Natalie Gehred
Dr. Elgin, Dr. Shaffer
Biology 434W
26 February 2016

Finishing of DFIC7492007

Abstract

Drosophila ficusphila has been sequenced using both Roche 454 pyrosequencing and Illumina paired-end sequencing. It is important that the final consensus sequence be constructed as accurately as possible, using all of the data available and taking into account the strengths and weaknesses of both sequencing technologies. DFIC7492007 is a highly repetitive 100,000 bp contig with one 20 bp gap, several regions of very low coverage, and many reads that mapped multiple places in the contig. The process of improving the quality of this contig's sequence started with an examination of mononucleotide repeats (MNRs) in High Quality Discrepancy regions (HQDs), which was challenging due to the large number of potentially mismapped reads. A gap at base 37,106 was resolved by removing the assembly piece and rejoining the resulting two contigs by searching for areas of overlap at their ends. The low-coverage region around base 69,000 was not resolved; this region needs the insertion of an assembly piece. MNRs in other low coverage regions were examined and improved. Any remaining ambiguous sites in low-coverage regions that could not be resolved were tagged as needing additional data. Primers could not be designed for further sequencing in most low coverage regions due to DFIC7492007's extremely repetitive sequence.

Introduction

The DFIC7492007 contig is a 100,000 bp region in the F element of the *Drosophila ficusphila* genome. The F element, also known as the dot chromosome, is unique due to its

unusual chromatin structure. While most actively transcribed genes are located in a euchromatic environment, it has been shown that the genes on the dot chromosome are located in an almost entirely heterochromatic environment. Despite being located in regions with silencing modifications, like histone H3 lysine 9 trimethylation, these genes are present at the same density and are expressed at levels comparable to those in euchromatic domains. The ultimate goal of this research project is to use comparative genomics to better understand heterochromatin and the dot chromosome in *Drosophila*. The *D. ficusphila* species is an important organism to sequence, as it is located in the 'sweet spot' of evolutionary divergence for the detection of small regulatory motifs. This means that *D. ficusphila* diverged from *Drosophila melanogaster* just recently enough to have recognizable motifs in regulatory regions that have not been subjected to a great amount of genetic drift.

Initial Assembly

The finishing project (DFIC7492007) is a 100 kb contig spanning bases 540,000-640,000 on the dot chromosome. Running Crossmatch on the initial assembly of the contig (Figure 1) shows that this region is very repetitive: under these circumstances forward/reverse read pairs will be mapped to different regions of the contig at a significant frequency, as seen by the red lines underneath the contig. There is also a sharp dip in coverage around base 69,000 that warrants investigation (Figure 2). In total, there were 19 low consensus quality positions, 280 highly discrepant regions, and 1 gap (not visible in Assembly View) of 20 bp.

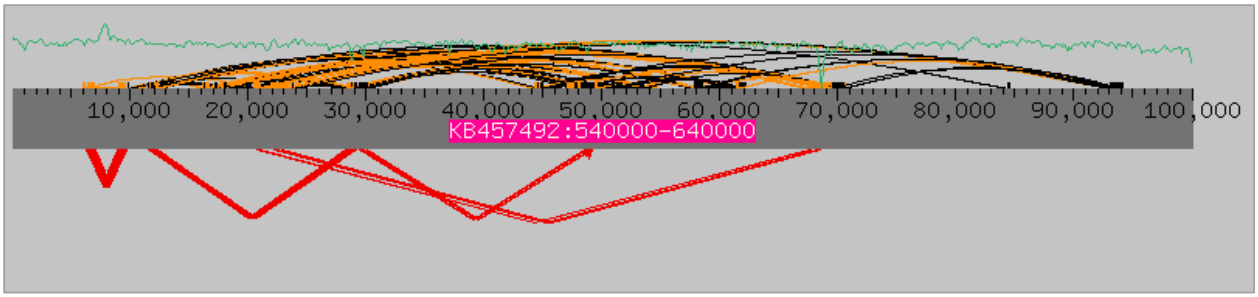


Figure 1. Initial Assembly View of Contig.
The contig appears as one full assembly. Red lines underneath the contig show that there are many discrepant forward and reverse mate pairs of reads. The black and orange arcs on the top of the contig map repeated areas. The green line shows that there is high coverage across the contig, excluding the region around position 69,000, where there is a sharp decrease in coverage.

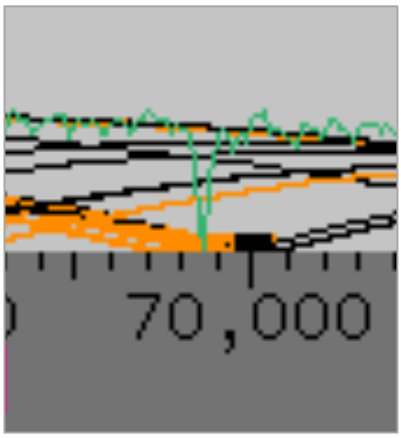


Figure 2. Dip in Coverage.
The green line signifies read number at a certain position. Assembly View shows a sharp decrease in coverage around 69,000 bp in a repetitive region of the contig.

HQDs

The contig was first examined for high quality discrepancies, defined as positions where there are at least three reads that disagree with the consensus sequence and have quality scores equal to or above 30. The list generated by Consed included 280 regions that were scanned for mononucleotide repeats. MNRs were specifically examined due to the fact that the technology

used by 454 sequencing is often inaccurate in these regions. Pyrosequencing measures the luminescence emitted when a specific nucleotide is incorporated, but the direct relationship between the amount of light detected and the number of bases added becomes error-prone after six of the same nucleotide are added consecutively. Thus, 454 sequencing often has errors in these regions, so high-quality Illumina reads (which do not use pyrosequencing technology) were used to resolve any discrepancies. Out of the 280 high quality discrepancy regions, 95 contained an MNR, 45 of which were changed to reflect the calls of the high quality Illumina reads. Five other regions were changed to reflect the high quality 454 and Illumina reads, where all but one or two reads called for a certain base that was not reflected in the consensus sequence (Table 1, Appendix). Most of the MNRs that were examined and corrected were monoA or monoT runs, which was expected because heterochromatic regions are enriched in A and T nucleotides.

The majority of the MNRs could be easily resolved by counting the number of bases in the Illumina and the 454 reads; often they could be reconciled. If not, the number of bases that the Illumina sequencing detected was understood to be more accurate and the consensus sequence was changed to reflect the Illumina call (Figure 3). However, there were some rather difficult cases due to areas of low coverage or confusion from mismapped reads, especially around bases 30,000 and 68,000. At base 29,539, for example, it was unclear if the consensus sequence should be changed (Figure 4). In these instances, 454 reads were given more consideration; the longer read length produced by 454 reads provided a lower probability of being mismapped in areas of high repetition. Regions with a high density of mismapped reads were tagged for consideration when making changes to the consensus sequence.

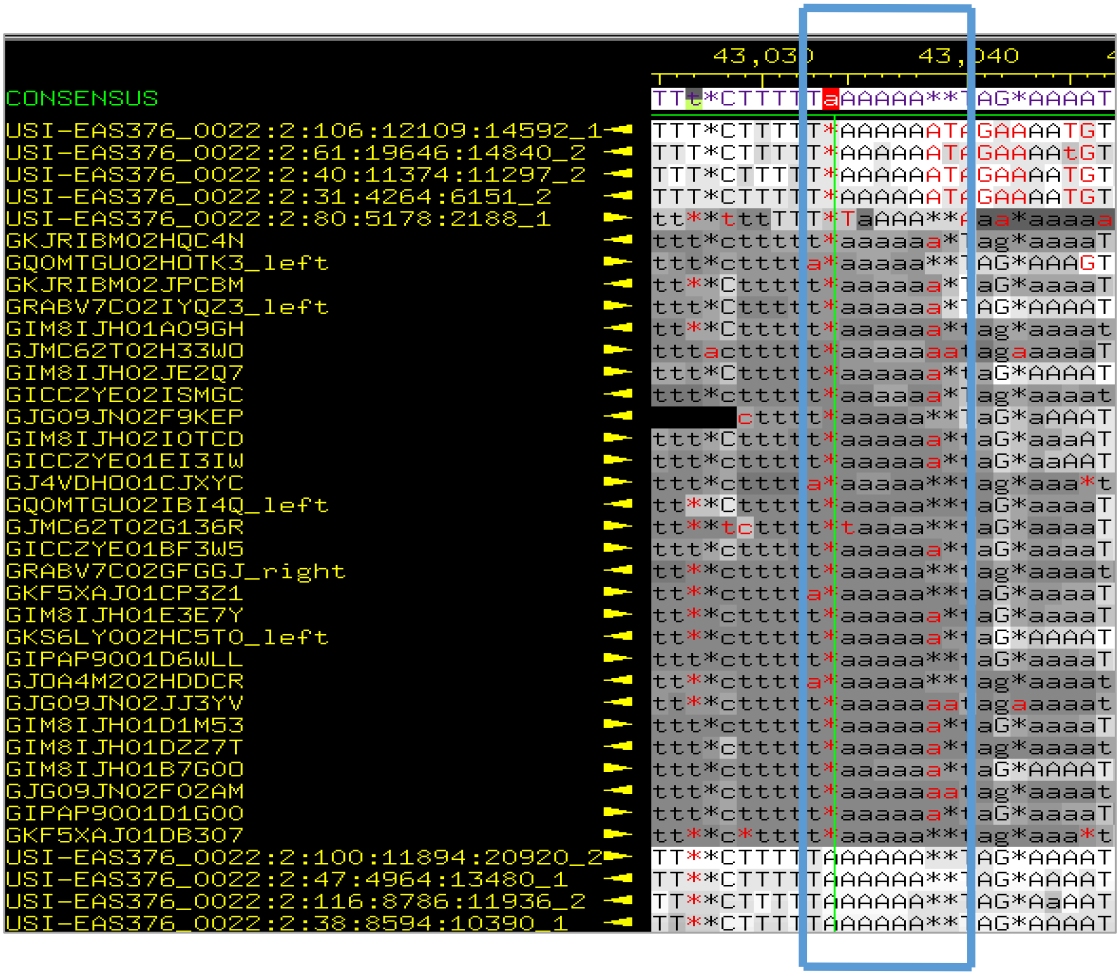


Figure 3. A Typical MNR Correction to Consensus Sequence at position 43,034.
At this position (listed as 43,055 in the ace.2 file in table above), counting the As in the Illumina-derived reads (distinguishable by "USI" in the read name) and a majority of the 454 reads (distinguishable by a "G" in the read name) agreed on a call of 6 As in this region, so an A was added to replace a pad in this position.

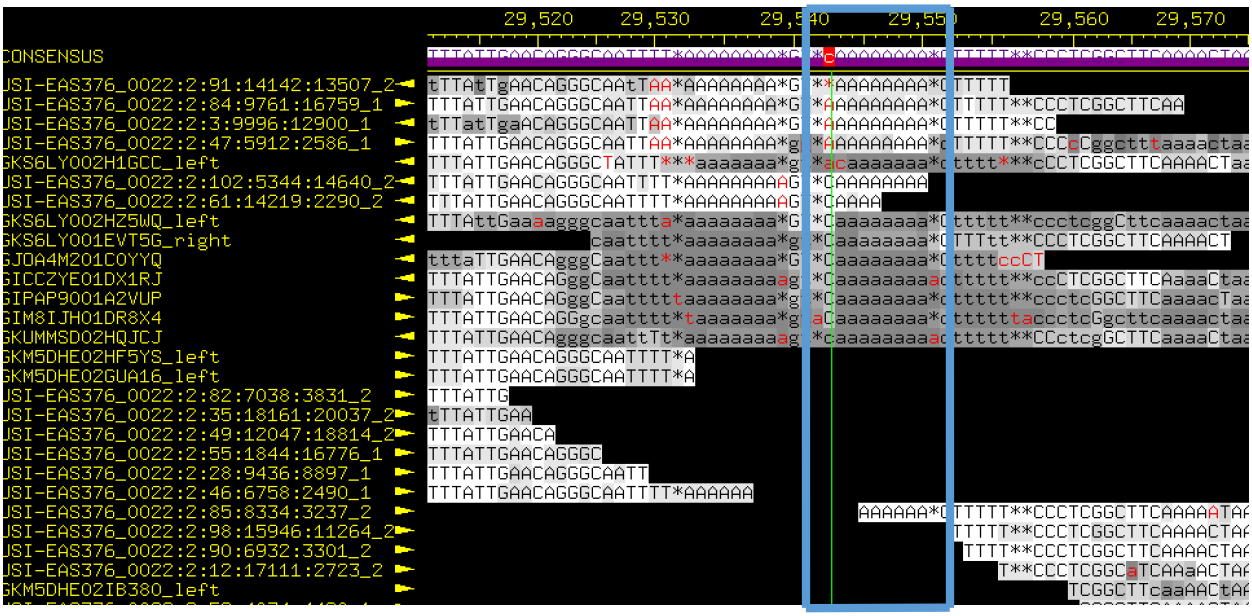


Figure 4. Ambiguous MNR in a HQDR at 29,542 (29,539 in file ace.2)
This base is an example of the more ambiguous HQDRs that resulted from a combination of low coverage and the possibility of mismapped reads. It is difficult to tell whether this base should have been changed to an A to match the number of As in the Illumina reads or should have remained a C according to the 454 reads. Because 454 reads are longer, they are less likely to be mismapped than the Illumina reads, so the consensus base should be a C.

Gap Resolution

A 20 bp gap, invisible on Assembly View, was discovered from bases 37,109-37,128. Because both sides of the gap were flanked with the same sequence, it seemed likely these reads should overlap. To resolve this, the assembly piece, a 'fake' read used to construct the assembly (KB457492:540000-640000), needed to be removed before the gap could be resolved. When this was removed, the single contig broke up into two contigs at base 68,707, creating contig A, of size 68,707 bp, and contig B, of length 31, 241 bp (Figure 5). This region had split apart because it was spanned only by the assembly piece (Figure 6). An examination of Crossmatch results in this region showed that the sequence around base 68,700 matches the sequence around base 22,000; both regions are tagged as an unknown repeat (Figure 8). Thus the reads supporting the

assembly sequence have likely mapped to the region around 22,000 (or elsewhere in the genome), and there is a lack of sufficient evidence to warrant changing the consensus sequence around base 68,700. Therefore, the assembly piece must remain in the contig to provide a consensus sequence for the region. However, because it was necessary to remove the assembly piece for the resolution of the gap spanning 37,109 -37,128 bp, an assembly piece needs to be reinserted into the final assembly to represent this repetitive region in the consensus sequence. This region was tagged and is recommended for further improvement.

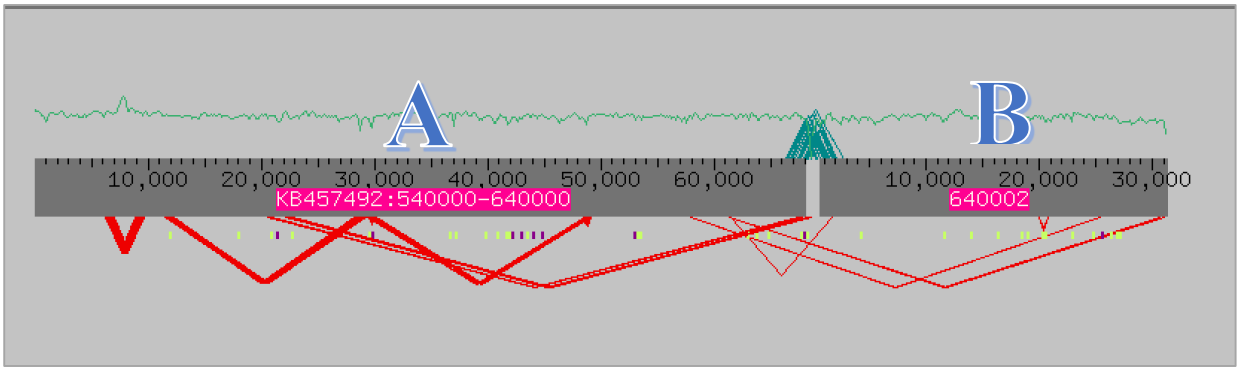
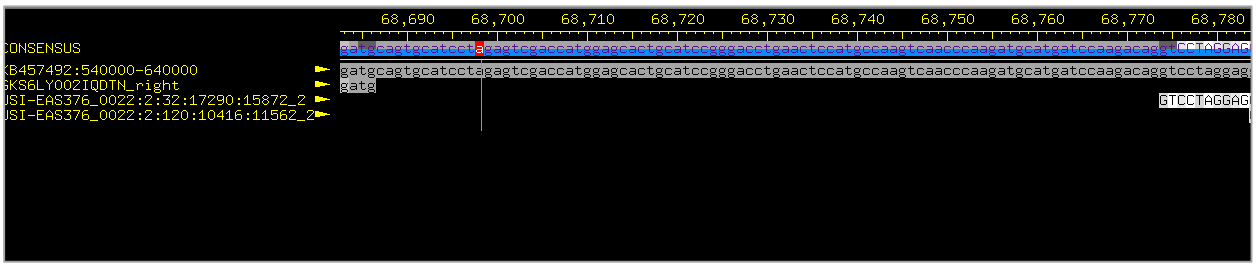


Figure 5. Break-up of the Single Contig on Removal of the Assembly Piece.
The removal of the assembly piece led to the contig splitting into two contigs of sizes 68,707 bp (contig A) and 31,241 bp (contig B) at the region of low coverage, where the assembly piece was essential. The gap is spanned by numerous mate pair reads.

Figure 6. Low Coverage Region around position 68,700.
The assembly piece is the only read spanning this low-coverage region, and the evidence used to create the assembly read is not shown. The consensus sequence must be left unedited here and tagged for further examination.



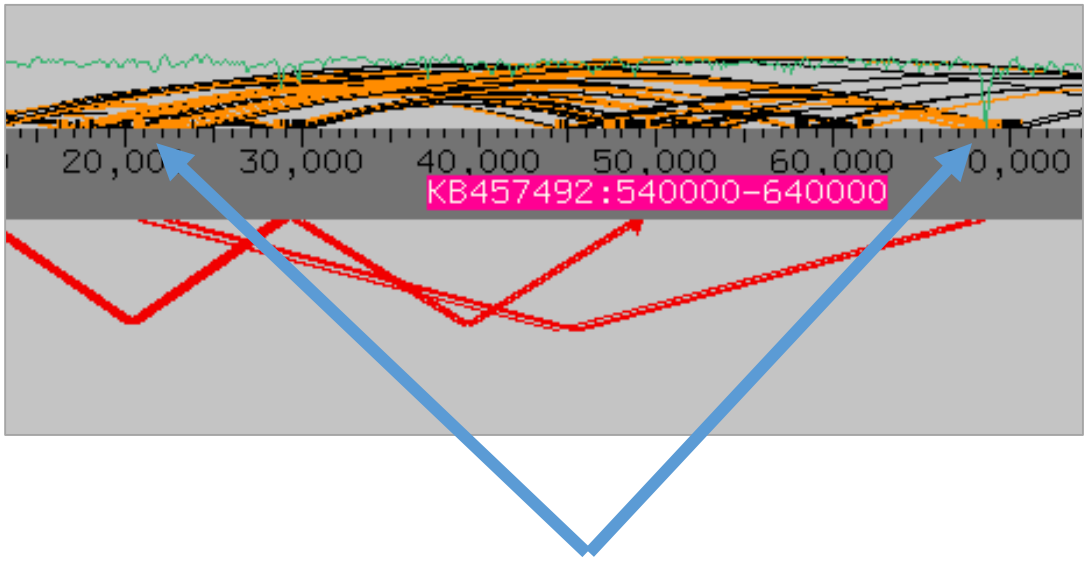


Figure 7. Crossmatch Results Linking Region 69,000 bp to Region 23,000 bp.
A closer examination of the Crossmatch results show that the low-coverage region is repetitive, mapping also to a region around 22,000 bp in the contig. The shorter Illumina reads that could have mapped to the low-coverage region may have mapped instead to the region around 23,000 bp due to the sequence similarity.

The 20 bp gap at position 37,109, however, was able to be resolved (Figure 8). To close this gap, the contig was torn apart at this region and Crossmatch was run to determine whether the sequences mapped onto each other at a repeat (Figure 9). “Search for String” was used to search for a sequence flanking the gap in both contigs: TTTTTTTTTTTTTATT. The comparison and alignment of these contigs is shown in Figure 11. By joining these contigs, the gap was resolved in contig 64006, now with a total of 99,896 bp (Figure 11).

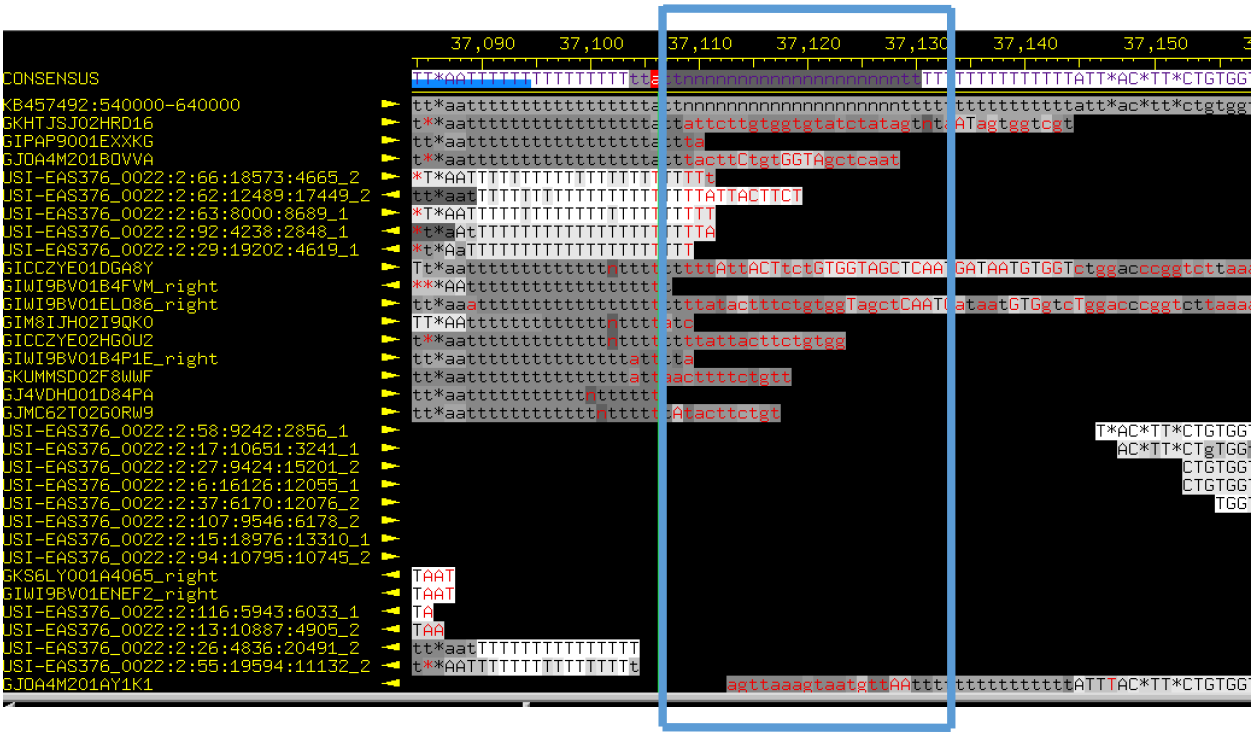


Figure 8. Gap Region from 37,109-37,128 bp.
The bases marked 'n' in the consensus sequence signify a gap in coverage. The repeating sequences on both sides of the gap, TTTTTTTTTTTTATT, suggest that the sequences to the left and right of the gap should be aligned on top of each other rather than next to each other.

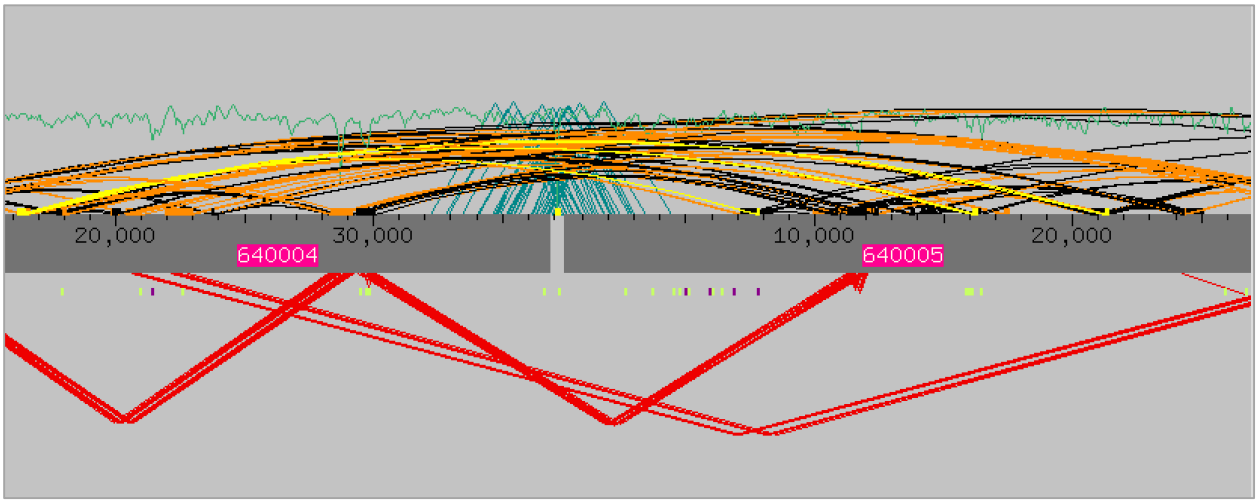


Figure 9. Contig 630003 Torn to Form Contigs 640004 and 640005.
The gap is flanked by repeats that could map onto each other, as seen by the results of Crossmatch in Assembly View.

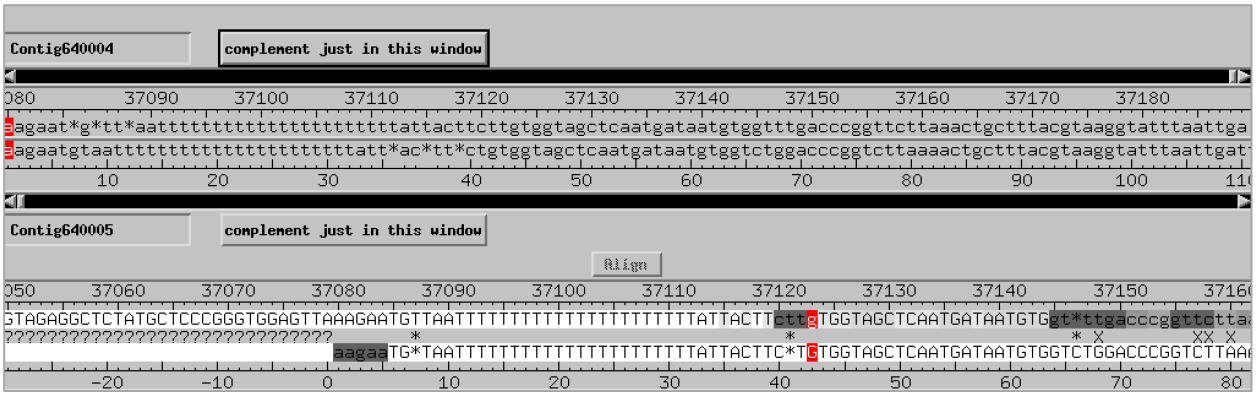


Figure 10. Alignment Results.
Contigs 64004 and 64005 are aligned by the sequence TTTTTTTTTTTTTATT, which appears on both sides of the gap. There are a few discrepancies, denoted as an 'X' between the alignment sequences. The long stretch of shared sequence, however, is strong evidence that these reads are from an overlapping, not sequential, region.

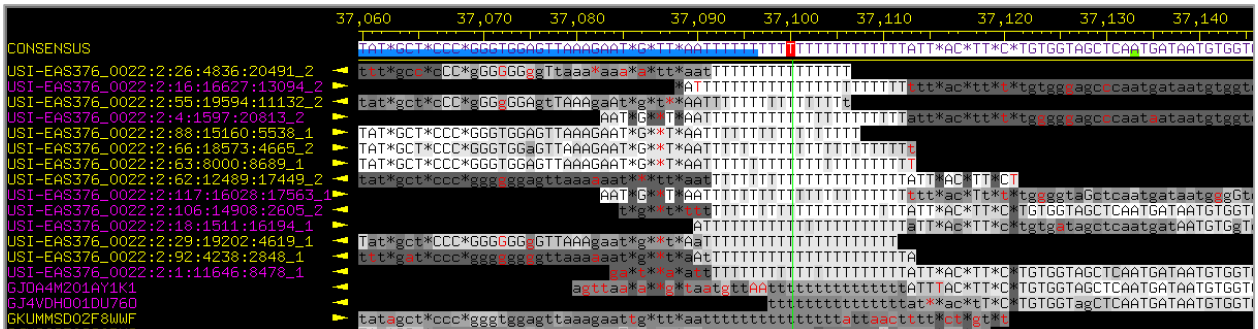


Figure 11. Resolution of Gap at 37,100.
The alignment of reads at 37,100 shows that the gap was successfully resolved in contig 640006.

Low Coverage Regions

A list of low coverage regions was generated using Consed to examine MNRs in these areas. It is necessary to search separately for MNRs in low coverage regions (defined as regions with less than 40 reads) because they might not be detected by the highly discrepant region search simply due to the lack of enough discrepant reads covering the region. Of the 267 MNRs in low coverage regions, 250 agreed with either the original consensus sequence or a region that had already been edited during the search for MNRs in the HQD regions. Of the seventeen regions that differed from the consensus sequence (Table 2, Appendix), five could be resolved by changing the consensus sequence to match the Illumina calls (Figure 12)

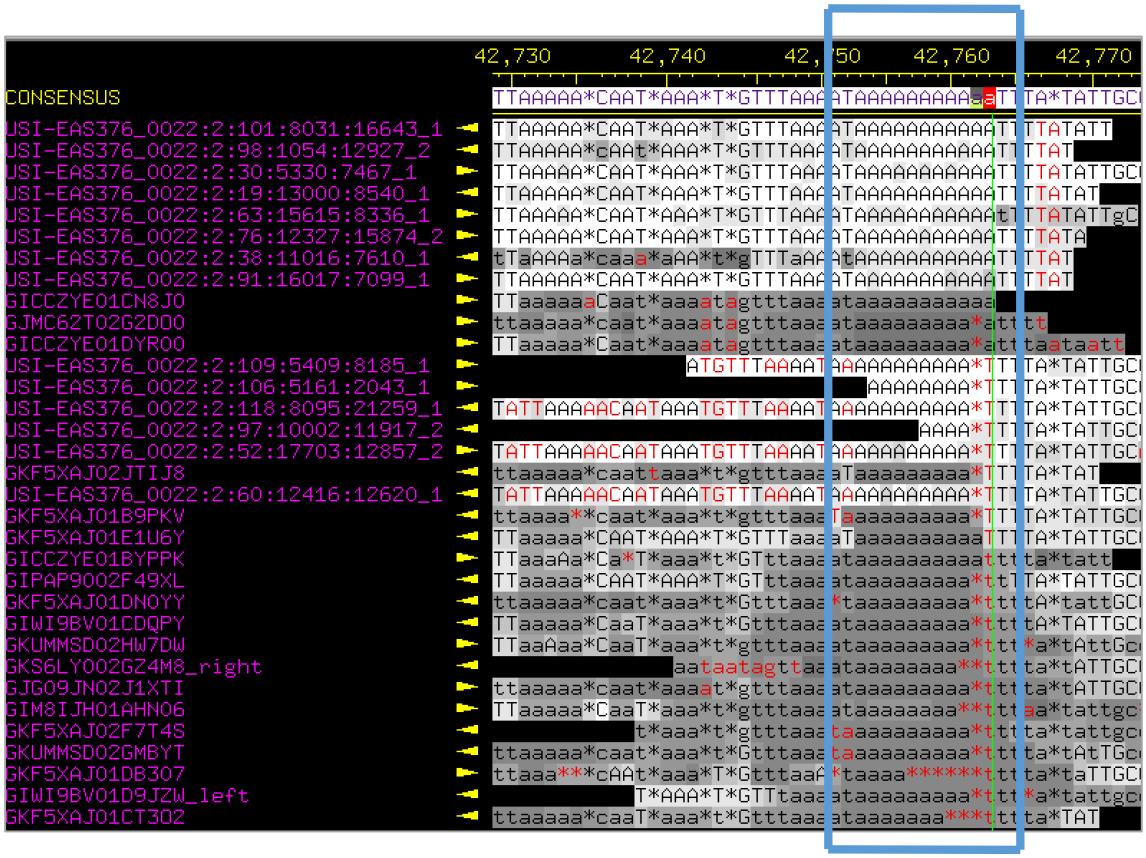


Figure 12. Modification of the Consensus Sequence in a Low Coverage Region.
This region had two As added to the monoA run due to the consensus among Illumina reads for 11 As in this region. There is ample evidence here for a change in the consensus sequence, despite being a region of low coverage.

The remaining 12 MNRs, however, occurred in regions of such low coverage that they could not be accurately evaluated without further data. The region around 48,815 bp is shown in Figure 13 as an example of one of these ambiguous regions. These regions were tagged as needing more data, with the idea that they could be resolved eventually if primers were created for Sanger sequencing. However, due to the number of repetitive regions, this proved challenging. Consed did not list any acceptable primer pairs for any of the regions listed in Table 2, explaining that any primers made would likely be of too low quality, or would anneal nonspecifically in these repetitive regions, rendering them useless (Figure 14). Thus, other methods besides Sanger sequencing might be necessary to resolve these regions.

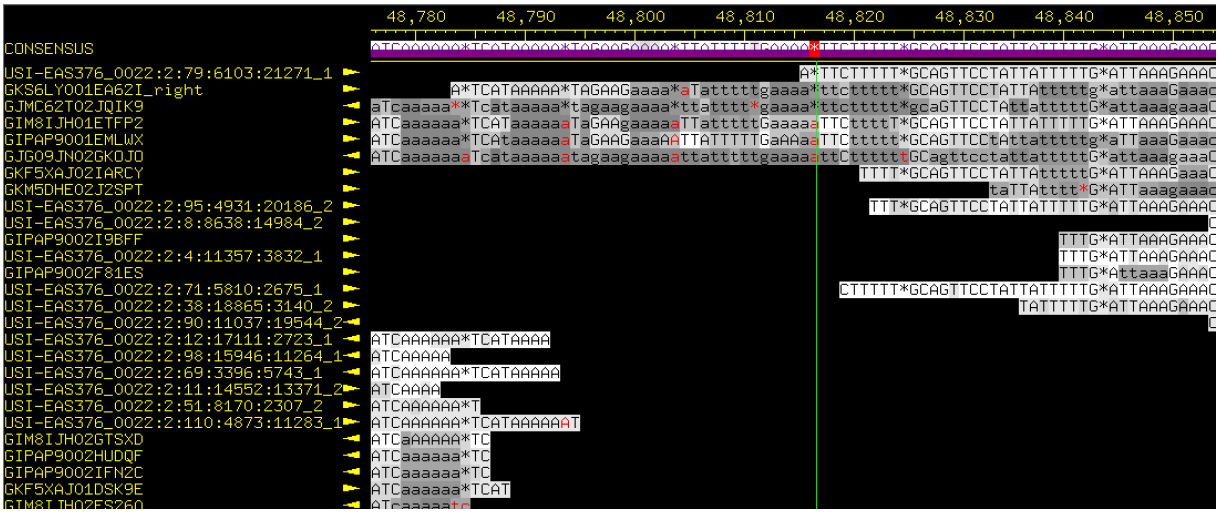


Figure 13. Low-Coverage Region Needing Additional Data.
These regions were tagged as needing further data.

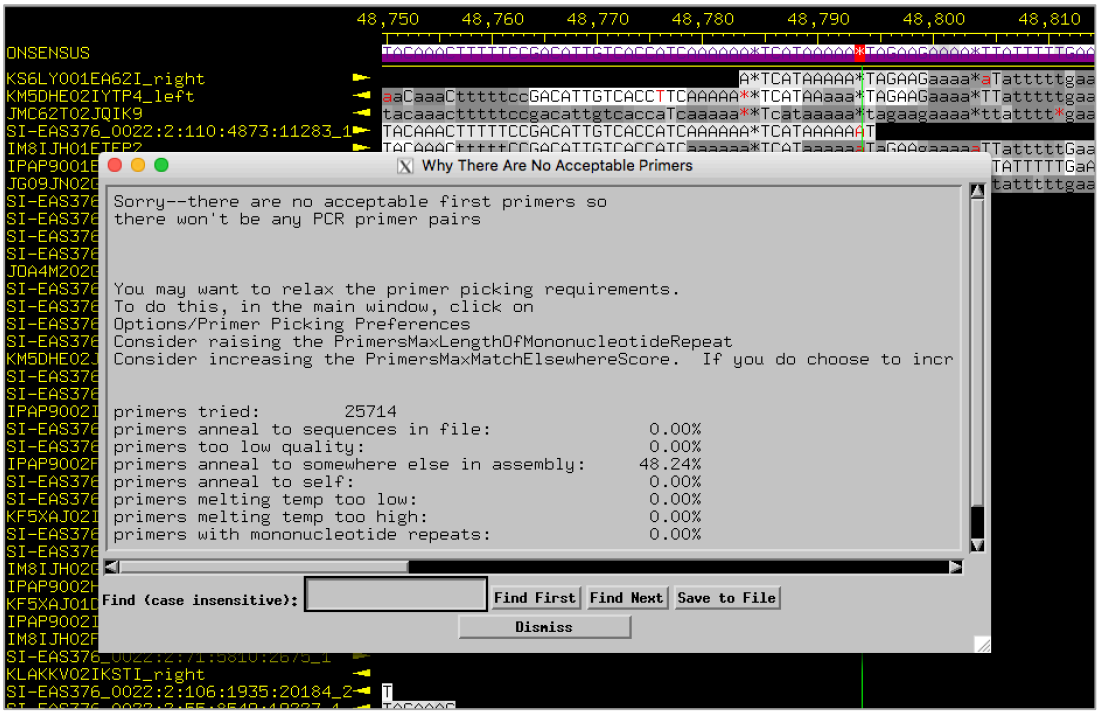


Figure 14. No Acceptable Primers Able to be Designed.
Due to the repetitious elements in the contig, no acceptable primers were able to be designed for the regions listed in Table 2. For the region from 68,488-68,575 bp, Consed declared that the primers were of too low quality; for the others, the primers would anneal elsewhere in the assembly.

Conclusion

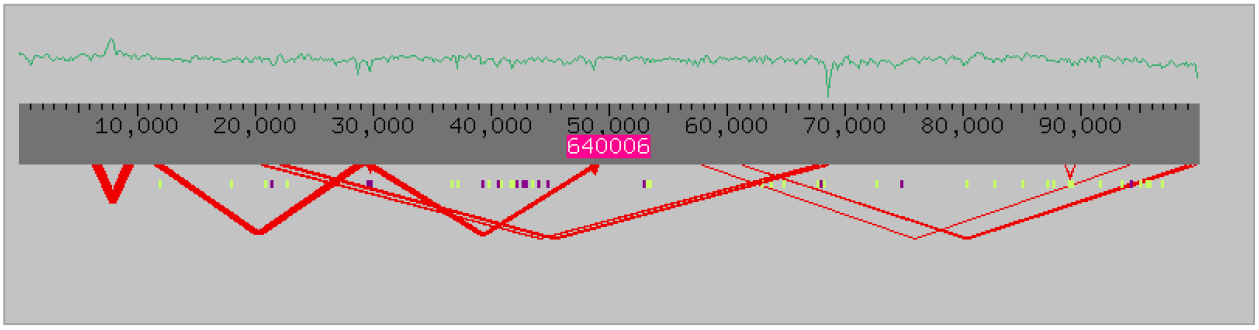


Figure 15. Final Assembly View of Finished Contig.

DFIC7492007, after finishing, appears as a single contig (Contig 640006) of 41,411 reads for a total length of 99,896 bp (Figure 16). Contig 640001 was not included in the final

assembly, as it contains the assembly read that was determined to have misalignments when the gap spanning bases 37,109-37,128 was resolved. However, work must be done to include data from the assembly read to restore the consensus sequence at the low-coverage region around 68,700 bp. A total of 55 base changes were made to the original consensus sequence, listed in Tables 1 and 2, due primarily to MNR errors by pyrosequencing. Many areas remain in the consensus sequence that need additional data, as described in Table 2. Designing primers was not possible due to the highly repetitive sequence of the contig. Before this sequence is annotated, it is recommended that some further data be gathered on the regions of low coverage. However, the repetitive areas of DFIC7492007 need not be a priority, as genes usually have a unique sequence.

Acknowledgements

Thank you to Dr. Elgin, Dr. Shaffer, Wilson Leung, and Lee Trani for their expertise and guidance in finishing this contig. Gratitude is also extended to Washington University in St. Louis and the Genomics Education Partnership for making this research possible.

Appendix

Table 1. List of High Quality Discrepancy Regions (HQDRs) Changed in Finishing Process

Region (bp) (ace.2)	Analysis	Conclusion
11,847	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 8 As
17,971	monoT run	Added a T due to the agreement of a majority of HQ Illumina reads for 8 Ts
20,957	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 7 As
21,452	monoA run	Added an A due to the majority of reads calling an A
21,455	monoT run	Added a T due to the agreement of HQ Illumina reads for 9Ts
22,634	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 8 As
29,767	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 8 As
29,881	monoT run	Added a T due to the agreement of HQ Illumina reads for 10Ts
36,609	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 15 As
37,229	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads
39,801	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 9 As
40,796	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 11 As
41,638	monoT run	Added a T due to the agreement of a majority of HQ Illumina reads for 8 Ts
41,895	HQDR	Added an A due the majority of HQ and Illumina and 454 reads
42,094	HQDR	Added a pad due to the majority of HQ 454 reads
42,117	HQDR	Added a T due to majority of HQ Illumina and HQ 454 reads
42,209	HQDR	Added an A due the majority of HQ Illumina and 454 reads
43,042	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 10 As
43,049	monoT run	Added a T due to majority of HQ Illumina and HQ 454 reads
43,055	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 6 As
43,113	HQDR	Added a C due to the agreement of a majority of HQ Illumina and HQ 454 reads

43,531	monoA run	Added a pad due to the majority of HQ Illumina reads for 6 As
43,967	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 8 As
44,960	monoT run	Added a T due to the agreement of a majority of HQ Illumina reads for 8 Ts
53,020	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 10 As
53,047	monoT run	Added a T due to agreement of HQ Illumina reads for 8 Ts
53,183	monoT run	Added a T due to agreement of HQ Illumina reads for 8 Ts
53,562	monoT run	Added a T due to the agreement of HQ Illumina reads for 9Ts
62,974	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 7 As
63,847	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 9 As
64,920	monoT run	Added a T due to agreement of HQ Illumina reads for 16 Ts
67,751	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 9 As
67,846	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 6 As
67,855	monoA run	Added a pad due to the majority of HQ Illumina reads for 6 As
68,353	monoA run	Added an A due the majority of HQ Illumina and 454 reads for 4 As
72,940	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 9 As
80,411	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 9 As
82,890	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 12As
85,242	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 9 As
87,347	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 8 As
87,891	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 9 As
89,074	monoT run	Added a T due to the agreement of HQ Illumina reads for 9Ts
89,371	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 9 As
91,716	monoT run	Added a T due to agreement of HQ Illumina reads and HQ 454 reads for 6As
93,671	monoT run	Added a T due to the agreement of HQ Illumina reads for 6Ts
94,360	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 9 As

94,425	monoT run	Added a T due to agreement of HQ Illumina reads for 8 Ts
95,043	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 13As
95,771	monoA run	Added an A due to the agreement of a majority of HQ Illumina reads for 11 As
96,002	monoT run	Added a T due to the agreement of a majority of HQ Illumina reads for 11 Ts

Table 2. List of Low Coverage Regions Changed or Tagged in Finishing Process

Region (bp)	Analysis	Conclusion
28715-28720	monoA run	Additional sequencing needed
28727-38732	monoA run	Additional sequencing needed
29532-29550	monoA run	Additional sequencing needed
39315-39326	monoA run	Added 2 As to match agreement of HQ Illumina reads for 12As
40609-40625	monoT run	Added 2 Ts to match agreement of HQ Illumina reads for 17Ts
42753-42763	monoA run	Added 2 As to match agreement of HQ Illumina reads for 11As
48789-48824	monoA run	Additional sequencing needed
68488-68514	monoA run	Additional sequencing needed
68533-68541	monoT run	Additional sequencing needed
68567-68575	monoA run	Additional sequencing needed
74684-74697	monoA run	Added 2 As for an agreement on number of As between reads
96893-96901	monoT run	Added a T to match agreement in number of Ts among Illumina reads