Finishing Fosmid DMAC-27a
of the *Drosophila mojavensis* third chromosome

Ruth Howe
Bio 434W
27 February 2010

Abstract

The fourth or "dot" chromosome of *Drosophila* species is composed primarily of highly condensed, heterochromatic DNA rich in repeats, making it difficult to sequence and study. However, evidence that approximately 80 genes are expressed from the dot chromosome raises questions about how heterochromatin and silencing work in this unusual area and makes this sequence a valuable target for study of the development and function of heterochromatin architecture. As part of a larger effort to sequence the dot chromosomes of multiple *Drosophila* species for comparison, this project aims to finish a euchromatic control region of the *D. mojavensis* third chromosome carried on fosmid DMAC-27a. This report will detail the finishing process from the initial assembly of reads by *phred/phrap* to confirmation of the final assembly and a description of any remaining difficulties or areas of interest.

Initial Assembly and Challenges

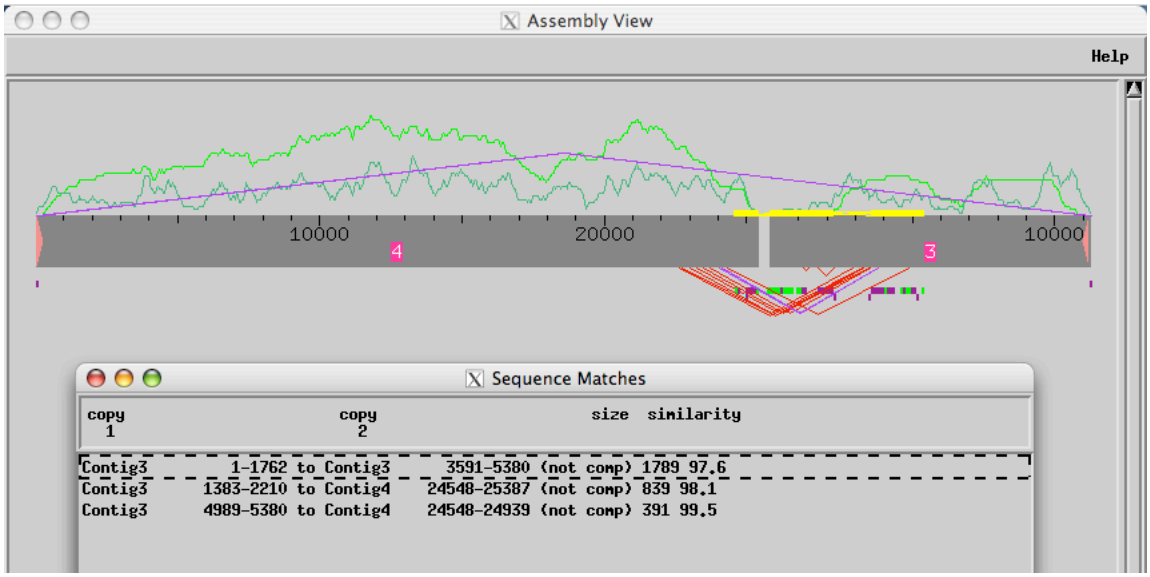


Fig. 1: the initial assembly view of DMAC-27a with list of tandem repeat regions (yellow boxes) generated by *crossmatch*. The depth of read coverage is represented by the light green lines, and high quality read coverage by the dark green lines. Red lines represent unresolved forward/reverse pairs. The pink arrows on either end of the contig represent the clone ends, which were arbitrarily defined as the first and last non-vector (as identified by phrap), high quality bases in the consensus.

The initial assembly of data by *phred/phrap* has several immediately evident problem regions, as can be seen in Figure 1. The depth of coverage crashes at the gap between contigs 3 and 4, which is surrounded by tandem repeat regions and unresolved forward/reverse pairs, suggesting a misassembly. Navigating to problem regions in the aligned reads windows revealed that the sequence also contains a few regions covered by only a single strand or chemistry, many high quality discrepancies, sequences below the consensus quality threshold of *phred* 30, and two contigs containing high quality reads that could not be incorporated into the assembly (Fig. 2).

ig. 2: the problem lists enerated for the initial ssembly (top) show large reas of single strand or ingle chemistry as well s low consensus quality egions and a large umber of high quality iscrepancies in contig 3. he traces for the reads ot incorporated into the ain assembly (bottom) how high quality reads nd a “match elsewhere” ag (teal highlight) in ontig 1. Note that *phred* as miscalled numerous ases in both reads of ontig 2, suggesting that anual editing could llow the reads to be ncorporated into the ain assembly.

In addition to the problems suggested by the assembly view, the restriction digests (eg. Fig. 3) suggested that there was too much data in the assembly, supporting the hypothesis of multiple misassemblies involving the tandem repeat regions. Notably, the errant fragment of the *in silico* digest linked the vector to two of the tandem repeat regions on contigs 3 and 4, suggesting that these apparent repeats in fact represented a single region that *phrap* had separated due to large numbers of single nucleotide polymorphisms (SNPs).

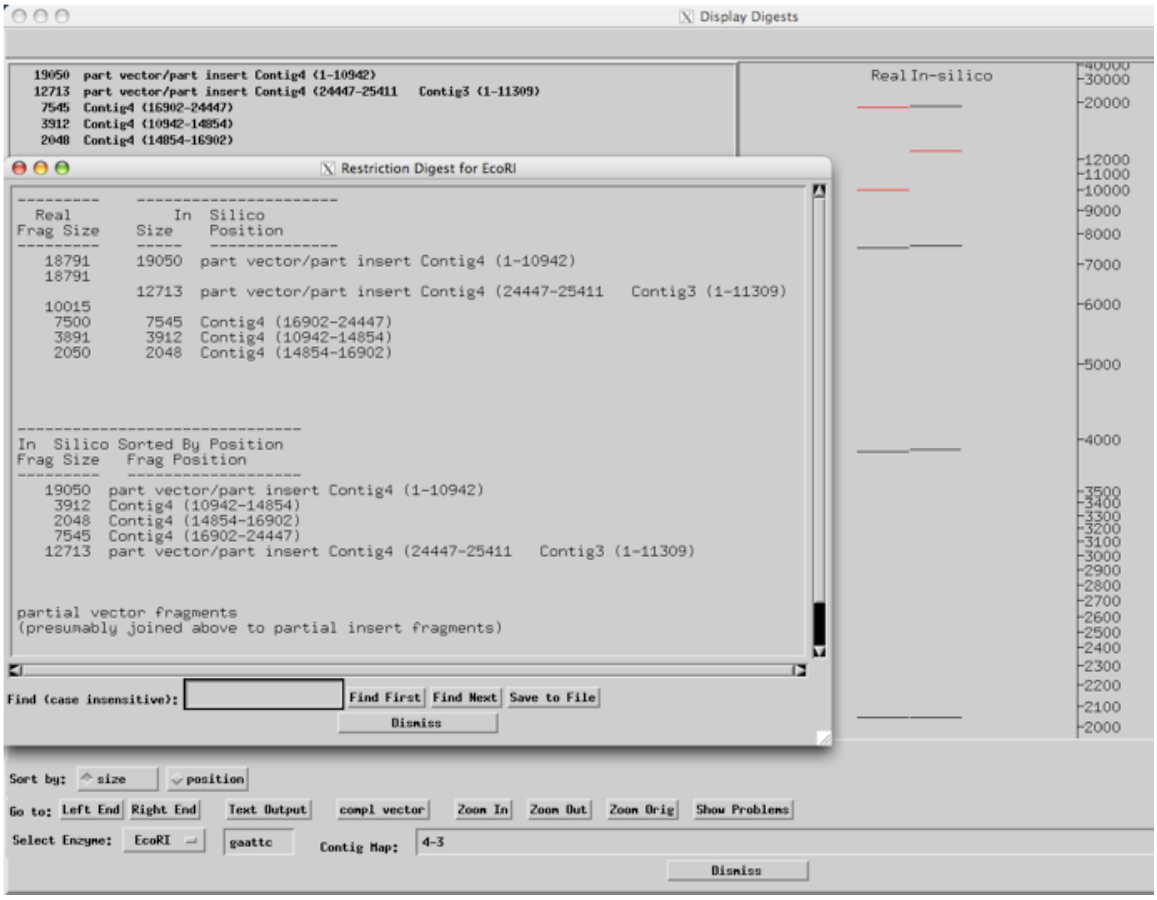


Fig. 3: the EcoRI digest exemplifies the consensus of the EcoRV, SacI, and HindIII digests that the computer assembly is larger than the real fosmid sequence by approximately 2 kb.

Resolution of Challenge Regions: Gaps

As noted above, both the restriction digests and assembly view strongly supported the idea of a misassembly by *phrap* that would allow a manual closing of the apparent gap. The structure of the tandem repeats suggested that the repeat at the 5′ end of contig 3 should be joined to the downstream region of sequence as the first step towards resolving the gap. A closer look at the read coverage in this region supported this approach (Fig. 4), so the consensus sequence was torn at the 3′ end of the first repeat region and the repetitious reads placed temporarily in their own contig 6.

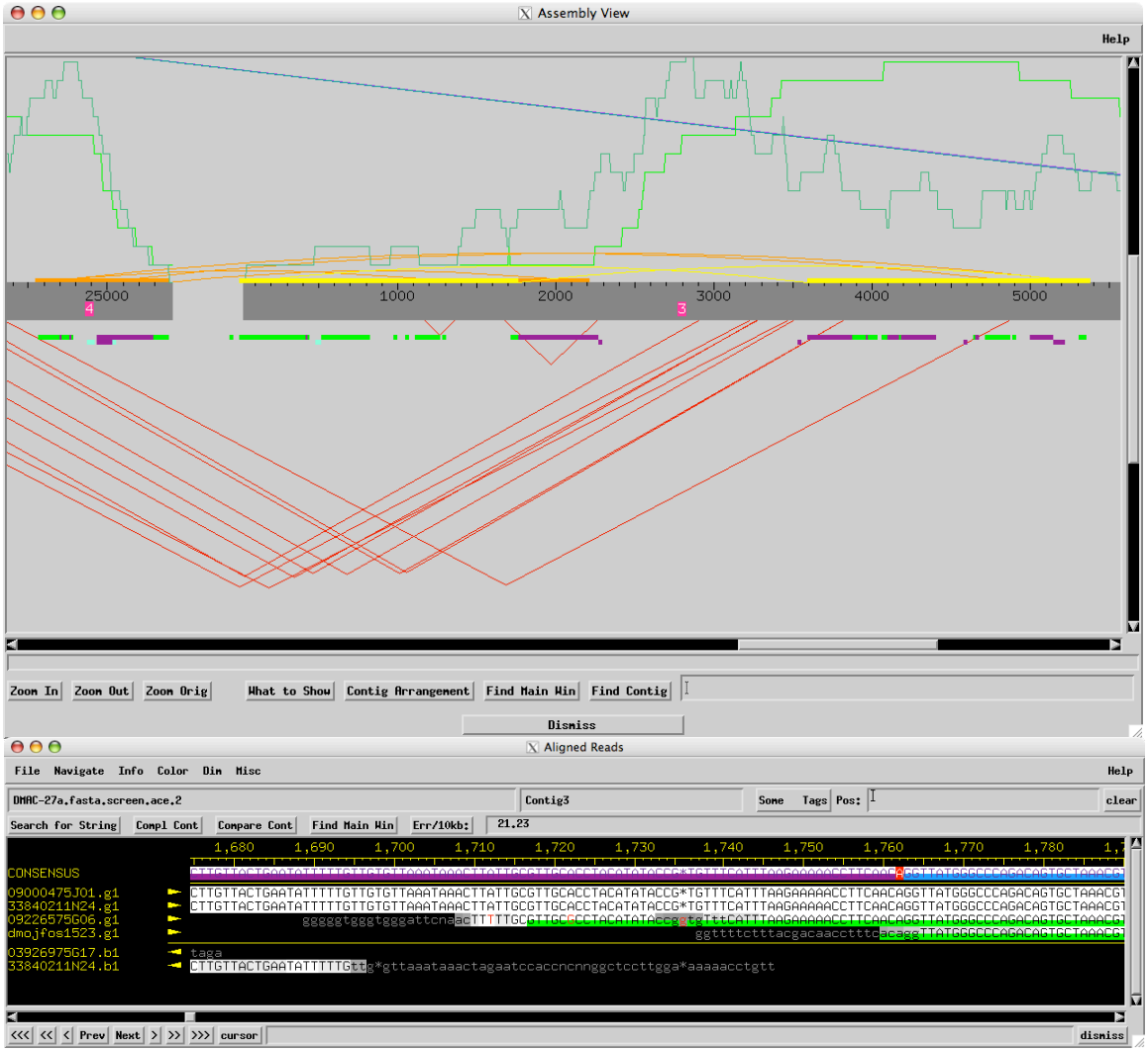


Fig. 4: The aligned reads window for the 3′ end of the first repetitious sequence in contig 3 (yellow highlight in top image) with the red cursor marking position 1762, the end of the tandem repeat and the point at which the contig was torn. The top 2 forward reads and the reverse reads were placed in their own contig, while the bottom 2 forward reads remained in contig 3.

The new contig 6 was then aligned to the matching region on contig 5 (previously contig 3) in the main assembly using *search for string* and the *compare contigs* function, as shown in Figure 5. The sequence alignment does show several high quality discrepancies, but both the restriction digests and the spacing of the unresolved forward/reverse pairs support a join between the two regions, so they were force-joined. As shown in Figure 6, this join resolved the forward/reverse pairs into gap-spanning subclones and improved the restriction digests.

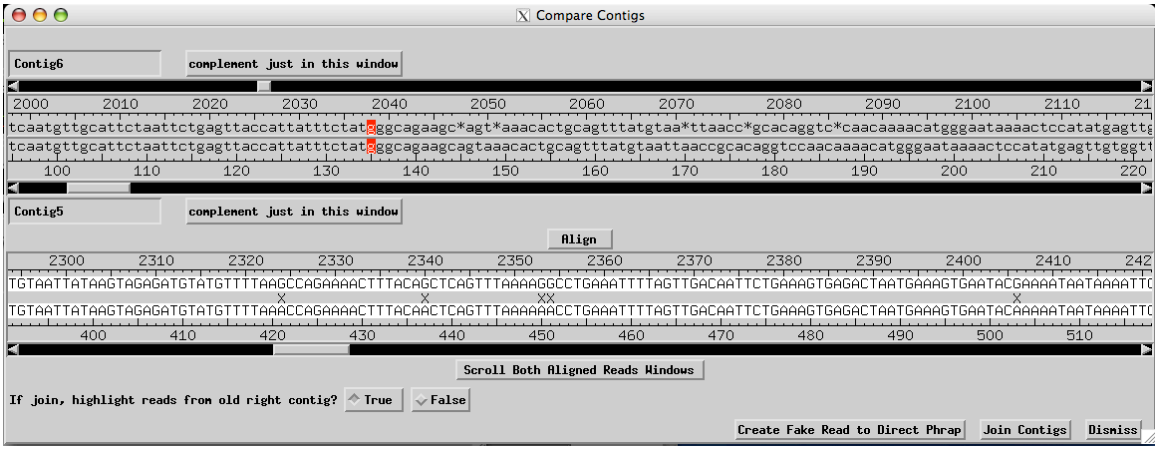


Fig. 5: the alignment between the new contig 5 and its corresponding repeat in contig 6 (previously contig 3). Xs between the sequences mark discrepancies, which here represent potential SNPs.

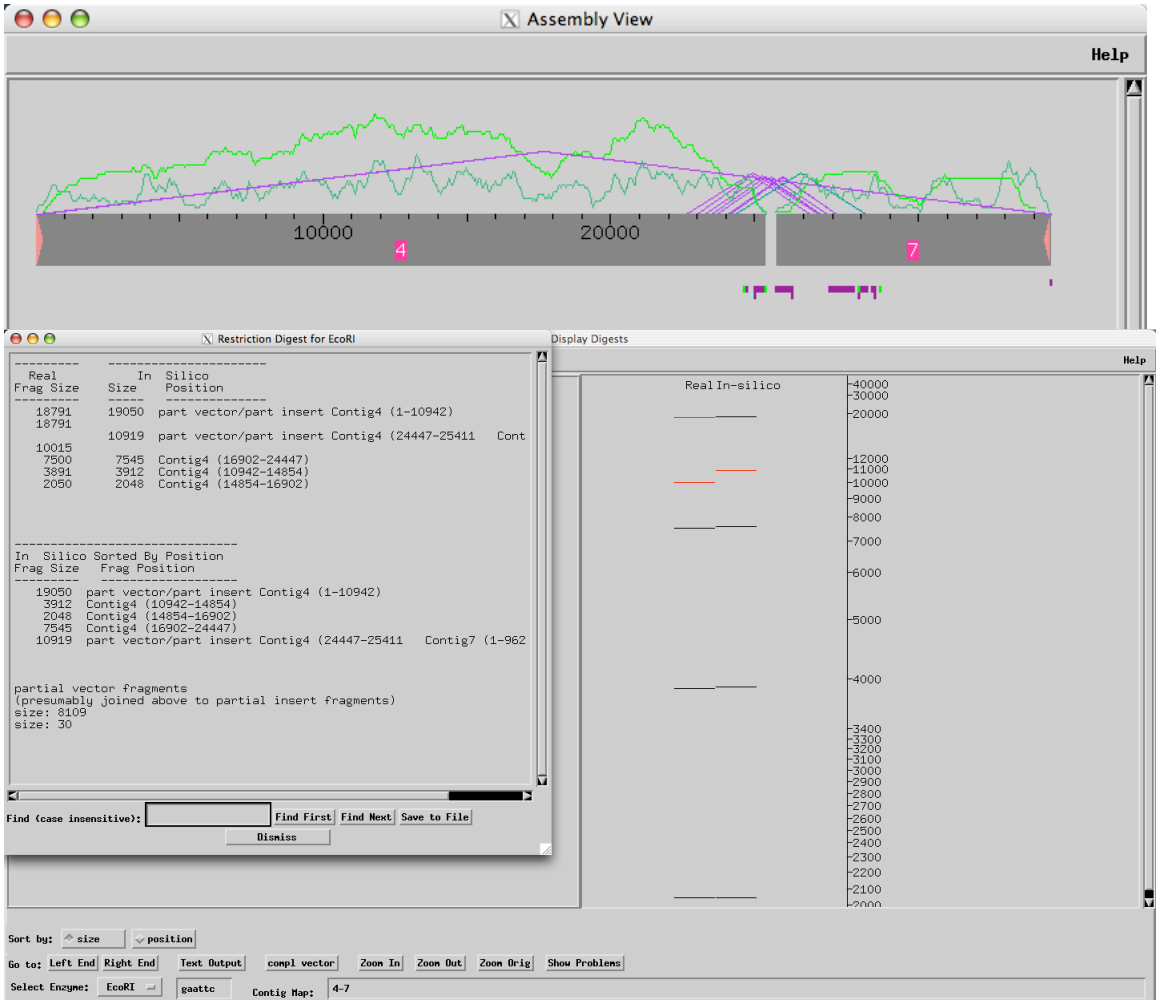


Fig. 6: the assembly view (*crossmatch* results not shown) and EcoRI digest after the forced join between the two largest tandem repeats. The truncated text describing the position of the 10919 bp *in silico* fragment again shows a misassembly around the gap between contigs 4 and 7.

The remaining discrepancy between the real and *in silico* digests shown in Figure 6 suggests a second misassembly by *phrap* that can be closed manually. The regions indicated by the digest correspond to the very ends of the two contigs and have extremely high sequence identity at 99.2%, indicating a high probability that they overlap. The repetitious regions were aligned as described above, revealing that the discrepancies were all in low quality regions or at the extreme ends of the reads, where the base calls are generally unreliable (Fig. 7). Thus, the contigs were force-joined, yielding a single contig in the main assembly and resolving regions of low consensus quality and the discrepancies in the enzyme digests (Fig. 8).

Fig. 7: the tandem repeats on the extreme ends of contigs 4 and 7 in the assembly view and the sequence alignment showing the four discrepancies between the sequences.

Fig. 8: the assembly view and (left to right) EcoRI, EcoRV, HindIII, and SacI digests after manually resolving the gaps in the assembly. Note that the red mismatch bars in the EcoRI and SacI digests were segments including vector sequence and were resolved by defining the *in silico* digest to include only those areas within the clone end tags.

Resolution of Challenge Regions: Calling and Incorporating Reads

Once the main assembly was reduced to a single contig, the two outstanding contigs remained to be incorporated, and reads needed to be called to both confirm the forced joins and supplement areas of single strand or chemistry coverage. Because both ends of DMAC-27a overlap by approximately 4 kb with other fosmids (data not shown), low consensus quality and single strand/chemistry issues in these regions will be resolved by the final assembly of the chromosomal sequence and do not need additional reads called at this stage. Regions with a consensus quality higher than *phred* 30 are also considered sufficiently covered.

After manually editing the bases miscalled by *phred* to give a reliable foundation for the string search, the two outstanding contigs were incorporated into the main assembly by doing a *search for string*, alignment, and force join. The only complications came in that the single read in contig 1 matched to both of the remaining tandem repeats (Fig. 9), and the reads in contig 2 had to be complemented to match the main assembly (Fig. 10). Neither contig matched perfectly with the consensus of the main assembly, but the restriction digest results, the fact that the matches for the outstanding contigs were so limited, and the already high incidence of SNPs supported the joins.

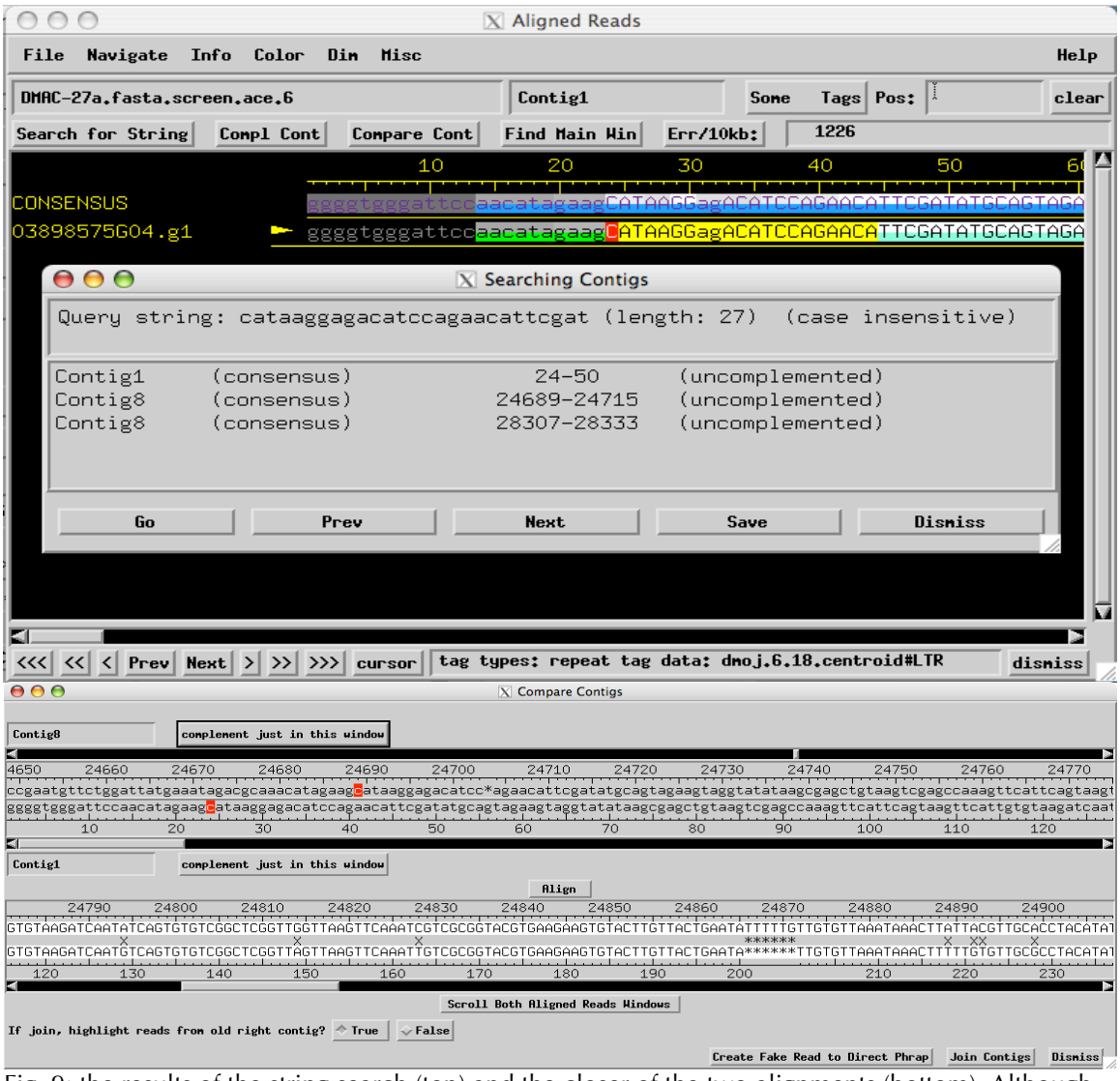


Fig. 9: the results of the string search (top) and the closer of the two alignments (bottom). Although there are several inconsistencies and one deletion, the enzyme digest data, the limited matches within the main assembly, and the already high number of SNPs in the repeats supports the join. Moreover, the 3′ end of the alignment extends over 100 bp past the end of the repetitious region identified in the initial *phred/phrap* assembly, suggesting that this sequence was properly placed into this assembly from the Whole Genome Shotgun data.

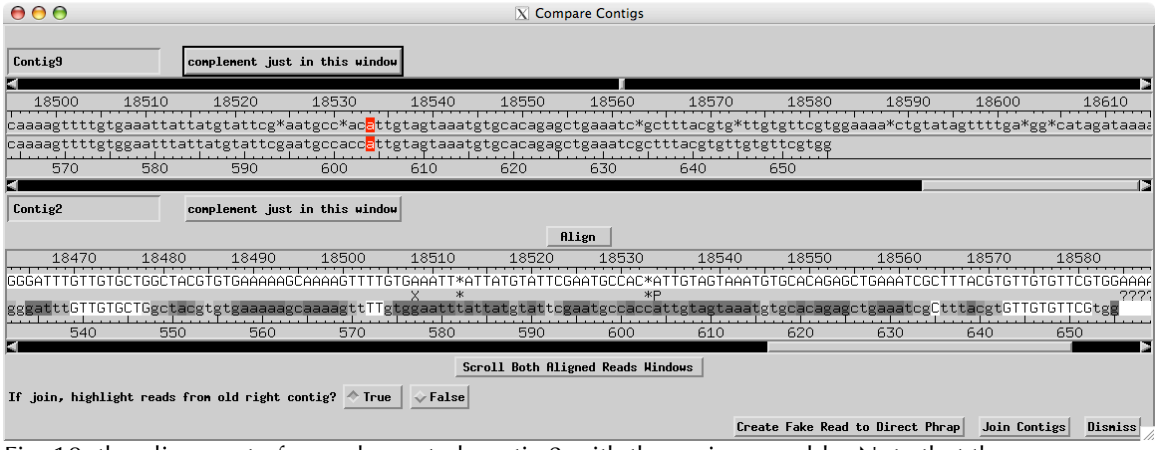


Fig. 10: the alignment of complemented contig 2 with the main assembly. Note that the discrepancies occur in areas of low consensus quality near the end of contig 2, making them of little concern.

With only one contig, the only reads that needed to be called were to verify the initial forced joins and to supplement regions with insufficient coverage in at least one direction. Figure 11 shows the table of reactions called and whether or not they were ultimately successful. Note, however, that the first plate of each reaction run had very low success rates, resulting in low quality reads that could not be reliably incorporated even with knowledge of the primer positions. Here, however, only the ultimate outcome of the called reads is displayed (Fig. 11).

Final Results of New Read Calls						
Primer No.	Reason for Call	Big Dye	dGTP	4:1	Target Sequence	Covered Sequence
5	Single strand or chemistry	Fail	Success	Success	19387-19431	19049-19277, 19296-19309
7	Single strand or chemistry	Fail	Fail	Fail	29854-29882	No Coverage
8	Confirm forced join	Fail	Success	Success	24546-25387	24551-24974
9	Confirm forced join	Fail	Success	Success	24546-25387	24993-25492
10	Confirm forced join	Success	Success	Fail	17930-18584	18400-18597
11	Confirm forced join	Fail	Success	Fail	17930-18584	17873-18294
12	Single strand or chemistry	Success	Success	Success	29854-29882	29463-30114

Fig. 11: the final results for all of the new reads called to cover forced joins and large single strand/chemistry regions. Primer 12 was designed after the results from the first round of reactions showed that primer 7 yielded no reads.

Several of these reads had consistently miscalled bases and were thus initially incorporated into the wrong regions of the assembly. Manual correction of the sequences did allow for

the correct assembly for some of the misplaced reads, but others were too low quality to be properly aligned and had to remain in their own contigs (Fig. 12).

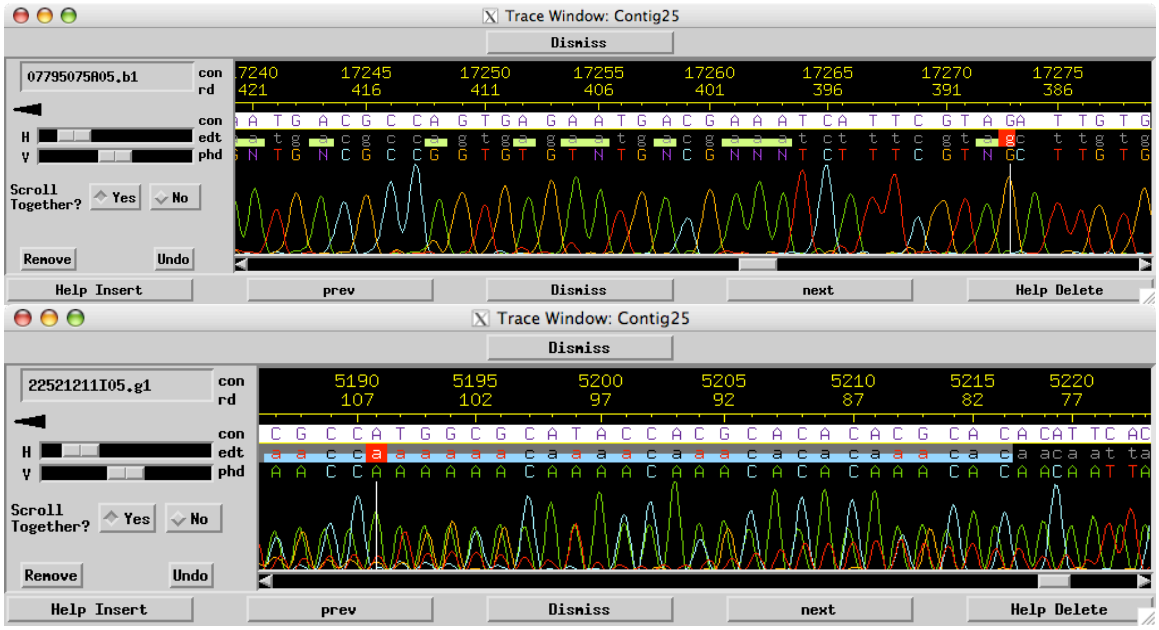


Fig. 12: manual editing of bases to allow for correct incorporation of new reads (top). Sequence contamination such as that in the bottom trace prevented the incorporation of other reads.

After the incorporation of these reads, several single-strand and single-chemistry regions remained in the main assembly. However, in each case the consensus was well above the quality threshold, and in several instances five or more high quality reads supported the consensus. The high quality of the consensus in these regions meant that no new reads were required, but the manually called reads were still contrasted with the results of *Autofinish* in order to test whether more reads may be necessary in other regions. While *Autofinish* did identify the same problematic regions in the assembly, it was unable to design any primers because of the repetitious regions (Fig. 13).

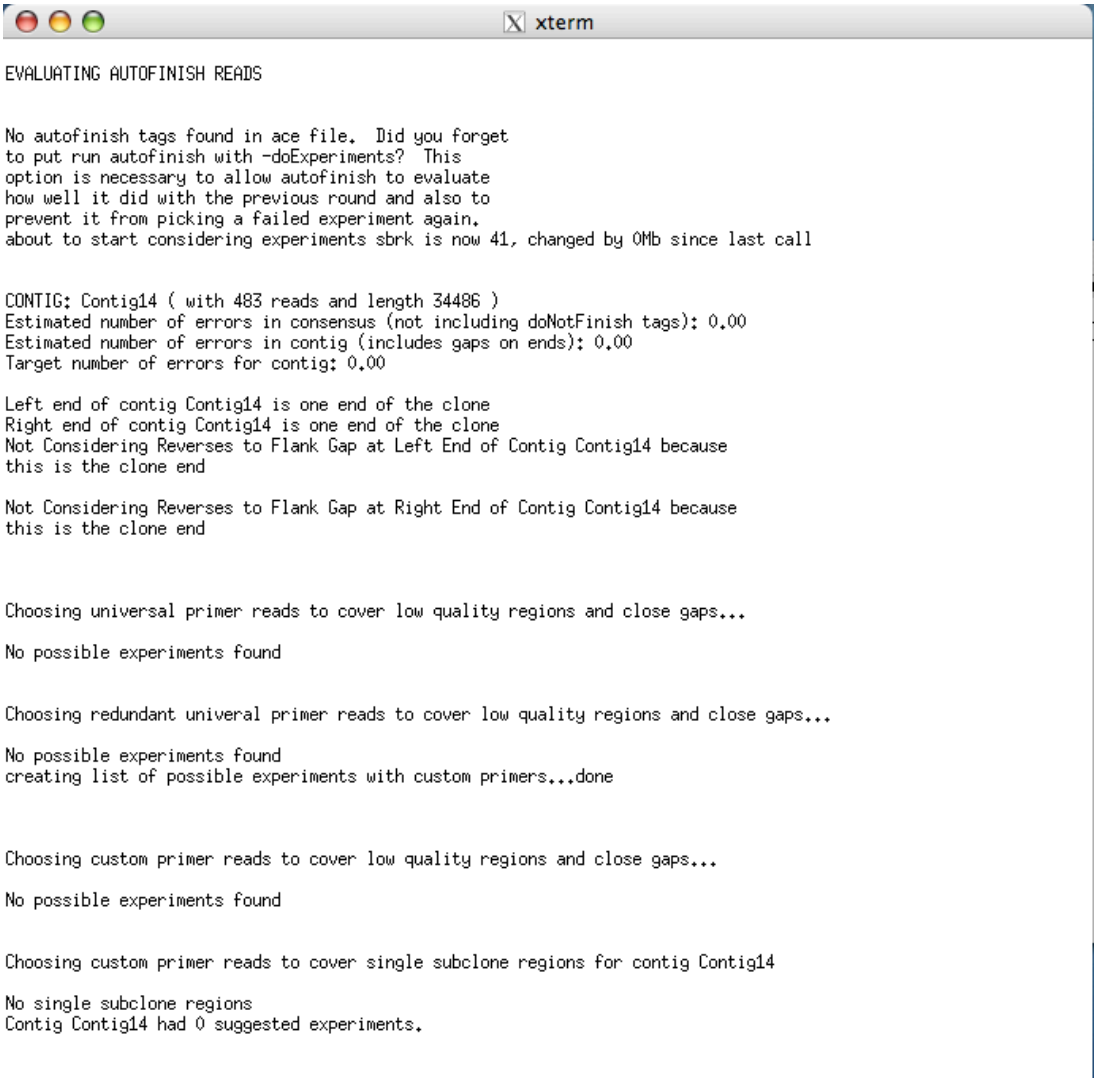


Fig. 12: Autofinish suggested no read calls for the edited assembly.

Resolution of Challenge Regions: SNPs and High Quality Discrepancies

The changes to the assembly and incorporation of new data dramatically increased the number of high quality discrepancies in the sequence, as can be seen in Figure 14. All of these regions were first included in a string search both of the consensus and of other reads to determine whether they could represent a misassembly. The unaligned high quality regions were found to be due to contaminated reads such as that shown in Figure 12, and were thus marked as low quality. High quality discrepancies supported by multiple reads were tagged as potential SNPs in the consensus sequence (Fig. 15), while those supported by only one read were either found to be condensed peaks or tagged as potential growth differences between plates (Fig. 16).

Fig. 14. the problem navigation window after the forced joins and incorporation of new reads into the main assembly shows the huge number of high quality discrepancies. Note that most of these are supported by multiple reads at each position, suggesting that they are SNPs rather than evidence of misassembly.

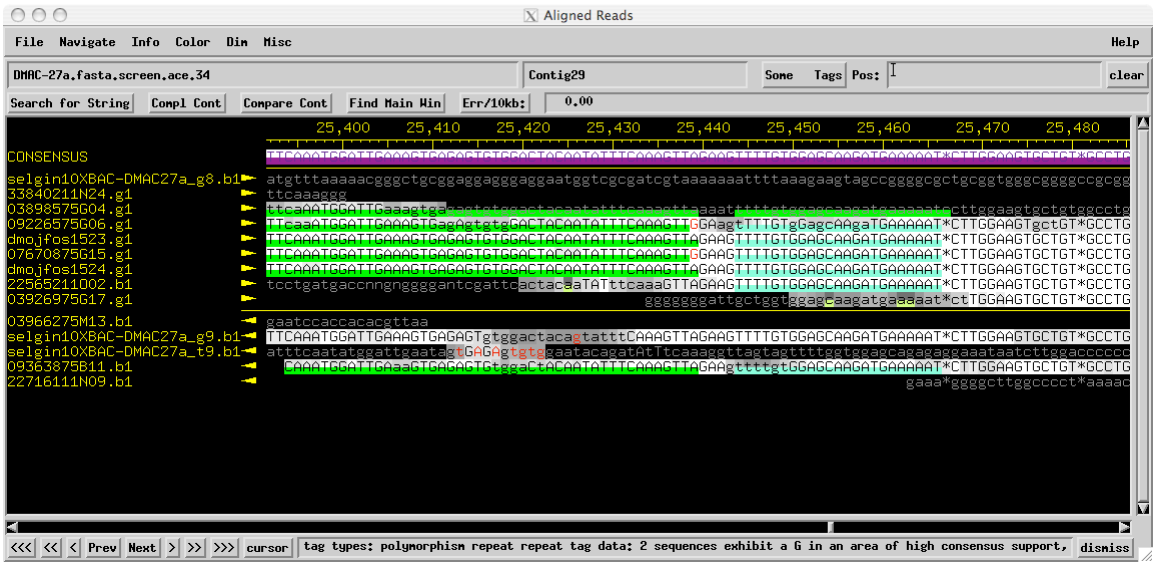


Fig. 15: SNP at position 25439 is supported by two high quality strands in an area of good consensus support.

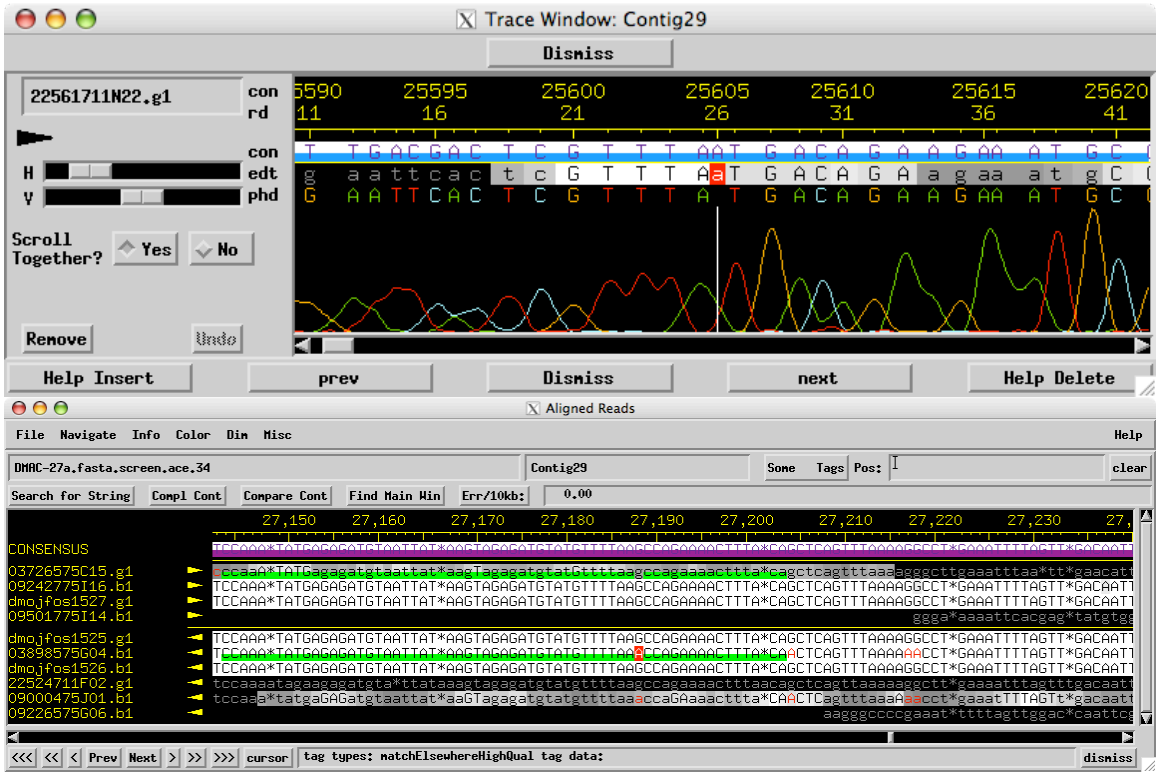
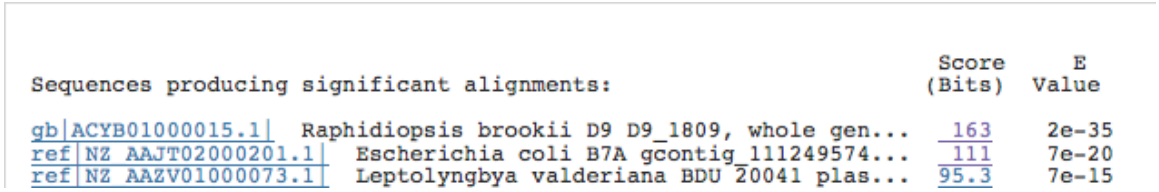


Fig. 16: A compression (top) of two adenosine signals into a single broad peak, and a SNP in a single read suggesting mutation occurring within the fosmid during amplification (bottom).

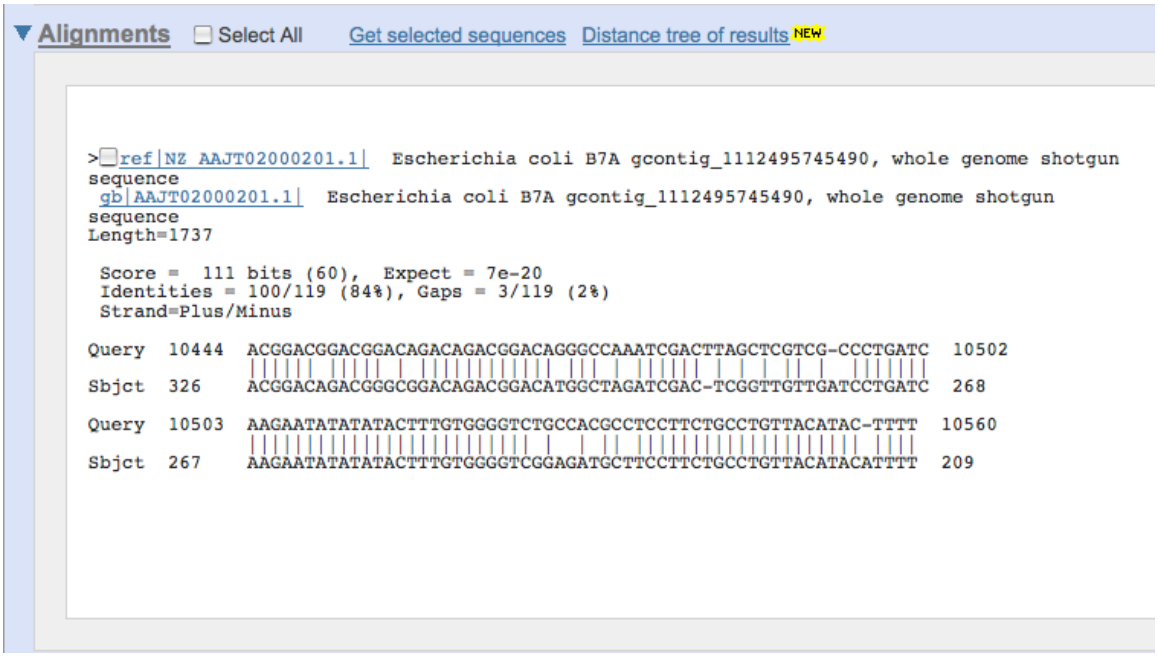
Confirmation of assembly

Once all of the challenge regions in the fosmid were resolved, the final assembly was tested for vector contamination by uploading the sequence to BLAST for comparison to all microbial genomes. The initial comparison yielded numerous high quality matches at the 5′ end to the *E. coli* vector sequence (Fig. 17). The 5′ clone end tag was thus moved to exclude the matched regions and the sequence was resubmitted to BLAST. This second search resulted in a single match (Fig. 18). However, the percent identity of the match is too low to act as convincing evidence of vector contamination, so the assembly may be considered confirmed.



Sequences producing significant alignments:		Score (Bits)	E Value
gb|ACYB01000015.1|	Raphidiopsis brookii D9 D9_1809, whole gen...	163	2e-35
ref|NZ_AAJT02000201.1|	Escherichia coli B7A gcontig_111249574...	111	7e-20
ref|NZ_AAZV01000073.1|	Leptolyngbya valderiana BDU 20041 plas...	95.3	7e-15

Fig. 17: the most significant matches from the initial BLAST alignment, with the first and third aligning to the 5′ end of the contig.



Alignments Select All Get selected sequences Distance tree of results NEW

>ref|NZ_AAJT02000201.1| Escherichia coli B7A gcontig_1112495745490, whole genome shotgun sequence
gb|AAJT02000201.1| Escherichia coli B7A gcontig_1112495745490, whole genome shotgun sequence
Length=1737

Score = 111 bits (60), Expect = 7e-20
Identities = 100/119 (84%), Gaps = 3/119 (2%)
Strand=Plus/Minus

```
Query  10444  ACGGACGGACGGACAGACAGACGGACAGGGCCAAATCGACTTAGCTCGTCG-CCCTGATC  10502
              |||||| ||||| | |||||||||||| ||| | |||||| | | | || |  |||||||
Sbjct  326    ACGGACAGACGGGCGGACAGACGGACATGGCTAGATCGAC-TCGGTTGTTGATCCTGATC  268

Query  10503  AAGAATATATATACTTTGTGGGGTCTGCCACGCCTCCTTCTGCCTGTTACATAC-TTTT  10560
              |||||||||||||||||||||||| |  | || ||||||||||||||||||| ||||
Sbjct  267    AAGAATATATATACTTTGTGGGGTCGGAGATGCTTCCTTCTGCCTGTTACATACATTTT  209
```

Fig. 18: the single match generated by the second BLAST alignment is fairly repetitious and short, making it too poor to constitute evidence of vector contamination.

Conclusions

Verification of correct assembly by the enzyme digests, which remain unchanged from Figure 8, and BLAST analysis concludes the finishing of this sequence. The sequence has been resolved into a single contig of approximately 34 kb with two small regions of low-identity tandem repeats (Fig. 19). There are no areas of sequence below the *phred* 30 threshold or covered by only a single subclone within the defined insert, and no vector appears in this region. The high number of SNPs present in this fosmid, while justified by the restriction digests and not surprising for *D. mojavensis*, may warrant further investigation. However, it appears that the fosmid is otherwise ready for annotation and incorporation into the final genome sequence.

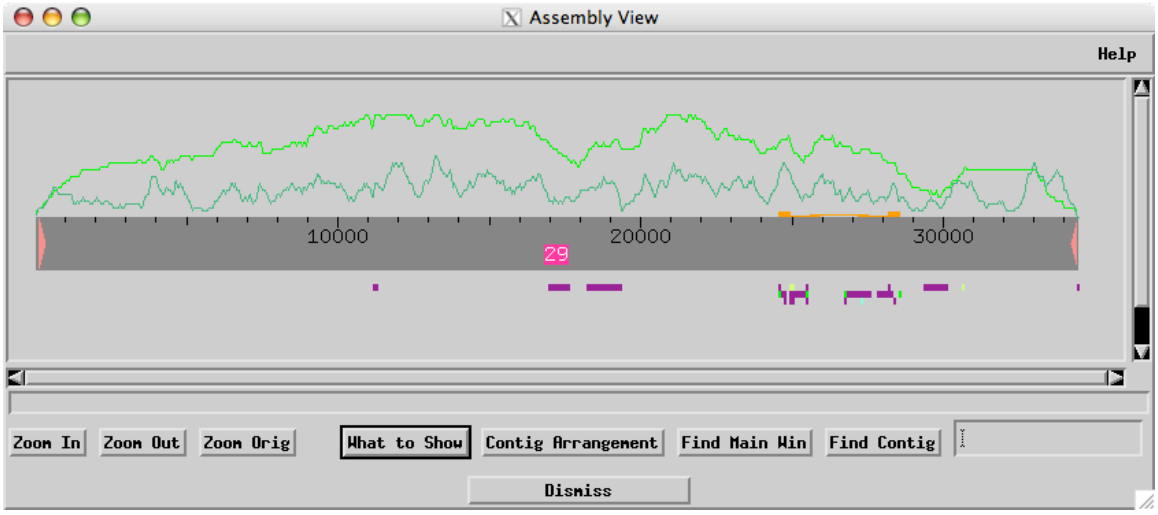


Fig. 19: final assembly view of DMAC-27a.

Addenda

In addition to the *D. mojavensis* third chromosome fosmid DMAC-27a, two *D. grimshawi* dot chromosome fosmids were also finished as part of this project: DGA28N19 and DGA16F05. Neither fosmid contained misassemblies, but rather required additional reads to supplement areas of single strand/chemistry coverage. Presented here is the outline of the finishing project for each fosmid.

DGA28N19

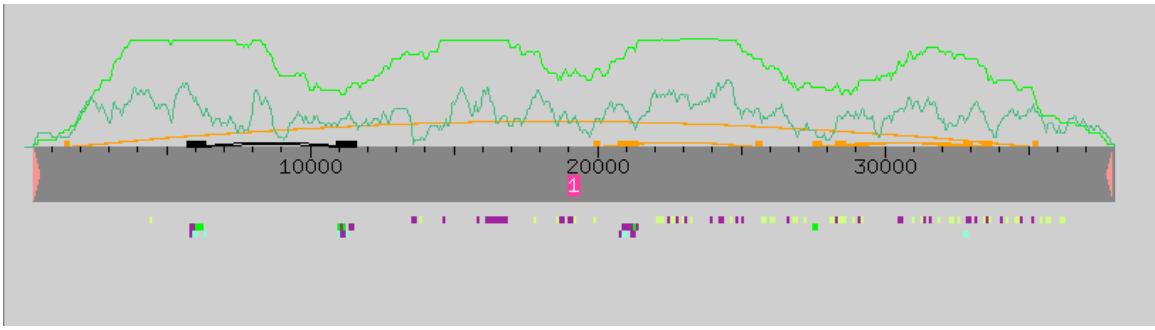


Fig. a1: the initial assembly view of DGA28N19 showing tandem and inverted repeats (orange and black, respectively).

DGA28N19 is a repeat-rich sequence on the *D. grimshawi* dot chromosome, as can be seen in the initial assembly view (Fig. a1). Despite the many tandem and single pair of inverted repeats, only one pair is of high enough quality to be a candidate for a misassembly, and no misassembly is evident in the restriction digests (eg. Fig. a2). The major challenges of this assembly were that the initial restriction digests were of poor quality and several stretches of sequence required additional reads (Fig. a3) to increase support of the consensus.

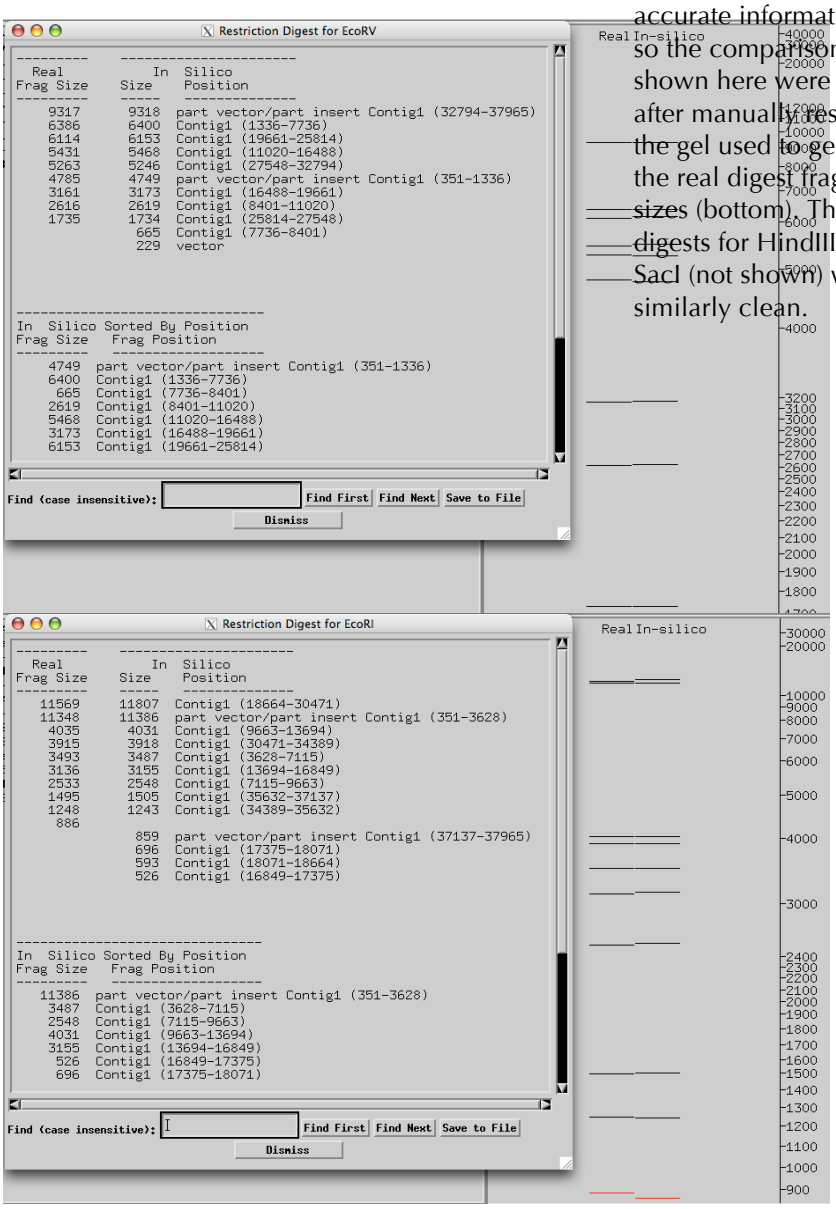


Fig. a2: the EcoRV and EcoRI restriction digests for DGA28N19. The initial real digest was too poor to give accurate information, so the comparisons shown here were made after manually resetting the gel used to generate the real digest fragment sizes (bottom). The digests for HindIII and SacI (not shown) were similarly clean.

Final Results of Reads Called for DGA28N19						
Primer No.	Reason for Call	Big Dye	dGTP	4:1	Target sequence	Covered Sequence
1	Single strand	Success	Fail	Success	13637-13830	13405-14409
2	Single strand	Success	Success	Success	27886-28156	27311-28331
3	Single strand	Success	Success	Fail	13637-13830	13453-13773
4	Single strand	Success	Success	Fail	13637-13830	13567-14000
5	Single strand	Success	Success	Fail	27886-28156	27569-28220
6	Single strand	Success	Success	Fail	27886-28156	27890-28296

Fig. a3: the final results for reads called for DGA28N19. As with the plates for DMAC-27a, the first run yielded no successes, while the second run of the same plate resulted in high quality reads for incorporation into the assembly. Reads 3-6 were called before the rerun of the initial plate

The reads incorporated successfully into the extant assembly with occasional manual editing such as shown in Figure 12. All other problem regions, such as high quality discrepancies, were resolved by the same methods detailed in the finishing report for DMAC-27a. Once these issues were resolved, the sequence was submitted to BLAST to check for vector contamination, returning no significant matches (Fig. a4). The assembly was thus considered to be finished, and is shown in Figure a5.

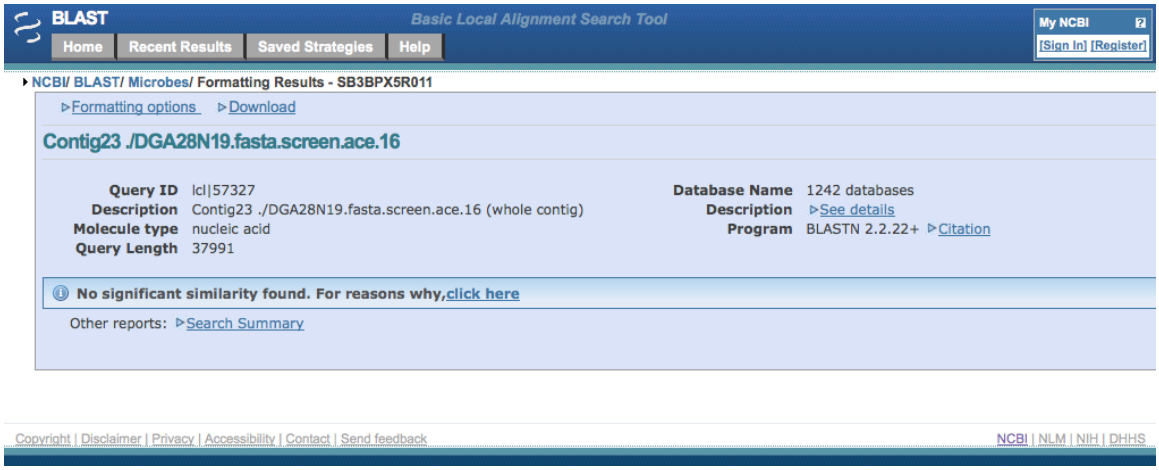


Fig. a4: BLAST found no significant matches between the finished sequence of DGA28N19 and any microbial genome, showing it to be free of vector contamination.

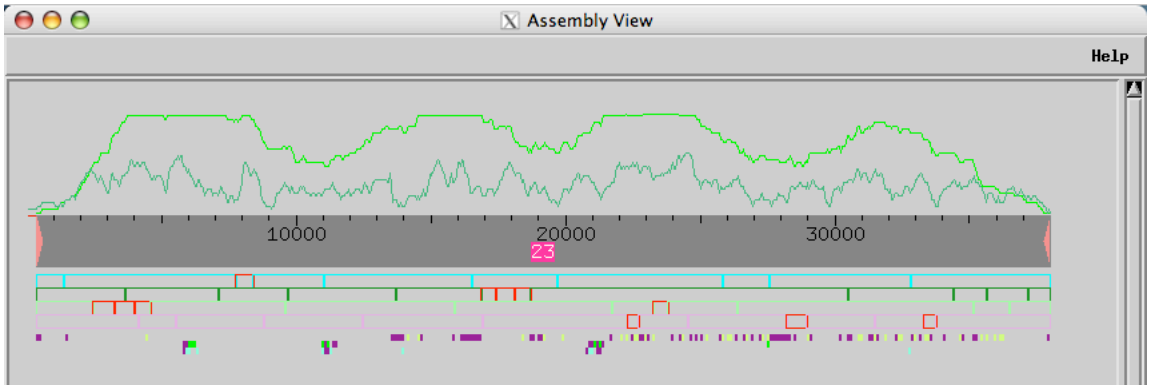


Fig a5: the final assembly view of DGA28N19, showing restriction digest fragments (*crossmatch* results not shown).

DGA16F05

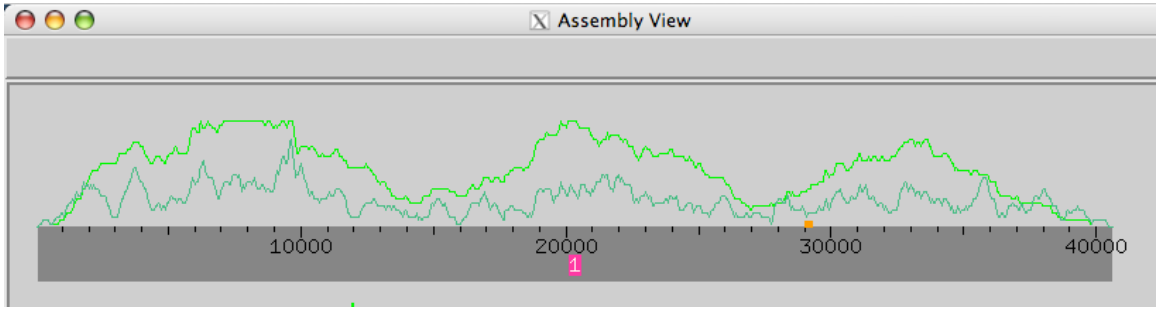


Fig. a6: the initial assembly view of DGA16F05.

The *D. grimshawi* dot chromosome fosmid DGA16F05 presented fewer challenges than DGA28N19, containing only one repetitious region and no misassemblies (figs. a6, a7). However, the coverage of several areas was too shallow and required numerous additional reads. Two regions noted by the problem navigation window as covered by only a single strand, positions 11724-11769 and 13407-13965, had sufficiently high *phred* scores and depth of high quality single-strand coverage to require no additional reads. Reads were called for all other areas of low consensus quality or single strand/chemistry coverage (Fig. a8)

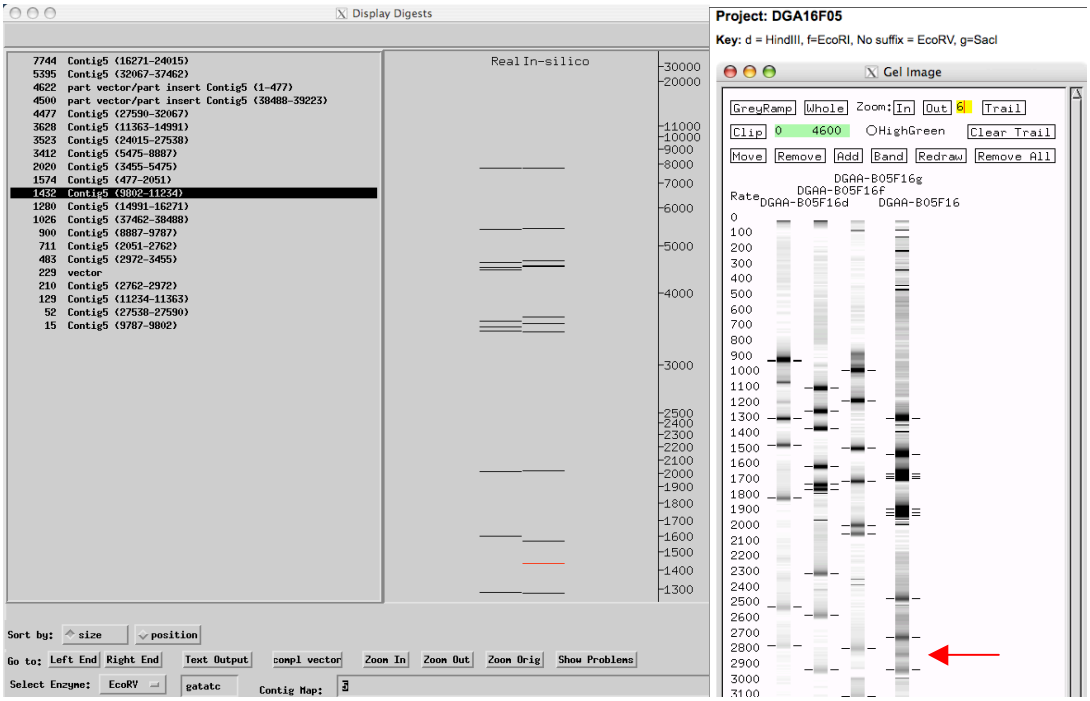


Fig. a7: After manually redefining which gel image should be used for the real digest restriction fragments, only one discrepant band remained in the EcoRV digest. However, on closer examination of the gel in consultation with a professional finisher, it was decided that this is due to a faulty computer analysis of the digest. The appropriate real band is indicated by the red arrow.

Final Results of New Read Calls						
Primer No.	Reason for Call	Big Dye	dGTP	4:1	Target Sequence	Covered Sequence
1	Single strand or chemistry	Success	Success	Success	2893-3034	1710-2720
2	Single strand or chemistry		Success		12236-12470	11326-12281
3	Single strand or chemistry		Success		14108-14937	13620-14547
4	Single strand or chemistry	Success	Success	Success	15784-16129	14950-15924
5	Single strand or chemistry		Success	Success	15784-16129	14708-15674
6	Base quality below threshold	Success	Success	Success	27820-27821	26270-27269
7	Single strand or chemistry	Success	Success		32227-32377	31420-32409
8	Single strand or chemistry		Success	Success	32227-32377	30767-31812
9	Single strand or chemistry	Success	Success		36516-36651	35572-36543
10	Single strand or chemistry		Success		39712-39798	38702-39654

Fig. a8: Final results of new reads called for DGA16F05. Reads were incorporated by re-running *phred/phrap*, and blank spaces indicate reads that were ignored by the program.

Although not all of the reads were incorporated by *phred/phrap*, the final problem navigation window (Fig. a9) no longer shows any unaddressed low quality regions outside of the 2 kb end regions, which will be resolved in the final assembly of the *D. grimshawi* sequence. Note that the sequences represented as unaligned high quality have actually been tagged as low quality because of sequence contamination in those reads.

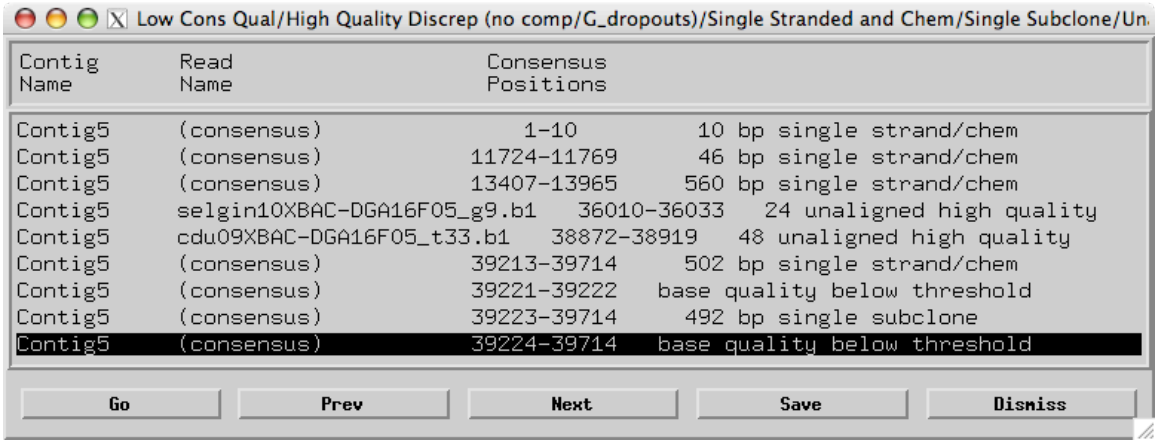


Fig. a9: Final problem navigation window.

The fosmid was then submitted to BLAST for comparison to microbial sequences. No matches were found (Fig. a9), indicating that the assembly was free from vector contamination and could be considered finished (Fig. a10). The assembly view does still show areas with markedly higher coverage than others, but the all interior regions are now well above the *phred* 30 quality threshold.

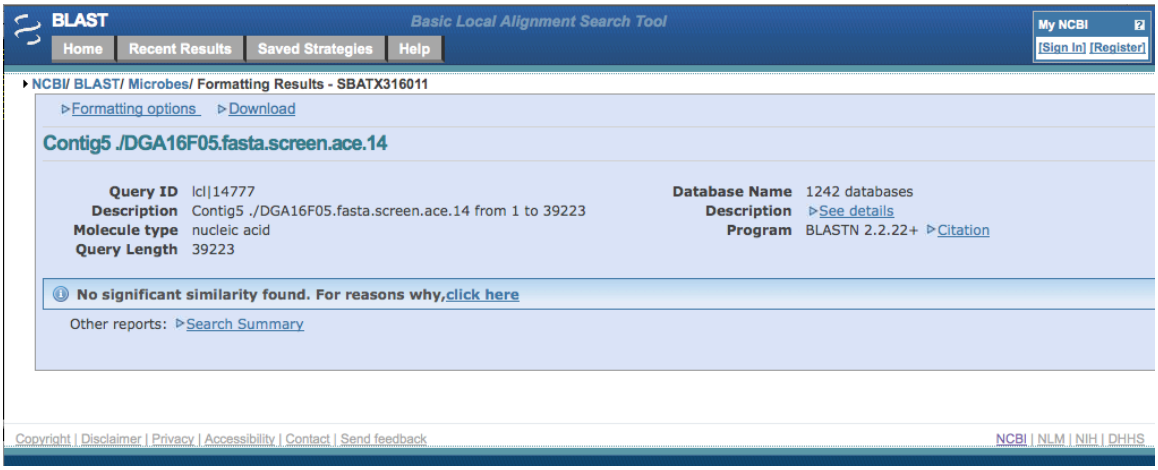


Fig. a9: a BLAST analysis found no matches for the DGA16F05 fosmid.

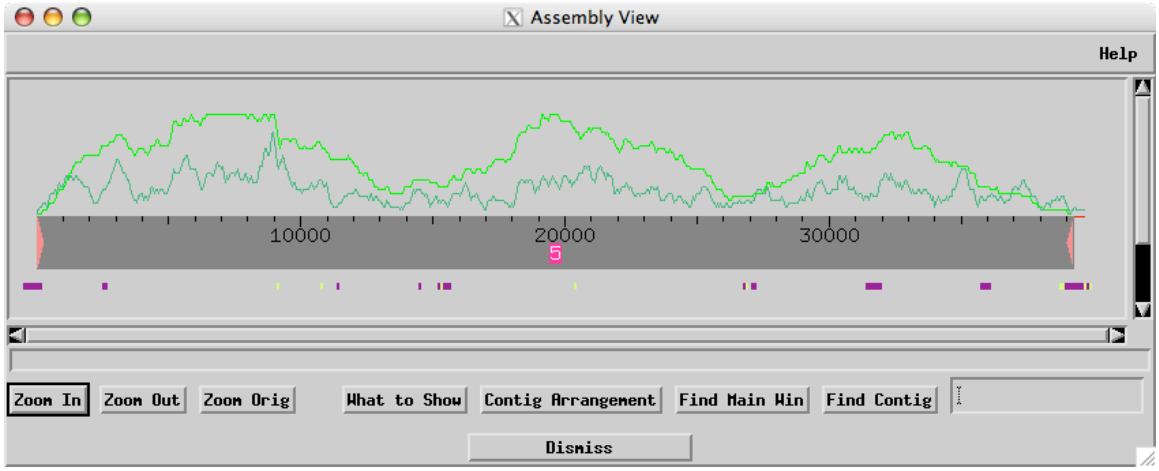


Fig. a10: the final assembly view of DGA16F05, showing improved sequence coverage.