Finishing *Drosophila mojavensis* Fosmid Clone DMAC-1a

Charlie Manchee
Bio 434W
Dr. Elgin
3/25/10

Fosmid Clone DMAC-1a

Abstract

Bio 4342 students are currently working on finishing the *Drosophila grimshawi* dot chromosome and a euchromatic region of *Drosophila mojavensis*. Project DMAC-1a (from *D. mojavensis*) began in two contigs with only a single macroscale problem and a few microscale problems. Using Consed and phredPhrap, this project has been finished to completion and is now in a single contig. This paper presents the progress towards finishing the fosmid clone DMAC-1a to completion.

Initial Analysis

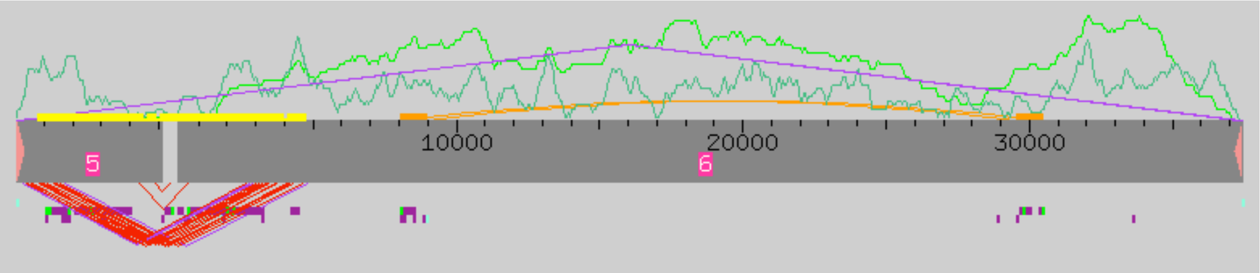
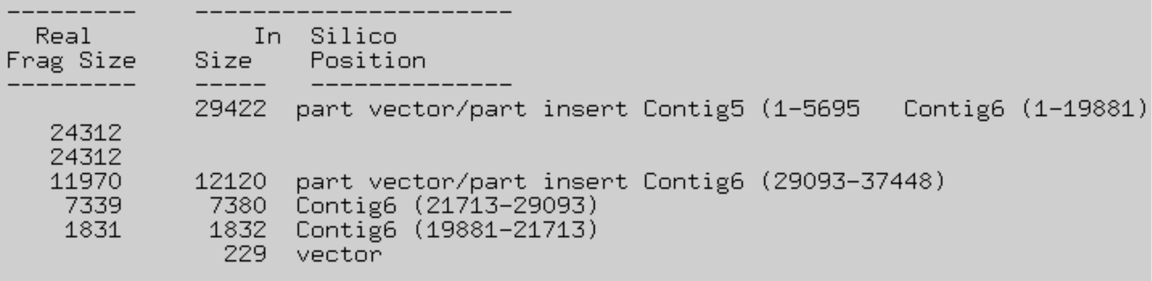


Fig. 1 Initial Assembly View.

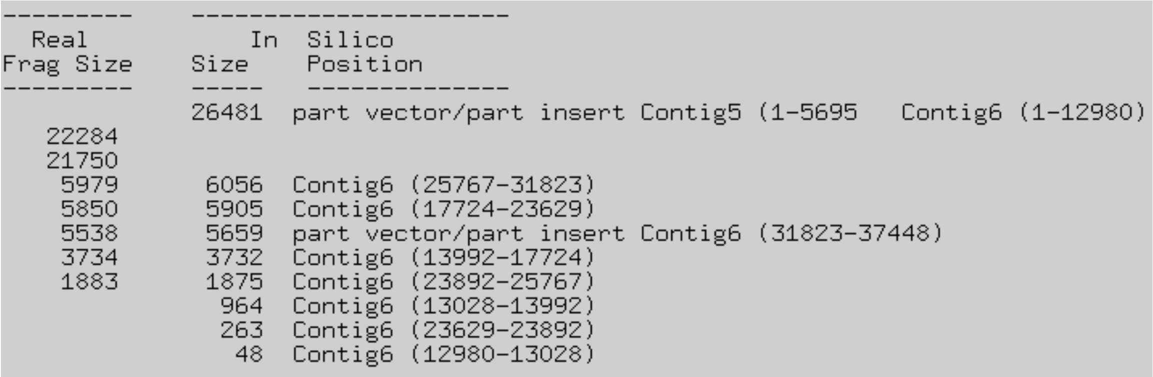
The initial assembly of the project contained two contigs: contigs 5 and 6. Many inconsistent forward/reverse pairs spanned this gap, but forward/reverse pairs were simply too far apart and were not in the wrong orientation. Other than the obvious problem of the gap, there were several high quality discrepancies and four low consensus quality regions.

After tagging the clone ends at the first high consensus base (27 and 38134) after the vector sequence (which was marked by Xs), running Crossmatch showed that the regions containing the inconsistent forward reverse pairs were also very similar sequences (as shown by the yellow bar in Fig. 1), suggesting the possibility that this region could actually be a single region. Before attempting to create a force join, I reviewed the restriction digests to see if the current assembly was larger than the real digest fragments (as it would be if the gap actually represented an overlap).



```
---------    ----------------------
  Real           In  Silico
Frag Size    Size    Position
---------    -----   ---------------
             29422  part vector/part insert Contig5 (1-5695   Contig6 (1-19881)
   24312
   24312
   11970     12120  part vector/part insert Contig6 (29093-37448)
    7339      7380  Contig6 (21713-29093)
    1831      1832  Contig6 (19881-21713)
               229  vector
```

Fig. 2 Text output for EcoRV digest of original assembly.



```
---------    ----------------------
  Real           In  Silico
Frag Size    Size    Position
---------    -----   ---------------
             26481  part vector/part insert Contig5 (1-5695   Contig6 (1-12980)
   22284
   21750
    5979      6056  Contig6 (25767-31823)
    5850      5905  Contig6 (17724-23629)
    5538      5659  part vector/part insert Contig6 (31823-37448)
    3734      3732  Contig6 (13992-17724)
    1883      1875  Contig6 (23892-25767)
               964  Contig6 (13028-13992)
               263  Contig6 (23629-23892)
                48  Contig6 (12980-13028)
```

Fig. 3 Text output for HindIII digest of original assembly.

Both EcoRV (Fig. 2) and HindIII (Fig.3) showed that the *in silico* fragment spanning the gap between contigs 5 and 6 was actually about 5000 bp larger than the real fragment from the digest performed on this fosmid. This suggested that there is, in fact, a 5000 bp overlap between contigs 5 and 6, justifying a possible force join between these contigs. After searching for string and aligning the sequences, I force joined contigs 6 and 7 to create contig 7 (Fig. 4).

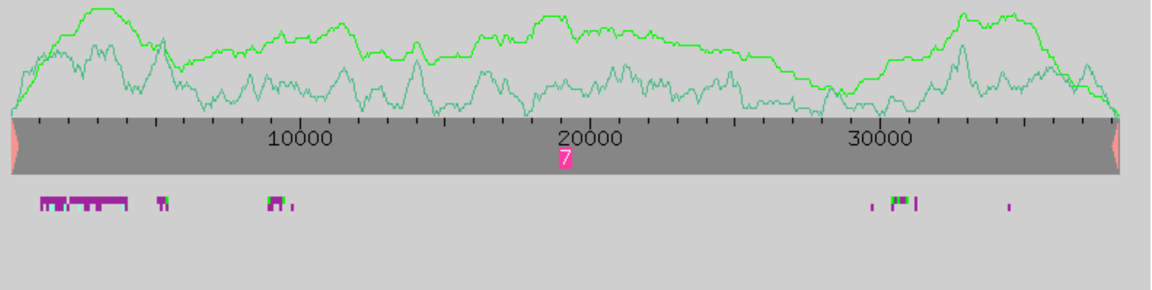
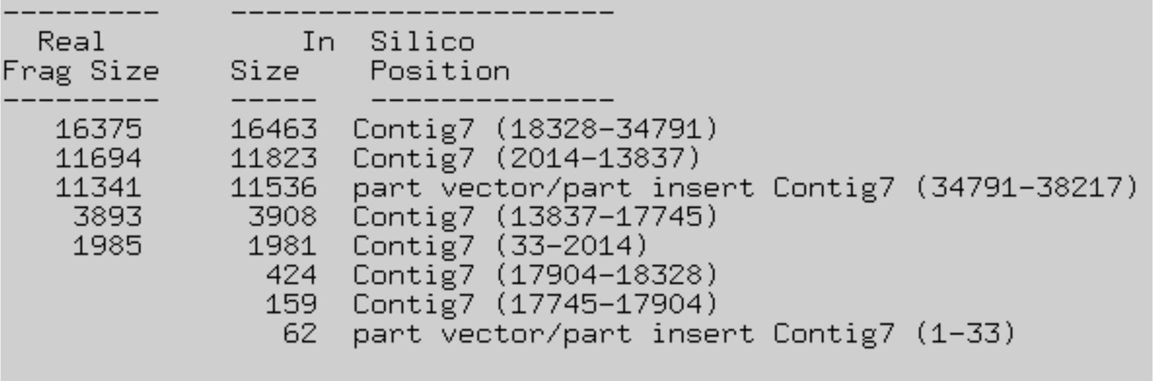
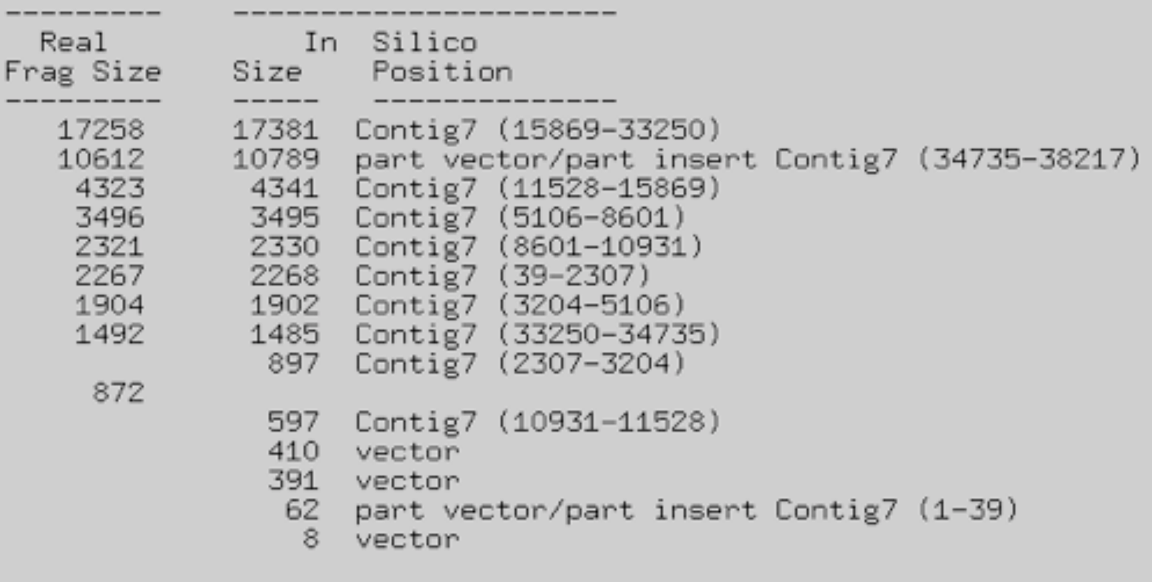


Fig. 4 Assembly view after joining contigs 5 and 6.



```
---------    ----------------------
  Real           In  Silico
Frag Size    Size    Position
---------    -----   --------------
   16375     16463  Contig7 (18328-34791)
   11694     11823  Contig7 (2014-13837)
   11341     11536  part vector/part insert Contig7 (34791-38217)
    3893      3908  Contig7 (13837-17745)
    1985      1981  Contig7 (33-2014)
               424  Contig7 (17904-18328)
               159  Contig7 (17745-17904)
                62  part vector/part insert Contig7 (1-33)
```

Fig. 5 Text output of EcoRI digest.



```
---------    ----------------------
  Real           In  Silico
Frag Size    Size    Position
---------    -----   --------------
   17258     17381  Contig7 (15869-33250)
   10612     10789  part vector/part insert Contig7 (34735-38217)
    4323      4341  Contig7 (11528-15869)
    3496      3495  Contig7 (5106-8601)
    2321      2330  Contig7 (8601-10931)
    2267      2268  Contig7 (39-2307)
    1904      1902  Contig7 (3204-5106)
    1492      1485  Contig7 (33250-34735)
               897  Contig7 (2307-3204)
     872
               597  Contig7 (10931-11528)
               410  vector
               391  vector
                62  part vector/part insert Contig7 (1-39)
                 8  vector
```

Fig. 6 Text output of SacI digest.

Contig 7 contained 600 reads and was about 38252 bp long. The new digests (Figs. 5 and 6) supported the join since all *in silico* strands larger than 1000 bp had a real fragment partner. However, to double-check this join I called two reactions (one on each end of the overlap region) (Fig. 10). Now that the obvious problem of the join was solved, it was time to solve the smaller order problems of the project.

The high quality discrepancies proved to be the most interesting. The assembly contained 46 such discrepancies, several of which spanned multiple bases. Twenty of the discrepancies were caused by five misaligned reads that were scattered throughout the assembly; removing these reads from the assembly solved these discrepancies. Five discrepancies represented wide peaks that were miscalled by the sequencer (Fig. 7) and one discrepancy resulted from a miscall caused by a slowly diminishing trace from the previous base overlapping the trace from the correct base (Fig. 8). Three discrepancies were created by growth differences in *E. coli*, as was determined by examining the traces and observing that the discrepancies' traces were high quality and normal (Fig. 9). The remaining seventeen discrepancies represented putative polymorphisms, which I determined to be different from growth difference because all of these discrepant bases were present in multiple reads at these locations.

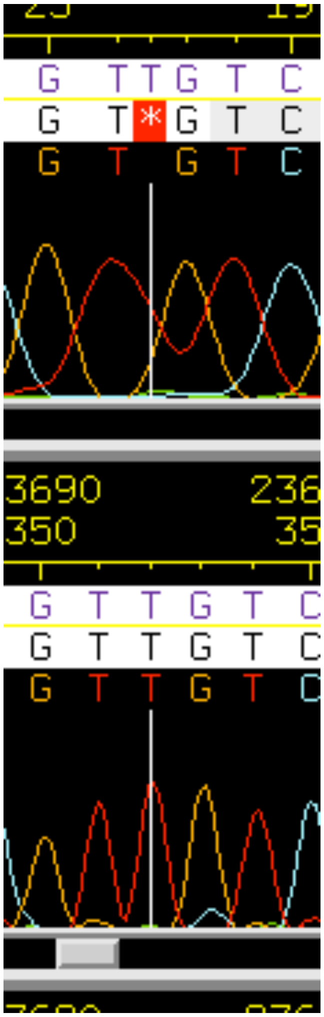


Fig. 7 Wide Peak.

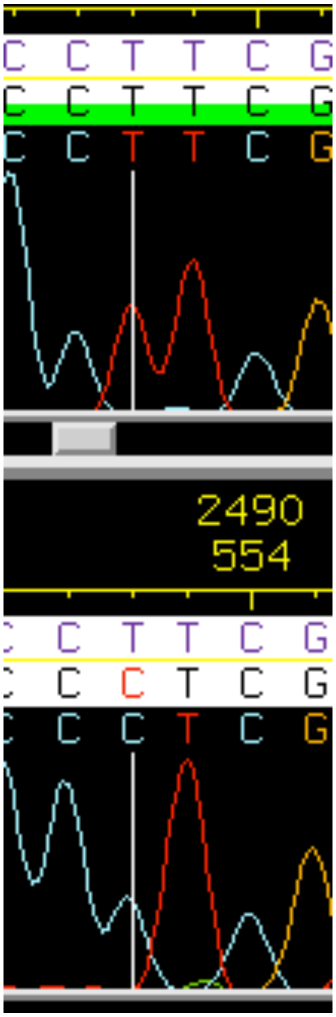


Fig. 8 Miscall.

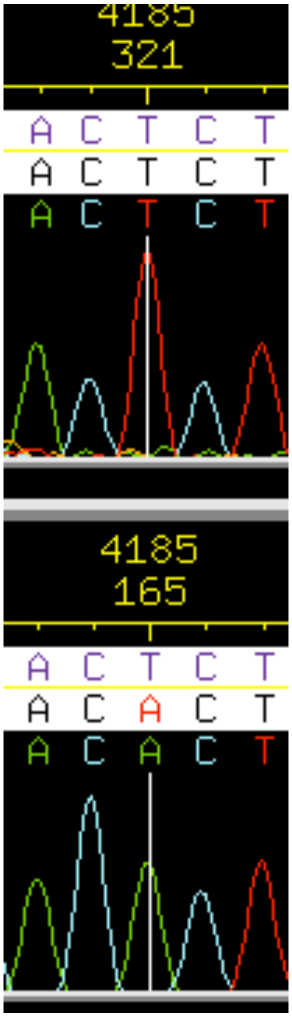


Fig. 9 Growth Difference.

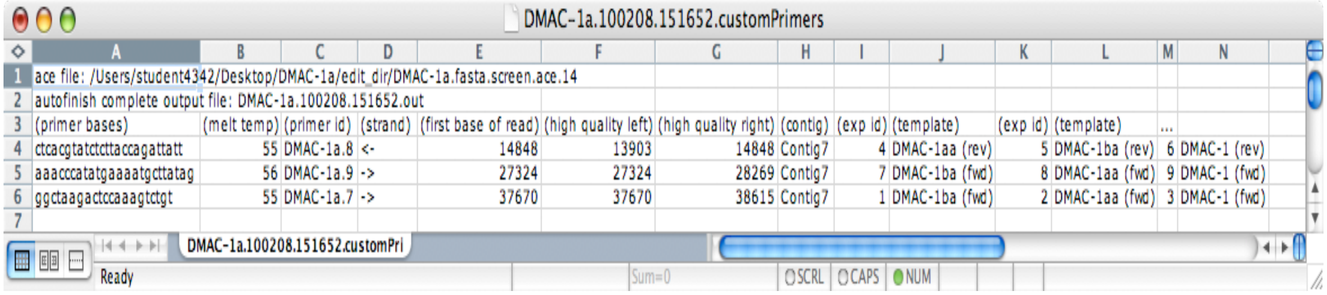
The polymorphisms were an interesting find since the fosmid is from a euchromatic region of *Drosophila mojavensis*. In a euchromatic region, very few polymorphisms were expected since the region is transcriptionally active and is therefore under evolutionary functional constraint. Also, all of the polymorphisms were found in the region that I had just force joined. Since the digests supported this join very well this join can be trusted despite the many polymorphisms. In fact, these polymorphisms were probably the reason phredPhrap did not initially join the region. However, it is possible that some of the discrepant reads in this region were actually reads from another, similar region of the genome that was placed in this assembly by mistake during the whole genome shotgun; this was not likely since the reads were spanned by forward reverse pairs and therefore must be very near each other in the genome. Another possibility is that this region should be a tandem repeat; this possibility warrants some consideration, but it is unlikely unless the repeat is small enough to not disrupt the digests. If the region does indeed contain this unusual amount of polymorphisms, it could mean that the region is an intron or a pseudogene that is not under any functional constraint.

I investigated the four low consensus quality regions: 1-25, 14651-14674, 29981-29983, and 38170-38252. Regions 1-25 and 38170-38252 represent the ends of the project. These ends were ignored since neighboring projects will overlap the ends of this fosmid. The remaining two regions, however, were in the middle of the project and therefore required attention. I called four oligos in total (one from each side of each low quality region) to resolve this problem (Fig. 10). All of the oligos from round one were called as 4:1 chemistry since there were no gaps in the assembly to be closed.

Primer	Primer pos.	Dir.	Reason	Autofinish	Success
Gcacggttagaagtggtaag	5465-5487	→	Check force join	None	Fail
Cagctcggctatgatgc	1052-1068	←	Check force join	None	Fail
Cgagtcgggtcgagtagtg	14195-14213	→	Low quality 14651-14674	None	Added, but didn't resolve problem
Tgatagatgttagctcacgtatctc	14897-14921	←	Low quality 14651-14674	Same region, different primer	Added, but didn't resolve problem
Cccagtccatctgtttcg	29791-29808	→	Low quality region 29981-29983	None	Added, but didn't resolve problem
Ggtgaggggtggatatagg	30068-30086	←	Low quality region 29981-29983	None	Added, but didn't resolve problem

Fig. 10 Table of first round primers with Autofinish comparison (Success based on re-run of plate with dGTP done during second week of project).

Comparison of 1st Round Primers with Autofinish


DMAC-1a.100208.151652.customPrimers
DMAC-1a.100208.151652.customPri
Ready
Sum=0
○SCRL
○CAPS
●NUM

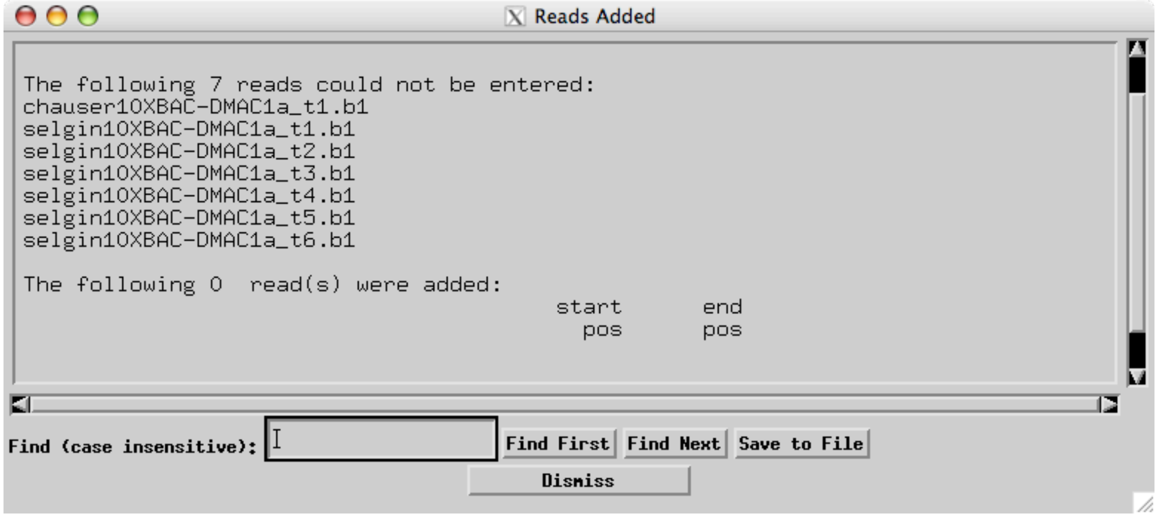
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ace file: /Users/student4342/Desktop/DMAC-1a/edit_dir/DMAC-1a.fasta.screen.ace.14													
2	autofinish complete output file: DMAC-1a.100208.151652.out													
3	(primer bases)	(melt temp)	(primer id)	(strand)	(first base of read)	(high quality left)	(high quality right)	(contig)	(exp id)	(template)	(exp id)	(template)	...	
4	ctcacgtatctcttaccagattatt	55	DMAC-1a.8	<-	14848	13903	14848	Contig7	4	DMAC-1aa (rev)	5	DMAC-1ba (rev)	6	DMAC-1 (rev)
5	aaacccatatgaaaatgcttatag	56	DMAC-1a.9	->	27324	27324	28269	Contig7	7	DMAC-1ba (fwd)	8	DMAC-1aa (fwd)	9	DMAC-1 (fwd)
6	ggctaagactccaaagtctgt	55	DMAC-1a.7	->	37670	37670	38615	Contig7	1	DMAC-1ba (fwd)	2	DMAC-1aa (fwd)	3	DMAC-1 (fwd)
7														

Fig. 11 Autofinish Primers.

Autofinish called only three primers as opposed to my six primers (Fig. 11). Autofinish's first primer was called to resolve a low quality region (14651-14674) that I called two primers (one from each direction) to resolve. Autofinish's second primer was called for a single subclone region that I had overlooked in my first analysis of the fosmid. The third Autofinish primer was designed to sequence a low quality region from 38170-38252, which I ignored because it is at the end of the project and will therefore be overlapped by neighboring projects. Overall, Autofinish missed one low quality region (29981-29983) for which I called primers and found a single subclone region which I had missed. I also called primers to double check the force join that Autofinish did not know had occurred (since I ran Autofinish on the consensus after I had made the join).

Analysis after Reads from First Week Primers Added

All of my reactions from week one used 4:1 chemistry initially since there were no gaps and there was no need to utilize more expensive resources. However, all of the reactions failed to give high quality data and no new reads were added (Fig. 12).



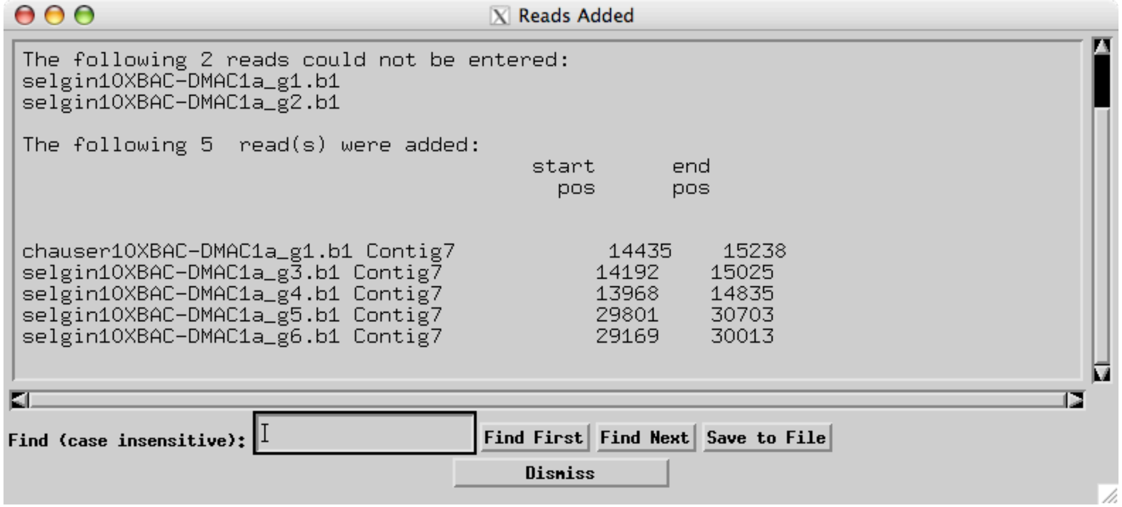
```
The following 7 reads could not be entered:
chauser10XBAC-DMAC1a_t1.b1
selgin10XBAC-DMAC1a_t1.b1
selgin10XBAC-DMAC1a_t2.b1
selgin10XBAC-DMAC1a_t3.b1
selgin10XBAC-DMAC1a_t4.b1
selgin10XBAC-DMAC1a_t5.b1
selgin10XBAC-DMAC1a_t6.b1

The following 0  read(s) were added:
                                        start      end
                                          pos      pos
```

Fig. 12 Reads added after week one primers (4:1 chemistry).

I did not call any new primers the second week since the first week primers were re-run with dGTP and there were no new reactions that would not be covered by these reactions.

Analysis after Reads from First Week Primers (dGTP) Added



```
The following 2 reads could not be entered:
selgin10XBAC-DMAC1a_g1.b1
selgin10XBAC-DMAC1a_g2.b1

The following 5  read(s) were added:
                                        start      end
                                          pos      pos


chauser10XBAC-DMAC1a_g1.b1 Contig7           14435    15238
selgin10XBAC-DMAC1a_g3.b1 Contig7           14192    15025
selgin10XBAC-DMAC1a_g4.b1 Contig7           13968    14835
selgin10XBAC-DMAC1a_g5.b1 Contig7           29801    30703
selgin10XBAC-DMAC1a_g6.b1 Contig7           29169    30013
```

Fig. 13 Reads added from first round primers (dGTP chemistry).

Using dGTP chemistry resulted in some success in adding reads to the correct position (Figs. 10 and 13). The reads designed to double check my force join failed, but since my digests are very strong support for this join, I will still trust the join. Also, since there were already many reads in the region, more reads would provide little additional information; thus I did not call any more reactions for this region. The rest of the reads were added, but they did very little to improve the quality of the regions for which they were designed since the reads were of fairly low quality.

After analyzing the new data provided by the reads, I designed new primers for the two low quality regions as well as the single subclone region identified by Autofinish previously. For these reactions I used all three chemistries (BigDye, dGTP, and 4:1) since this is the final round of calling reads (Fig. 14).

Primer	Primer pos.	Dir.	Reason	Chemistry	Success
Tgatagatgttagctcacgtatctc	14862-14886	←	Low Quality Region 14651-14674	BigDye	Added
				dGTP	Added
				4:1	Fail
cgtatatatagcgcagatatcgag	29847-29870	→	Low Quality Region 29981-29983	BigDye	Added
				dGTP	Added
				4:1	Added
gcttcatcggatcatgc	27555-27571	→	Single Subclone Region 27864-27880	BigDye	Added
				dGTP	Added
				4:1	Added

Fig. 14 Second Round of Primers.

Analysis after Reads from Second Week Primers

The single subclone region now has multiple high quality reads, so the region was resolved. The low quality region from 14651-14674 has been raised above a phred score of 30 at all bases, so this region is resolved. The last low quality region consists of only three bases (Fig. 15). The phred scores of the bases are 29, 28, and 28, respectively, and many traces show a very consistent pattern of bases, so I can trust the consensus in this region (Fig. 16).

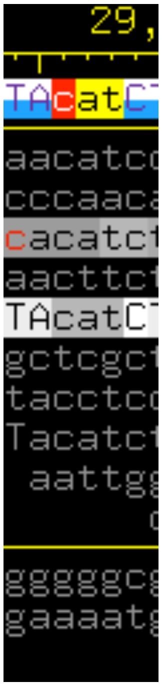


Fig. 15 Aligned Reads Window of Remaining Low Quality Region.

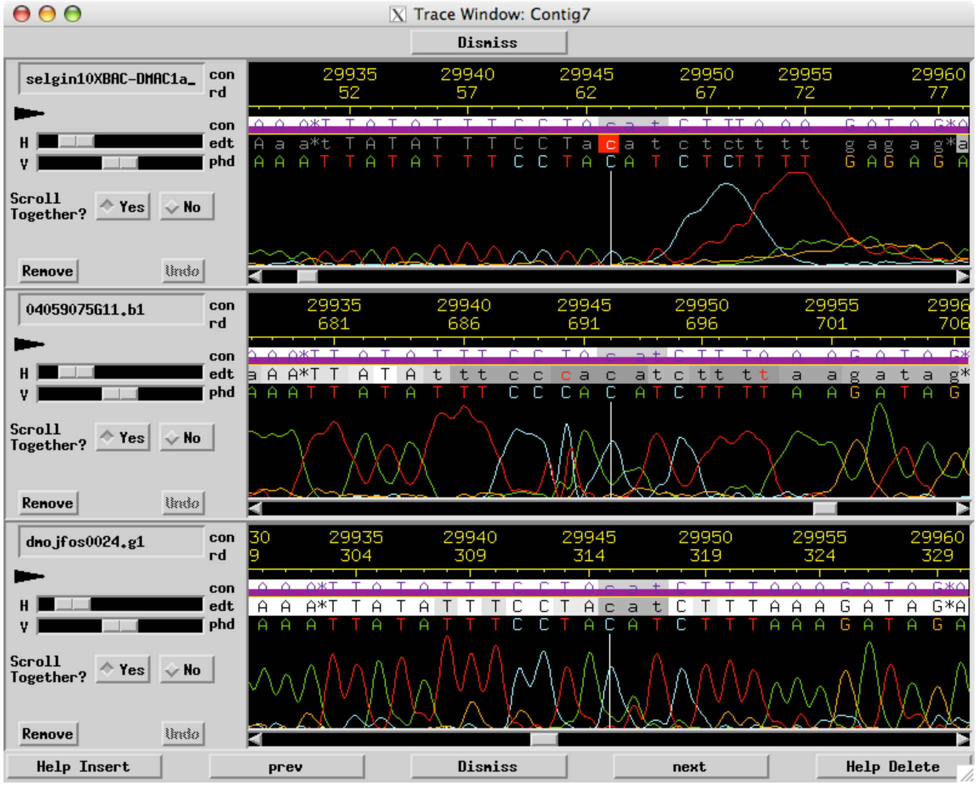
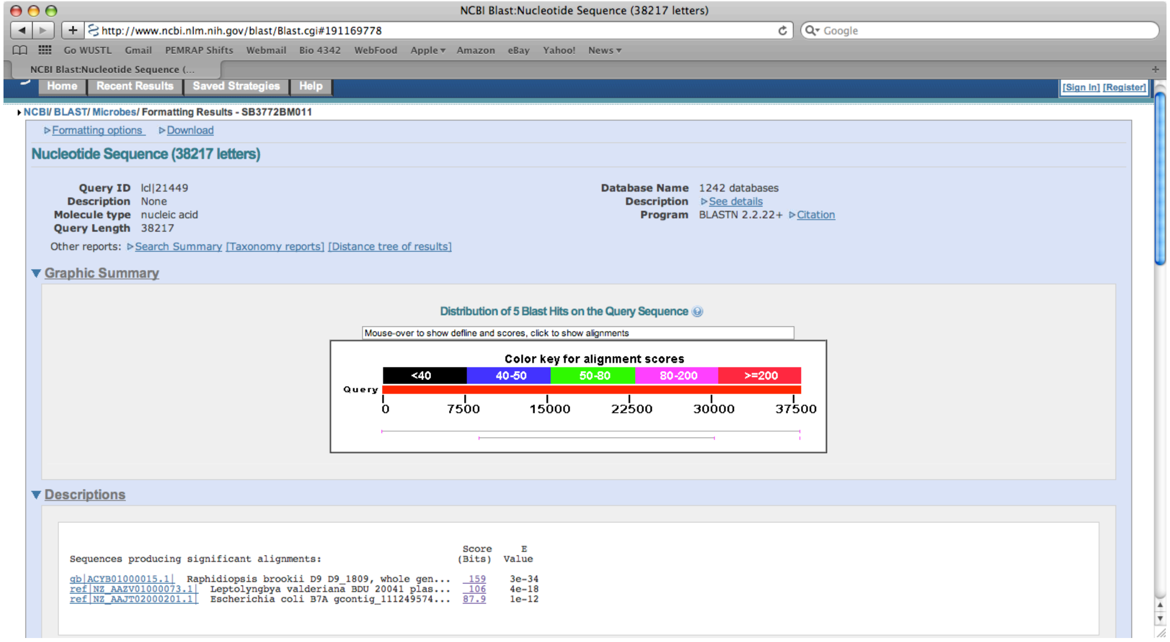


Fig. 16 Traces of Remaining Low Quality Region.

Lastly, I analyzed and tagged all single chemistry or single strand regions; these regions are all of high quality and therefore do not pose any problems. I found three mononucleotide runs of A, but the traces on both sides of these runs were very good, so the consensus can be trusted since it is obvious that the quality did not crash due to these runs. I also ran BLAST to determine whether any microbial DNA had contaminated my assembly (Fig. 17). BLAST found only weak, small matches to microbe DNA, none of which were both significant and long enough to warrant attention.



▸ NCBI/ BLAST/ Microbes/ Formatting Results - SB3772BM011

▷ Formatting options ▷ Download

Nucleotide Sequence (38217 letters)

Query ID	lcl|21449	**Database Name**	1242 databases
Description	None	**Description**	▷ See details
Molecule type	nucleic acid	**Program**	BLASTN 2.2.22+ ▷ Citation
Query Length	38217		

Other reports: ▷ Search Summary [Taxonomy reports] [Distance tree of results]

▼ Graphic Summary

Distribution of 5 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

▼ Descriptions

```
                                                          Score     E
Sequences producing significant alignments:               (Bits)  Value

gb|ACYB01000015.1|  Raphidiopsis brookii D9 D9_1809, whole gen...   159    3e-34
ref|NZ_AAZV01000073.1|  Leptolyngbya valderiana BDU 20041 plas...   106    4e-18
ref|NZ_AAJT02000201.1|  Escherichia coli B7A gcontig_111249574...  87.9    1e-12
```

Fig. 17 BLAST Results.

Final Analysis

The project was successfully completed. The final assembly (Fig. 18) contains only a single contig. All bases are above phred 30 except for three bases, which can be trusted based upon their traces, and there are no single subclone regions remaining in the problem. Remaining high quality discrepancies are most likely polymorphisms and have been tagged as such. The digests of EcoRI (Fig. 19) and SacI (Fig. 20) are consistent with my final assembly.

Further work on the clone could investigate the highly polymorphic region. It is rare to see such a high rate of polymorphism in a euchromatic region, since transcribed genes are generally under functional constraint. Such a high rate of polymorphisms may indicate that this region is not transcriptionally active and may represent a pseudogene or an intron. It is not likely that the polymorphic region actually contains reads mistakenly placed in my assembly during the whole genome shotgun since many forward/reverse pairs spanned this region.

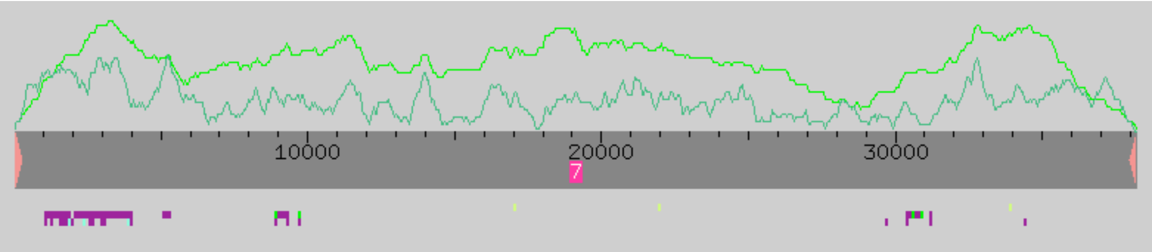


Fig. 18 Final Assembly View.

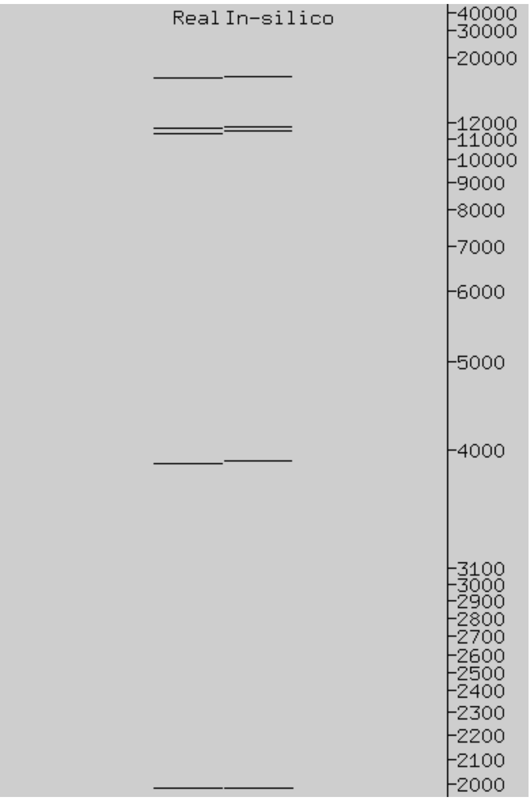


Fig. 19 EcoRI Final Digest.

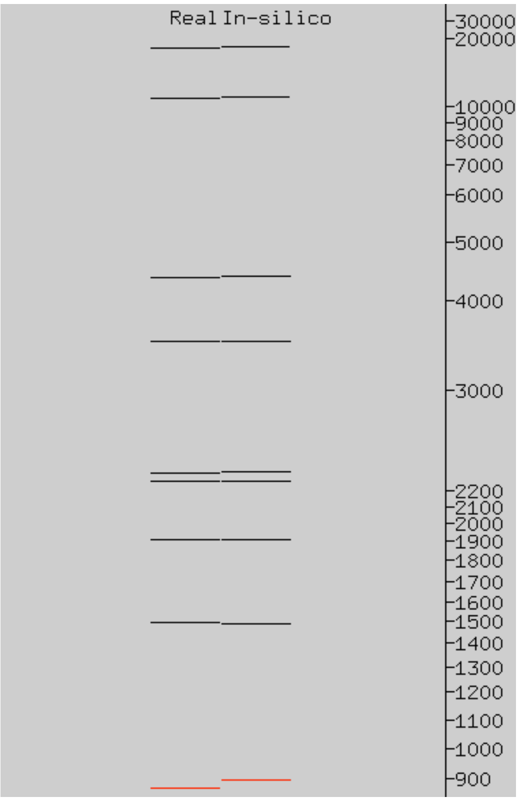


Fig. 20 SacI Final Digest.

Addendum

Project DGA15M06 has been finished to completion. The initial assembly consisted of only a single contig. There were no low quality regions or single subclone regions other than those at the end of the project and only twelve high quality discrepancies, all of which were actually bad traces. No additional reads were required for this project.

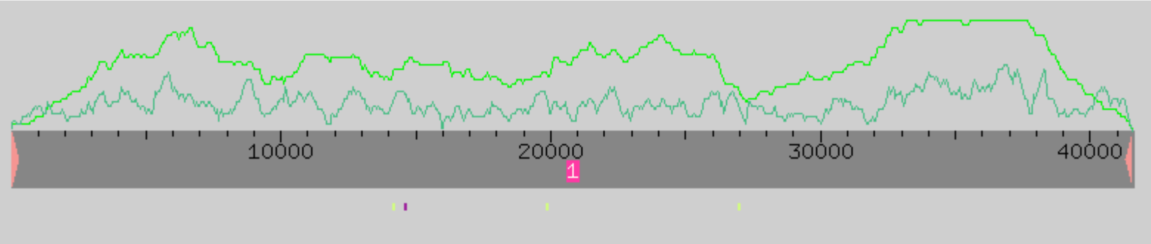


Fig. 21 Original Assembly View for DGA15M06.

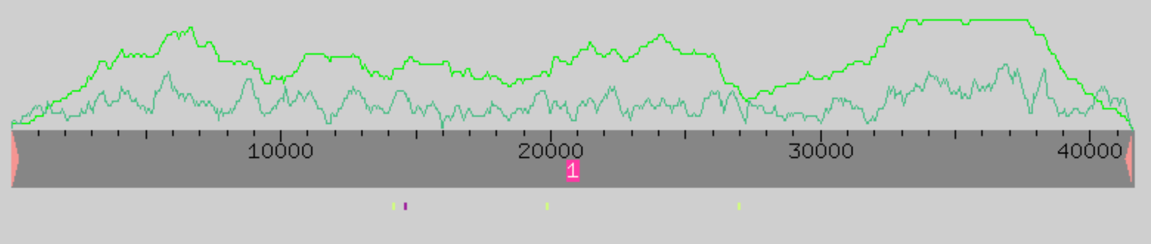


Fig. 22 Final Assembly View for DGA15M06.