Finishing *Drosophila grimshawi* Fosmid: DGA43A19

Varun Sundaram

2/16/09

Abstract

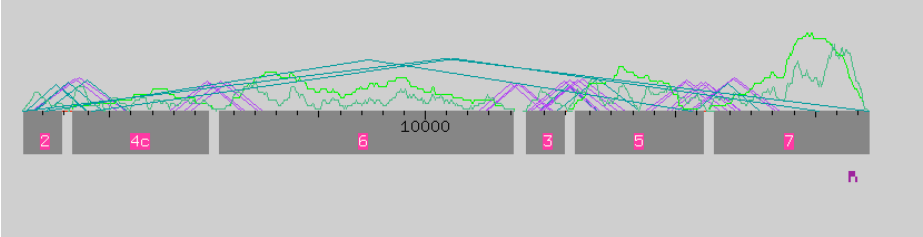
My project focused on the fosmid clone DGA43A19. The main problems with the fosmid were the five gaps. Using Consed, Phredphrap, and Autofinish, I was able to resolve the gaps by ordering reactions specific to these areas. After ensuring the fosmid was in one contig I focused on resolving the low quality areas and high quality discrepancies. The high quality discrepancies all turned out to be mis-called bases that needed editing. My final contig has some single strand/single chemistry regions; however, these regions are all above Phred score of 30 and further reactions are unnecessary. There is one single subclone region left unfinished, which needs further resolution.

Introduction

The purpose of the overall project is to finish and annotate the dot chromosome (chromosome 4)of the *Drosophila grimshawi.* This chromosome is primarily heterochromatic. After the fosmids are finished and annotated, analysis of the data should provide some important information on the differences between heterochromatin and euchromatin by comparing the finished regions to known euchromatic regions. Annotation allows for an in depth comparison of the genetic material to other annotated genes or genomes. Heterochromatin packaging is an area of interest because of its relationship to transposable elements, repeats, and silenced genes.

Finishing of Fosmid DGA43A19

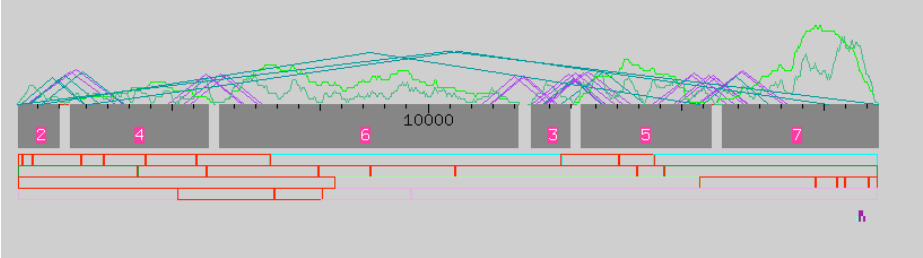
Figure 1 Initial Assembly view



My initial assembly *in-silico* digest did not match the real digest, and I attributed the problem to the fact that contig 4 was complemented in the assembly. Un-complementing the 4th contig resolved the major issues with the digest.

Searching for the paired-end reads dga43a19.b1 and dga43a19.g1 identified the ends of the assembly. These paired end reads were used because they spanned the furthest region of the assembly (as indicated by the large triangles). Using Consed I was able to search for these paired end reads and tag the ends of the individual contigs as the end of my fosmid.

Figure 2 Assembly View (un-complemented contig 4) with restriction sites

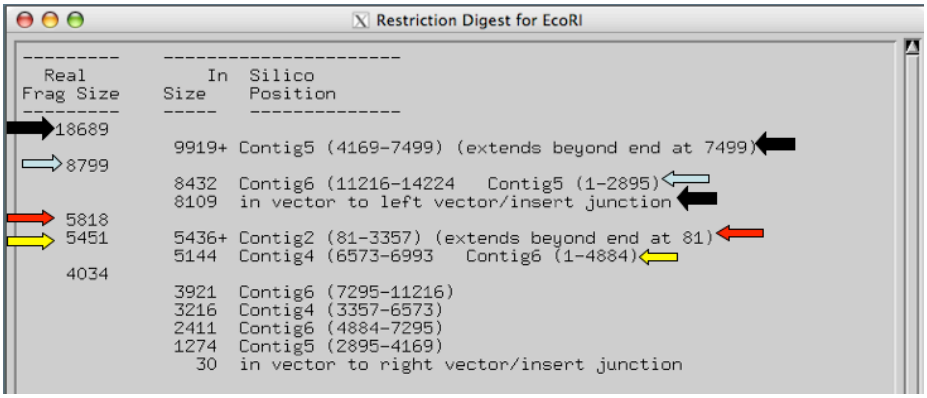


Based on the new assembly, the size of the gaps between the contigs could be estimated to be about 300bp based on the difference between band 2 (blue) *in silico* and band 2 (blue) in the Digest (Fig 3). A similar process was used to estimate the sizes of gaps 2-4, 4-6, and 5-7 to be 400bp, 300bp, and 500bp respectively. For the 5-7 gap the vector has to be added to the gap band because the real digest also contains vector.

To ensure that the gaps required additional reads I searched for potential overlaps in the gaps, which could be manually joined. After searching the ends of the contigs for matches, there were no forced joins to be made, and this was verified by the data from the digest, which indicated that the gaps were at least 300bp.

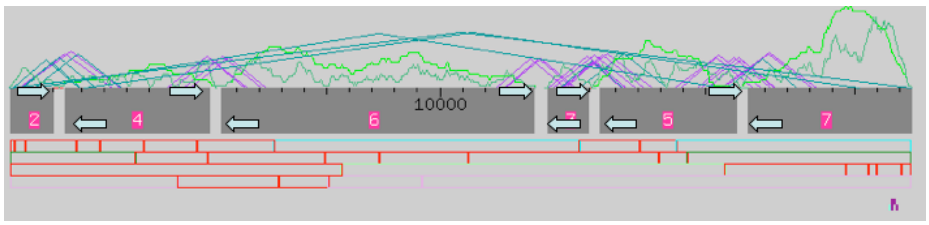
Top and bottom strand primers were designed for both ends in order to cover the entire gap (Figure 4). For all primers, all 3 sequencing chemistries were used in the project because there was no reason to suspect that using only 4:1 chemistry would yield superior results.

Figure 3 Restriction Digest EcoRI



Black→ gap 5-7
Blue→ gap 3-6 and gap 3-5
Red→ gap 4-2
Yellow→ gap 4-6

Figure 4 Assembly view with oligos designed to cover gaps.



After designing primers for the gaps, the navigator function of Consed was used to identify the low quality areas. Low quality areas were defined to be a Phred score lower than 30, or lower than 25 if covered by multiple strands/sequencing chemistries. Because the low quality areas were typically between five and twenty bases, only one primer was called to cover the regions (Red arrows in Figure 5).

For one particular low quality region (figure 6a-b) there was a repeat region upstream. Therefore, the primer chosen was a bottom strand primer downstream to the problematic area. Had an upstream primer been chosen it is possible the DNA polymerase would have encountered problems, due to the extensive repeat segment, and the read would not be complete.

Figure 5 Assembly view with Primers for low quality regions identified

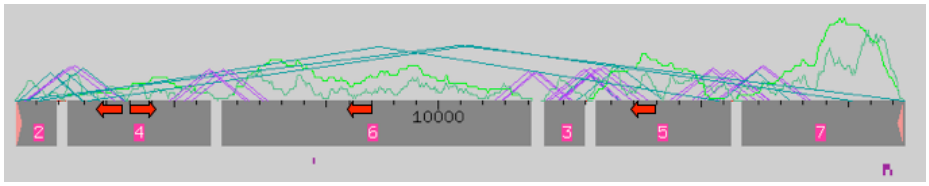
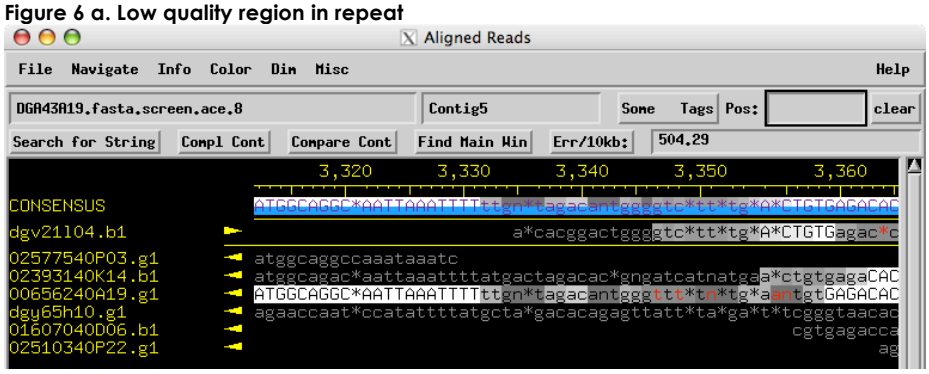
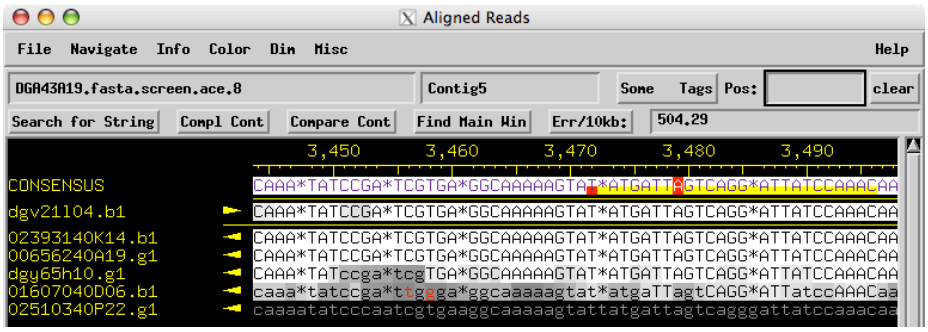


Figure 6 a. Low quality region in repeat



6 b. Primer designed to resolve low quality repeat region



Finally the high quality discrepancies needed to be resolved (Figure 7a). All the high quality discrepancies were caused by bases incorrectly interpreted in the sequence, which were easily fixed by editing the bases. After locating the high quality discrepancies the trace window was opened and the base calls were examined. It is evident in figure 7b that the pad placed in the third read is a mis-call, so I re-labeled it as low quality. A similar procedure was used for all other high quality discrepancy edits (figure 7c).

Office 2004 Test Drive…, 3/26/09 1:29 PM
Comment [1]: Comment about commas – after long introductory phrases (longer than five words) or in a compound sentence in which one or both or the sentences are long, it helps to have commas in those situations

Figure 7 a. High Quality Discrepencies

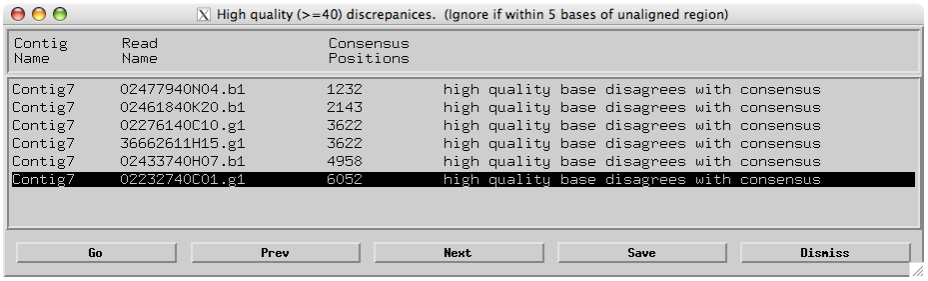
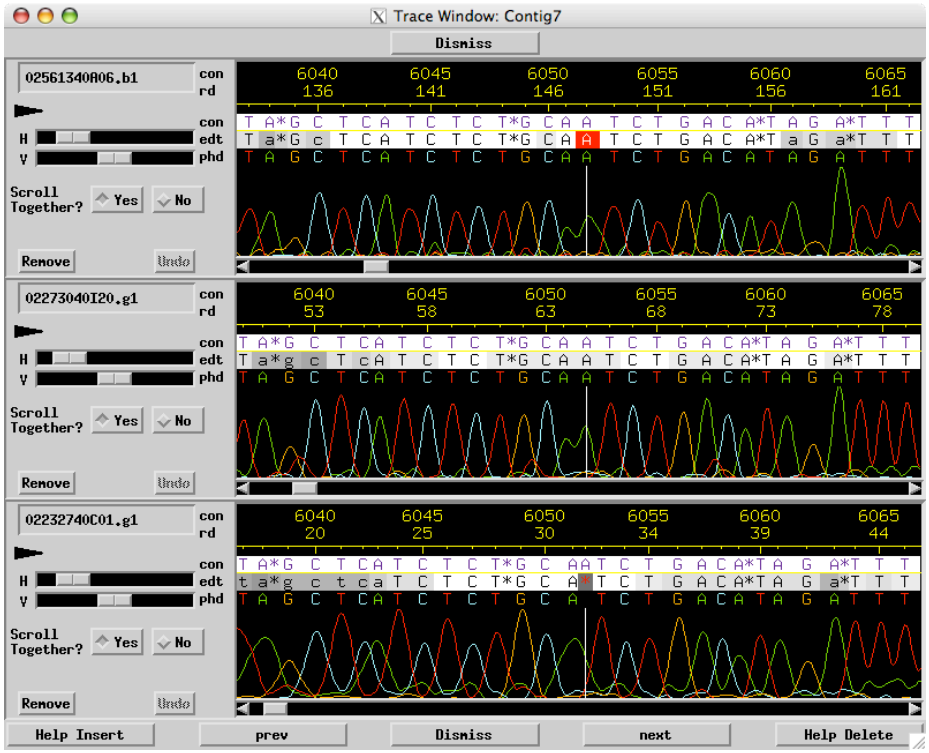


Figure 7 b. Trace window of a High Quality Discrepancy.



7 c. All high quality discrepancies resolved by editing during the project (positions refer to the final contig)

Contig Name	Read Name	Consensus Positions	Tag Length	Comment
Contig2	selgin09XBAC-DGA43A19_11.b1	4191		1 edit
Contig2	selgin09XBAC-DGA43A19_g23.b1	4879-4880		2 edit
Contig2	02314640E13.g1	8929		1 edit
Contig2	02559940P03.b1	13868		1 edit
Contig2	02559940P03.b1	13869		1 edit
Contig2	02329240D14.b1	14269		1 edit
Contig2	02577540P03.b1	25837		1 edit
Contig2	02579540E17.g1	26622		1 edit
Contig2	02477940N04.b1	34200		1 edit
Contig2	02461840K20.b1	35111		1 edit
Contig2	02276140C10.g1	36590		1 edit
Contig2	02433740H07.b1	37926		1 edit

Since there were no high quality discrepancies that needed resolution via additional reads, the only primers designed were for low quality regions or gaps. A comparison between the Autofinish primers and the primers chosen by examination of the assembly is shown in figure 8. All primers designed by Autofinish, for the first round, covered the same sections intended by the manual selection of primers or unnecessary segments of the fosmid such as the ends or areas already covered by another primer. Autofinish created a primer to resolve the low quality region at the end of a repetitive section upstream to the repetition. However, calling the primer downstream of the repetitive section ensures a higher likelihood of success due to the fact that traversing a long repetitive sequence can cause a read to fail. There were some other instances where Autofinish called a primer too far away from a gap to adequately resolve it, and the manual primers offered a better chance of success.

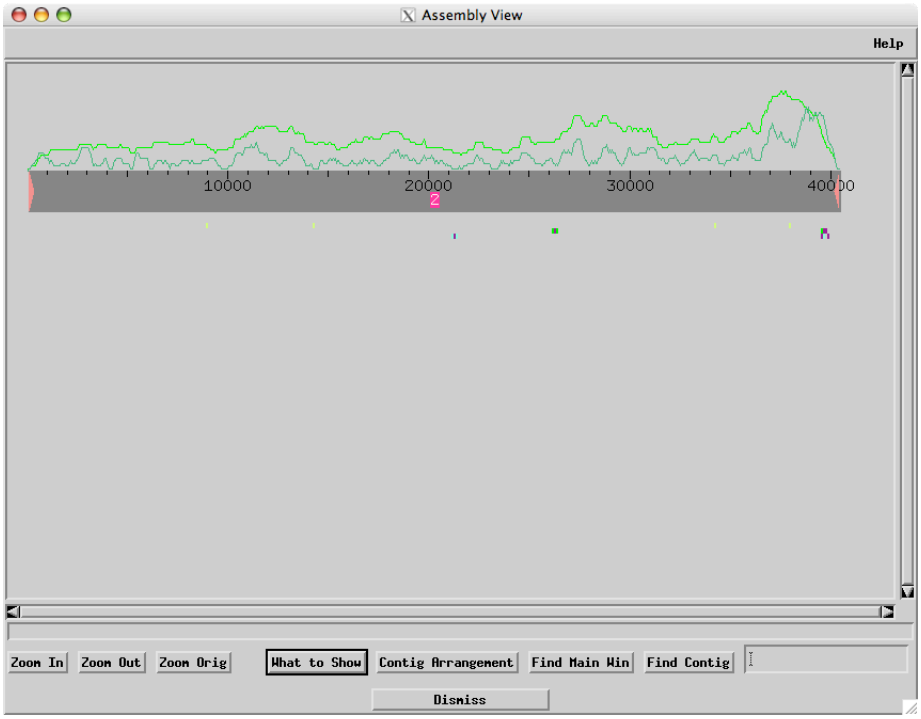
Examination of the problematic regions gave no reason for using just 4:1 chemistry as opposed to all three sequencing chemistries (Big-Dye, 4:1, and DGTP). Additionally due to the low cost of using all three chemistries and the high cost of ordering a new primer, it was deemed prudent to ensure at least one successful read for each primer ordered by using all three chemistries. As a result there was at least one successful read for each primer ordered. DGA43A19_t1, a poor read, was incorrectly incorporated into the beginning of the contig; however, this did not affect the consensus. DGA43A19_t4, DGA43A19_t11, DGA43A19_t13, DGA43A19_t14, were failures and no read was obtained. The failure of these 4:1 reads supports the reasoning to order all three chemistries. If these primers were ordered with just 4:1 chemistries, five primers would have been a complete waste and new primers would have been required.

Figure 8. Autofinish vs. manual primer selection, colored squares indicate similarities between manual calling and Autofinish. White boxes indicate superfluous primers called by Autofinish.

Name	Contig	Oligo	R base	Reason
Autofinish 1	2	aatcgcagcgcaaaa	154	resolve LQ at beginning of contig
Autofinish 2	2	ttttatgcagagcaaatcaaa	800	resolve 2-4 gap topstrand
Autofinish 3	2	aaaacatcaatatatatggcgaata	2357	resolve 2-4 gap topstrand
Autofinish 4	2	agacgtgttaagacatgttcact	2961	resolve 2-4 gap topstrand
Autofinish 5	3	caaaatgattgtgttgataaggtc	217	resolve 3-6 gap bottomstrand
Autofinish 6	3	tctgttcccaaaagcga	2373	resolve 3-5 gap topstrand
Autofinish 7	3	acgtgagcaataagttacatttt	2891	resolve 3-5 gap topstrand
Autofinish 8	4	acagcaatggtgatgtaagttat	376	resolve 2-4 gap bottomstrand
Autofinish 9	4	aattcacctgtcaggctctt	4586	resolve single subclone region
Autofinish 10	4	ctgcttcaatttacaaacttaaaa	5521	resolve single subclone region
Autofinish 11	4	acgcgactcgatcatatatact	7856	resolve 4-6 gap topstrand
Autofinish 12	5	tctttatgccagccatacttat	129	resolve 3-5 gap bottomstrand
Autofinish 13	5	tgatcctccgctcctc	3747	resolve LQ near repeat topstrand
Autofinish 14	5	agctatgagacaacatagcgtat	7388	resolve 5-7 gap topstrand
Autofinish 15	6	cgaatcgtaaacatttcgtc	156	resolve 4-6 gap bottomstrand
Autofinish 16	6	ccaaggaagcctccg	1036	LQ region topstrand
Autofinish 17	6	ttatatctacaacaaaacgtgcata	5266	resolve single subclone region
Autofinish 18	6	gcgagaggagagaacatgg	6399	resolve single subclone region
Autofinish 19	6	aataaataattcatcagcttctcg	12580	resolve single subclone region
Autofinish 20	6	catttatcaaacgcaaccc	14335	resolve 3-6 gap topstrand
Autofinish 21	6	acttaatgaacaaaataccaacaac	14976	resolve 3-6 gap topstrand
Autofinish 22	7	tctcagctgaattgctttgt	96	resolve 5-7 gap bottomstrand
Autofinish 23	7	ccaccccgactaaataatgt	8363	resolve LQ at end of contig
Name	Contig	Oligo	R base	Reason
XBAC-DGA43A19_2	2	tcatgctttaatgtagacgtgtta	2014	resolve 2-4 gap topstrand
XBAC-DGA43A19_6	3	agttttaaggcttgatgagttaaa	321	resolve 3-6 gap bottomstrand
XBAC-DGA43A19_7	3	cattaatttaaacgtgagcaataag	1948	resolve 3-5 gap topstrand
XBAC-DGA43A19_1	4	ttcgatgtgcgaggc	262	resolve 4-2 gap bottomstrand
XBAC-DGA43A19_11	4	ccttatccaaagcaatgattt	1899	resolve LQ region
XBAC-DGA43A19_12	4	gaacaggaggccgaaat	2774	resolve LQ region
XBAC-DGA43A19_4	4	cgacagcaatggtgatgtaag	6613	resolve 4-6 gap topstrand
XBAC-DGA43A19_8	5	gcataaagaatcgtgaaacg	162	resolve 3-5 gap bottomstrand
XBAC-DGA43A19_14	5	tatgattagtcaggattatccaaac	3496	resolve LQ region
XBAC-DGA43A19_9	5	acgattggctaatgcgta	6381	resolve 5-7 gap topstrand
XBAC-DGA43A19_3	6	gggacgaaatgtttacgatt	174	resolve 4-6 gap bottomstrand
XBAC-DGA43A19_13	6	tcactgtgtaactgtttctgtctc	6480	resolve LQ region
XBAC-DGA43A19_5	6	ttattgcggaactcgtattacta	14017	resolve 3-6 gap topstrand
XBAC-DGA43A19_10	7	ataactcgatcggtgtacatataaa	340	resolve 5-7 gap bottomstrand

Adding the files to the edit_dir and chromat_dir directories and re-running PhredPhrap incorporated the reads from these primers. The resulting assembly was composed of a single contig (see figure 9).

Figure 9. Assembly after round 1 reads were incorporated



This new assembly had some high quality discrepancies which were easily resolved and then three low quality areas which needed additional reads. Additionally primers were designed for some single subclone regions. A list of second round primers is given in figure 10. All three chemistries were used again based on the same reasoning as in round one.

Figure 10 round 2 primers

2nd round					
Name	contig	Oligo	R base	Reason	failed reads
XBAC-DGA43A19_17	2	tgcacatactcgtaactaacgaa	13975	topstrand single subclone 14870	bigdye fail
XBAC-DGA43A19_18	2	cgatcaacaaattaccagattatt	21011	topstrand single subclone 21770	
XBAC-DGA43A19_19	2	aaatggctgtccaactacct	22929	topstrand single subclone 23130	
XBAC-DGA43A19_21	2	gaggatgaacatcatggttg	4004	topstrand low quality 4140	
XBAC-DGA43A19_23	2	aactccttcatctacctaccaat	5497	bottomstrand low quality 5230	bigdye fail
XBAC-DGA43A19_24	2	cctggtccaaattctatgtga	29443	topstrand low quality 29630	
XBAC-DGA43A19_25	2	gtgatgtggggacgaaatgtttacgatt	9589	topstrand single subclone 9670	
XBAC-DGA43A19_26	2	cgttcttatgtatgcaacgagt	14100	topstrand single subclone 14330	
XBAC-DGA43A19_28	2	tgtggaagtatgcgcg	21807	bottomstrand single subclone 2490	

After incorporating the second round of reads, the low quality areas were completely resolved. Some high quality discrepancies were created but easily resolved by editing. The final list of problems is shown in figure 11. All of the single strand or single subclone regions indicated in the navigator window possess a Phred score above 30 and therefore do not require additional reactions. These have been tagged in the final ace file as "single subclone but above Phred 30." The unaligned high quality sequence is a low quality read which is slightly offset from the consensus. The trace window for this read is

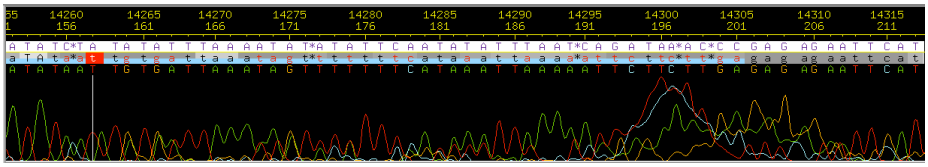
shown in figure 12, which demonstrates the read's low quality and allows for this problem to be tagged in the final ace file. Since this is a low quality read which is causing the mis-alignment there is no reason to suspect problems with the consensus. Had this read been a high quality read, there could have been a potential mis-assembly or polymorphic region in the fosmid.

Contig 1 still contained a single read around 830 base pairs, which partially aligned at the right end of the assembly. This is because the read in contig 1 is an extremely low quality read. It is not problematic that it does not fit into the assembly, especially because its forward read does not even exist in the sequence of this assembly. The ends of our fosmids will be improved when overlapping fosmids are assembled later in the project.

Figure 11. Final list of problems for assembly

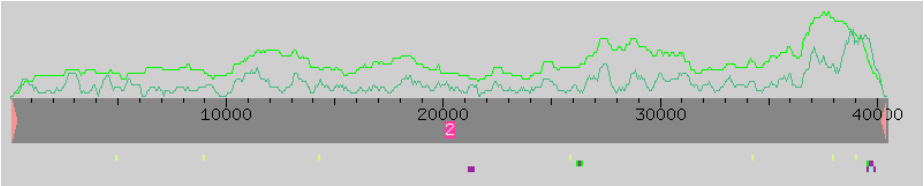
Contig Name	Read Name	Consensus Positions	
Contig2	(consensus)	1144-1311	170 bp single strand/chem
Contig2	(consensus)	1947-2038	95 bp single strand/chem
Contig2	(consensus)	5442-6198	791 bp single strand/chem
Contig2	(consensus)	7396-7456	62 bp single strand/chem
Contig2	(consensus)	9563-9917	362 bp single strand/chem
Contig2	(consensus)	9835-9917	83 bp single subclone
Contig2	(consensus)	12647-13037	394 bp single strand/chem
Contig2	selgin09XBAC-DGA43A19_g25.b1	14238-14305	68 unaligned high quality
Contig2	(consensus)	16456-16610	158 bp single strand/chem
Contig2	(consensus)	19683-19706	24 bp single strand/chem
Contig2	(consensus)	21749-22109	368 bp single strand/chem
Contig2	(consensus)	26399-26826	438 bp single strand/chem
Contig2	(consensus)	29374-29472	99 bp single strand/chem
Contig2	(consensus)	32368-32440	73 bp single strand/chem
Contig2	(consensus)	34445-34573	133 bp single strand/chem

Figure 12. Unaligned high quality read



Conclusion

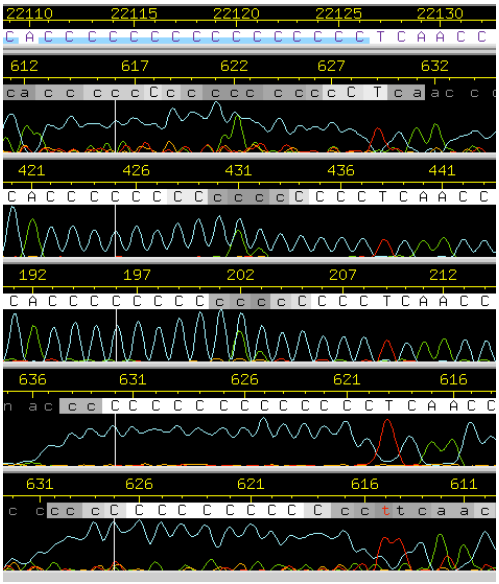
Figure 13. Final assembly view



The second round of reads was necessary to eliminate the last of the low quality areas. After resolving the contig minor editing was necessary to explain the high quality discrepancies. Many single strand regions with a Phred score above 30 were tagged but additional reads were not called to resolve them. However, there is one 83bp single sublclone at position 9835-9917 that needs to be resolved.

Finally there was a poly C run from 22111-22126. Examining the trace windows suggest that this is truly a poly C run due to the quality of the reads. However there is a suspicious A peak at bp 22120 (figure 13), which could be a potential polymorphism. This region was tagged as a poly C run.

Figure 13. Poly C run 22111-22126



To confirm the absence of vector from the sequence, the Basic Local Alignment Search Tool (BLAST) was used. The search compared the consensus sequence to known microbes. The BLAST results provided no matches, and therefore it is assumed that no bacterial contamination is present in the fosmid.

Another *in silico* digest was run to confirm the assembly and the results are shown in figures 14 (a-d). It is apparent that the assembly is correct from these digests. The

discrepancy with the final band in both digests can be explained by looking at the gel images. Both the HindIII and the EcoRI digests show a smear for the fourth band and in the case of HindIII there is actually a darker band, which appears further down on the gel (potentially explaning the 2192 band in the *in silico* digest). Based on this digest, and the overall high quality of read depth in the assembly. It is safe to assume that this fosmid has been adequately finished. However, fosmid DGA43A19 still remains to be annotated.

Office 2004 Test Driv…, 3/28/09 10:04 AM
Comment [2]: Reference figure

Figure 13 a. EcoRI final digest

Real Frag Size	In Silico Size	Position
18689	18153	part vector/part insert Contig2 (30448-40491)
8799	8531	Contig2 (20642-29173)
5818	5739	part vector/part insert Contig2 (1-5710)
5451	5384	Contig2 (8926-14310)
4034		
	3921	Contig2 (16721-20642)
	3216	Contig2 (5710-8926)
	2411	Contig2 (14310-16721)
	1275	Contig2 (29173-30448)

Figure 13 b. EcoRI gel image of 4034 band

Office 2004 Test Driv…, 3/28/09 10:03 AM
Comment [3]: Figures 13b and d need to have basepair numbers next to the bands so audience knows how big bands are

Figure 13 c. HindIII final digest

Real Frag Size	In Silico Size	Position
22667	21898	part vector/part insert Contig2 (18627-40491)
15913	15685	part vector/part insert Contig2 (1-7580)
4755	4713	Contig2 (7580-12293)
4151	4142	Contig2 (14485-18627)
	2192	Contig2 (12293-14485)

Figure 13 d. HindIII gel image of 4151 band

****

* GEP Finishing Checklist
****
Clone Name: DGA43A19
Student Finisher: Varun Sundaram

Goal: all sequence in one contig, all bases in the consensus having Phred >25, optimally both strands sequenced or two chemistries run on every region. For regions covered by a single clone, bases must have a Phred score >30. Assembly is confirmed by at least two restriction digests.

Project Status:

_V_ Completely finished
__ Finished except for questions regarding possible SNPs
__ Projects need more data to cover gaps
__ Not finished

Single Nucleotide Polymorphisms (SNPs)

_V_ Tagged all putative SNPs
_V_ Tagged all potential SNPs (e.g. unresolved High Quality Discrepancies)

Verify the Assembly

_V_ Project is in a single contig
_V_ Comment tags on any contigs over 2 kb that are not in the assembly
_V_ Cloning ends identified and tagged
_V_ Comment tag on any Assembly pieces used (fake reads)
_V_ Comment tag on any "PCR only" regions
_V_ Run BLAST (check for contamination from vector, host)

Use "search for string" to check the following:
_V_ mononucleotide runs (> 15 A's, >15 C's), scan traces to verify the consensus
of the region flanking the mononucleotide runs
_V_ X's in consensus (none should be present , indicates vector!)
_V_ N's in consensus (none should be present , ambiguous!)

Navigators

Check the following Navigators, please comment on any problems that are unresolved:

_V_ Low consensus quality
(Phred < 25 for double stranded regions, Phred < 30 for single stranded regions)
_V_ High quality discrepancies (use the navigator in the Aligned Reads Window,

not the Consed Main Window)
_V_ Single strand, single chemistry regions
_V_ Single subclone regions

Restriction Digests

Require at least two digests that match the in-silico digest. Please explain any discrepancies:
_V_ HindIII
_V_ EcoRI
__ EcoRV –incomplete digest
__ SacI -incomplete digest

Final Checks

_V_ Save final version with the suffix ace.0
_V_ Add comment tag at the left end of the contig with your name, professor and school