results. I also added more coverage by ordering two different primers to cover each region. One of the more challenging aspects of my fosmid was the large low quality region at the 3' end of contig 4 and at the ends of the gap between contigs 2 and 3. This was a difficulty because both regions represented around 1 kb of DNA, and the upper limit of quality data for a new sequencing read is around 600 bases. Thus one read would not have been sufficient to resolve the area.

Kenneth Smith 3/31/09 2:40 PM
Deleted: s

For these regions I had to order a read in Round 1 and then order a new read in Round 2 whose primer was contained in the information from Round 1. I ran into this problem in the second round, because several of my Round 1 reactions failed. Thus I did not have a strong consensus sequence to order the Round 2 primers. To alleviate this problem, I used all three reaction chemistries on two different primers from each end of the gap. Two primers came from the 5' end of the gap and the other two from the 3'. This allowed me to achieve 12 times coverage for each of the problem regions (4 primers per gap and 3 chemistries per read).

Round 2 Results:

When I received my new reads and added them into my Consed file, I reran PhredPhrap again. This time the three contigs had been reduced to one and the only problem area that remained was a low quality region between bases 8406 and 8936 (Figure 7). The gap-

spanning reads were very successful. Instead of having to order new reads in Round 3 like I discussed in the previous paragraph, the reads coming from both sides had just enough overlap in the middle of the gap for PhredPhrap to complete the join.

Figure 7 shows a list of the remaining problems after the Round 2 reads had been incorporated. It should be mentioned that even though this list looks long, the only areas we considered to be a problem are high quality discrepancies and base quality below threshold. The single stranded/single chemistry regions that are labeled are all above a Phred score of 30.

Figure 7. Problem List After Round 2

Contig Name	Read Name	Consensus Positions	
Contig2	(consensus)	1–22	22 bp single subclone
Contig2	(consensus)	1–22	22 bp single strand/chem
Contig2	(consensus)	297–1534	1250 bp single strand/chem
Contig2	(consensus)	1955–1973	19 bp single strand/chem
Contig2	(consensus)	4671–5453	803 bp single strand/chem
Contig2	(consensus)	4786–4788	base quality below threshold
Contig2	(consensus)	4864–4918	55 bp single subclone
Contig2	(consensus)	5912–6135	228 bp single strand/chem
Contig2	(consensus)	7274–7441	168 bp single strand/chem
Contig2	(consensus)	7864–7979	116 bp single strand/chem
Contig2	(consensus)	7864–7979	116 bp single subclone
Contig2	(consensus)	8202–8351	150 bp single subclone
Contig2	(consensus)	8202–9060	884 bp single strand/chem
Contig2	(consensus)	8406–8415	base quality below threshold
Contig2	(consensus)	8449–8459	base quality below threshold
Contig2	(consensus)	8888	base quality below threshold
Contig2	(consensus)	8896–8900	base quality below threshold
Contig2	(consensus)	8902	base quality below threshold
Contig2	(consensus)	8905–8916	base quality below threshold
Contig2	(consensus)	8920–8925	base quality below threshold
Contig2	(consensus)	8933–8936	base quality below threshold
Contig2	(consensus)	9670–10148	478 bp single strand/chem
Contig2	(consensus)	10460–11098	643 bp single strand/chem
Contig2	(consensus)	11929–12101	177 bp single strand/chem
Contig2	(consensus)	18811–19209	399 bp single strand/chem
Contig2	(consensus)	20077–20282	210 bp single strand/chem
Contig2	(consensus)	21328–21367	40 bp single strand/chem
Contig2	(consensus)	22502–22653	152 bp single subclone
Contig2	(consensus)	22502–22653	152 bp single strand/chem
Contig2	(consensus)	22983–23558	576 bp single strand/chem
Contig2	(consensus)	24215–24326	117 bp single strand/chem
Contig2	(consensus)	25123–25403	281 bp single strand/chem
Contig2	(consensus)	26414–28060	1661 bp single strand/chem
Contig2	selgin09XBAC-DGA23F17_g24.b1	28668–28696	29 unaligned high quality
Contig2	(consensus)	28669–29160	519 bp single strand/chem
Contig2	selgin09XBAC-DGA23F17_g19.b1	29422	high quality base disagrees with consensus
Contig2	(consensus)	30245–30350	106 bp single strand/chem
Contig2	(consensus)	31919–31996	78 bp single strand/chem
Contig2	(consensus)	32918–32944	27 bp single strand/chem
Contig2	(consensus)	37565–37689	136 bp single strand/chem
Contig2	(consensus)	38847–38853	7 bp single strand/chem
Contig2	(consensus)	44437–44439	base quality below threshold

I ordered four additional reads for round 3 in hopes of resolving the 530 base pair region represented by the orange box: two primers on each side of the low quality region to provide ample coverage

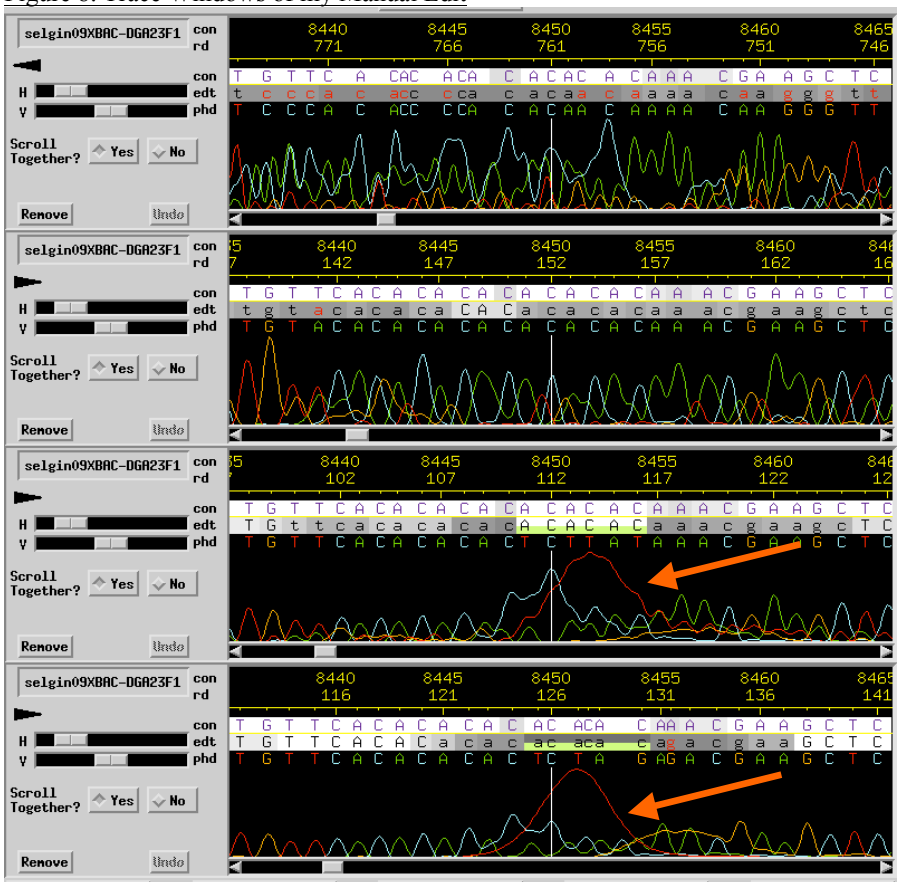
Round 3 Results:

Once the Round 3 data has been incorporated, all the problem areas have been resolved. There are no more high quality discrepancies and no more low quality bases. I ran the consensus sequence through a BLAST search to check for any contamination from the cloning source. No results appeared meaning my sequence was free from any foreign DNA. There are three interesting regions of mononucleotide repeats over 15 base pairs

long, two of which are A regions with the other being a C region. Although read quality usually breaks down after these regions, there are no surrounding low quality bases.

There was one six base pair AC repeat region located at 8449-8454 that was under the Phred threshold. But by examining the trace windows (Figure 8), it is apparent that the bottom two reads have a T dye blob that threw the sequence off, while the underlying peaks show the ACACAC pattern that agrees with the consensus. I manually edited this region and the consensus changed to high quality.

Figure 8. Trace Windows of my Manual Edit

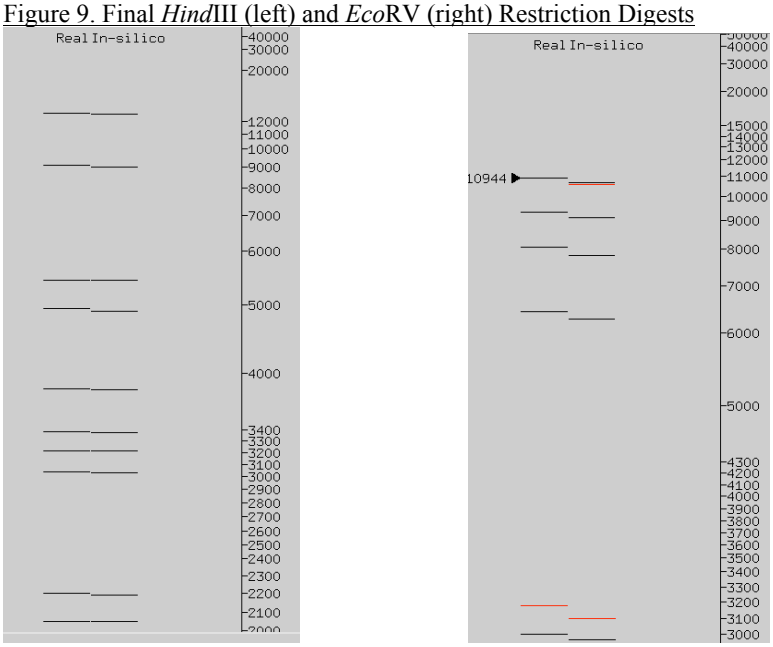


In addition to checking for mononucleotide runs, I also scanned the entire fosmid and discovered there are no Xs or Ns present. Xs would signify that vector sequence had been incorporated into the sequence and Ns represent an unknown nucleotide.

The final digests (Figure 9) look as though my fosmid is correctly assembled. There are no discrepant bands in the *Hind*III digest and the *Eco*RV *in-silico* digest looks very

similar to the real digest, just shifted slightly. This slight shift in the *Eco*RV digest may be a systematic error in the digest process.

Figure 9. Final *Hind*III (left) and *Eco*RV (right) Restriction Digests



Conclusion:

I started with four contigs, separated by three gaps and widespread poor read depth and quality, but after incorporating three rounds of new sequences, my fosmid appears to be finished. Figure 10 shows that the assembly view appears to have sufficient read depth and high quality bases spanning the whole contig. Due to a slight misunderstanding there are three single subclone regions that need further sequencing data. Simply calling a few additional reads would solve this problem rather easily. Besides that slight problem, I have fixed all the gaps and resolved every region to an above-threshold quality score.

Figure 10. Final Assembly View