Sakura Oyama

Bio 434W

Dr. Elgin

February 12, 2016

Finishing DFIC7494001

Abstract

DFIC7494001 consists of the first 100 kb of the dot chromosome in *Drosophila ficusphila*. Since the project did not include any gaps, finishing involved resolving discrepancies in mononucleotide runs and checking low coverage and low consensus quality regions. Further PCR sequencing will be necessary at three regions in order to resolve discrepancies at mononucleotide runs in low coverage regions. The last 2500 bp were not finished, according to protocol.

Introduction

Chromatin in eukaryotes can be classified as either euchromatin or heterochromatin. Euchromatin is heavily enriched in genes and is associated with active transcription, as its loose packaging allows easy access for RNA polymerase. In contrast, heterochromatin is a tightly packed form of DNA that is generally inaccessible to RNA polymerase and thus silenced. Heterochromatin can be further classified into two categories: constitutive and facultative. Constitutive heterochromatin, found in all cell types of the organism, usually contains many repeats and can function at centromeres or telomeres, in addition to acting as an attractor for other gene-expression or repression signals. Facultative heterochromatin, found in some cell types of the organism, often contains pseudogenes, which are genes that have been silenced through mechanisms such as histone deacetylation.

The dot chromosome in *Drosophila*, also known as the Muller F element, appears to be mostly constitutive heterochromatin. Yet, despite being surrounded by high quantities of the repetitious sequences usually associated with heterochromatin formation, most of the approximately eighty genes found on the 1.3 Mb arm of the chromosome are expressed. The immediate goal of this project is to improve the sequence and annotate the F element of *D. ficusfila.* The long-term goal is to understand how these genes are able to alter chromatin structure in order to recruit transcription machinery and ultimately become expressed in a heterochromatic environment.

Initial Assembly

The initial assembly of DFIC7494001 displayed a single contig with zero low coverage regions (Figure 1). No gaps were observed.

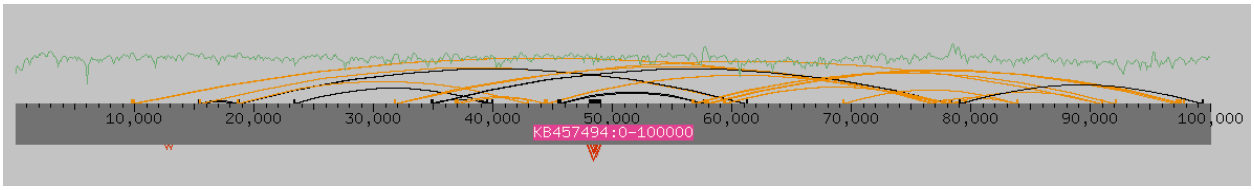

Figure 1: Initial assembly of DFIC7494001. Green lines represent high quality coverage. Red lines represent inconsistent mate pairs. Orange lines represent uncomplemented sequence matches while black lines represent complemented sequence matches. These orange and black lines correspond to repetitious sequences.

High Quality Discrepancies

Single nucleotide discrepancies were reviewed during this finishing process in order to ensure that all discrepancies at mononucleotide runs (MNR) were resolved. 454 pyrosequencing is prone to making errors at MNRs because the brightness of the light flash produced on incorporation of dNTPs becomes disproportionate to the number of nucleotides if the MNR is more than 6 bases long. Thus, it is necessary to ensure that Illumina reads are used to determine the consensus at MNRs. Illumina and 454 reads are easily distinguishable using Consed, as Illumina reads begin with “USI-." Using a high quality discrepancy search, seventy-eight MNRs

that required additional evaluation were found. Of these, forty-eight MNRs required a change in the consensus (Table 1).

Table 1: High quality discrepancy base changes

Location	Analysis	Edit	Source of Data
1257-1260	monoT run	+A	HQ Illumina reads
3714-3720	monoA run	+A	HQ Illumina reads
3918-3928	monoA run	+A	HQ Illumina reads
12713-12723	monoT run	+T	HQ Illumina reads
13556-13562	monoT run	+T	HQ Illumina reads
15407-15412	monoA run	+A	HQ Illumina reads
17021-17028	monoA run	+A	HQ Illumina reads
21131-21139	monoA run	+A	HQ Illumina reads
21673-21682	monoA run	+A	HQ Illumina reads
27646-27653	monoA run	+A	HQ Illumina reads
30391-30399	monoT run	+T	HQ Illumina reads
32254-32265	monoA run	+A	HQ Illumina reads
32679-32687	monoT run	+T	HQ Illumina reads
32689-32694	monoT run	+T	HQ Illumina reads
33656-33666	monoA run	+A	HQ Illumina reads
40418-40428	monoT run	+T	HQ Illumina reads
40635-40642	monoA run	+A	HQ Illumina reads, reads were misaligned
40643-40648	monoT run	+T	HQ Illumina reads, reads were misaligned
40794-40802	monoT run	+T	HQ Illumina reads
41081-41088	monoT run	+T	HQ Illumina reads
45305-45328	2 monoT runs divided by 2As	+TTT	HQ Illumina reads, reads were misaligned
45854-45863	monoT run	+T	HQ Illumina reads
47927-47932	monoT run	+T	HQ Illumina reads
48076-48083	monoT run	+T	HQ Illumina reads
53308-53316	monoA run	+A	HQ Illumina reads
53540-53547	monoA run	+A	HQ Illumina reads
55024-55030	monoA run	+A	HQ Illumina reads
60579-60590	monoA run	+A	HQ Illumina reads
63342-63350	monoT run	+T	HQ Illumina reads
63437-3445	monoT run	+T	HQ Illumina reads
65830-65939	monoT run	+T	HQ Illumina reads
65934-65945	monoA run	+A	HQ Illumina reads
67649-67657	monoA run	+A	HQ Illumina reads
72310-72321	monoT run	+TT	HQ Illumina reads, reads were misaligned
72981-72992	monoT run	+T	HQ Illumina reads
73049-73058	monoT run	+T	HQ Illumina reads
78711-78717	monoA run	A -> T	HQ Illumina reads

79206-79211	monoG run	+G	HQ Illumina reads
79752-79764	monoT run	+TT	HQ Illumina reads, reads were misaligned
82589-82594	monoT run	+T	HQ Illumina reads
83017-83022	monoT run	+T	HQ Illumina reads
85812-85818	monoA run	+A	HQ Illumina reads
86910-86919	monoT run	+T	HQ Illumina reads
88985-88993	monoT run	+T	HQ Illumina reads
92863-92874	monoT run	+T	HQ Illumina reads
94112-94123	monoT run	+T	HQ Illumina reads
95393-95410	monoA run	+A	HQ Illumina reads
96960-96970	monoT run	+T	HQ Illumina reads

The most common error encountered was the lack of one or more bases in the MNR due to Consed's determination of the consensus using poor quality 454 reads. These errors were resolved easily by manually replacing pads (*) with the correct number of bases as determined by HQ Illumina reads. For example, the monoA run at around base 3720 originally had seven As in the consensus due to 454 reads. However, close counting of the Illumina reads show that there should be eight As in the consensus (Figure 2).

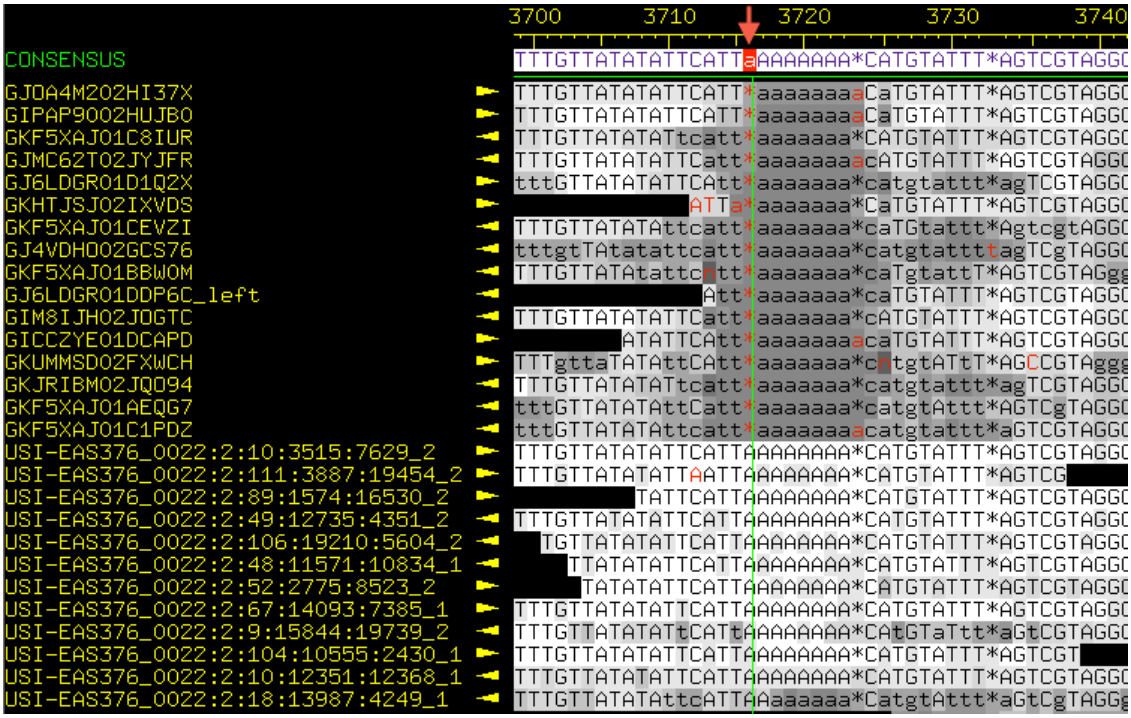

Figure 2: HQ Illumina reads used to add an A to the consensus at base 3716, as marked by red arrow.

Particularly in regions with multiple MNRs, many discrepancies were also caused by misalignments. These errors were resolved in either one of two ways. The first method involves manual counting of bases in HQ Illumina runs. For example, the original consensus at the monoA run followed by a monoT run at around base 40645 called for seven As and five Ts (Figure 3). However, counting bases in the HQ Illumina reads reveals that the monoA run actually contains eight As and the monoT run contains six Ts. The HQ discrepancies in the Illumina runs flanking both sides of the MNR region show that these reads were misaligned and were not reflected properly in the consensus. This method does not fix the misalignment but focuses simply on ensuring that the consensus sequence reflects the majority of HQ Illumina reads.

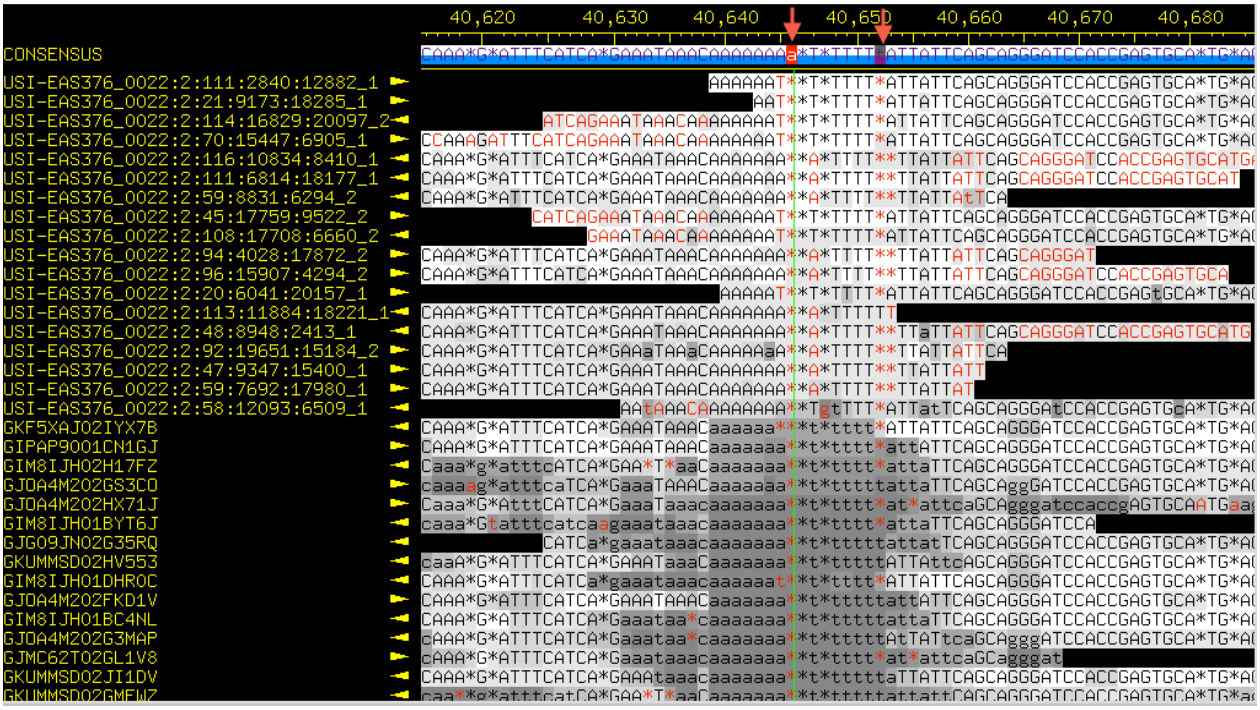

Figure 3: Example of MNR discrepancy due to misalignment resolved by manual counting of bases. An A was added at position 40646 and a T was added at position 40652, as marked by red arrows. Close counting reveals that majority of the HQ Illumina reads show eight As followed by six Ts.

The second method involves pulling out groups of similar discrepant reads, creating a miniassembly, and then realigning them to the consensus sequence. This method does not require manual edits to the consensus sequence as Consed will create a new consensus sequence based on the new aligned reads. For example, there are two clear sets of reads at the monoA run around position 3515 (Figure 4).

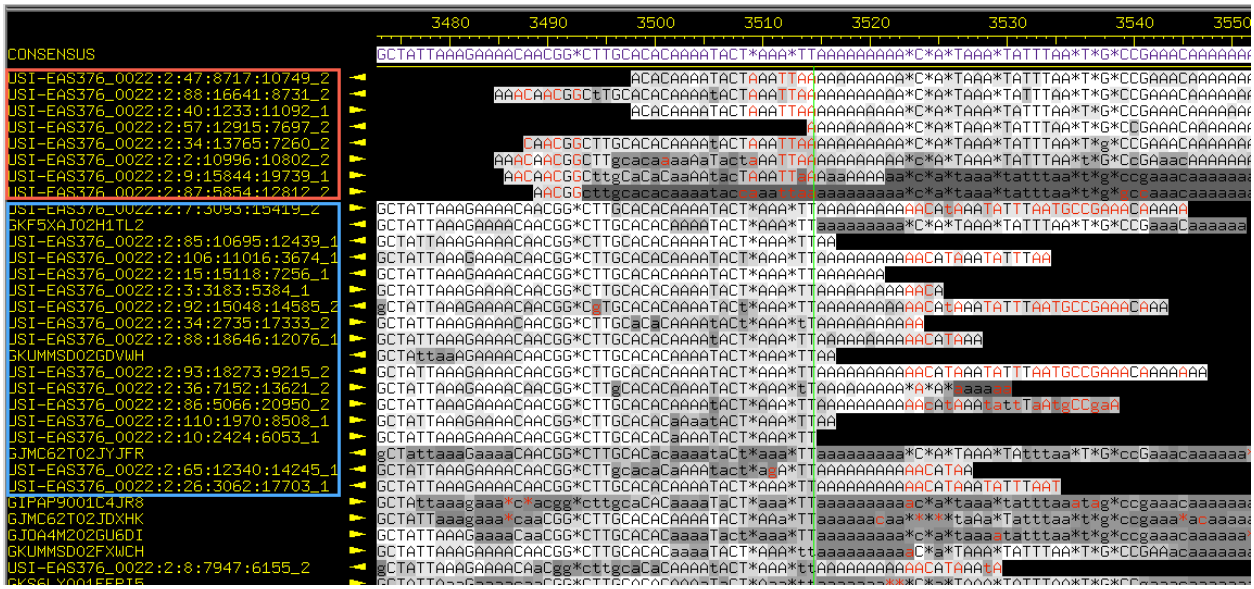

Figure 4: Original alignment of reads at monoA run around position 3515. Two distinct sets of reads are readily identifiable. One group, outlined in the red box, has discrepant left ends. The group below, outlined in the blue box, has discrepant right ends.

All reads at the top of the page with discrepant left ends were pulled out of the assembly as a group. The same procedure was used to pull out the set of reads below with discrepant right ends. A mini assembly was obtained for each set of reads in order to determine new consensus sequences. These smaller contigs were then compared to the main contig using assembly view (Figure 5) and complemented if necessary.

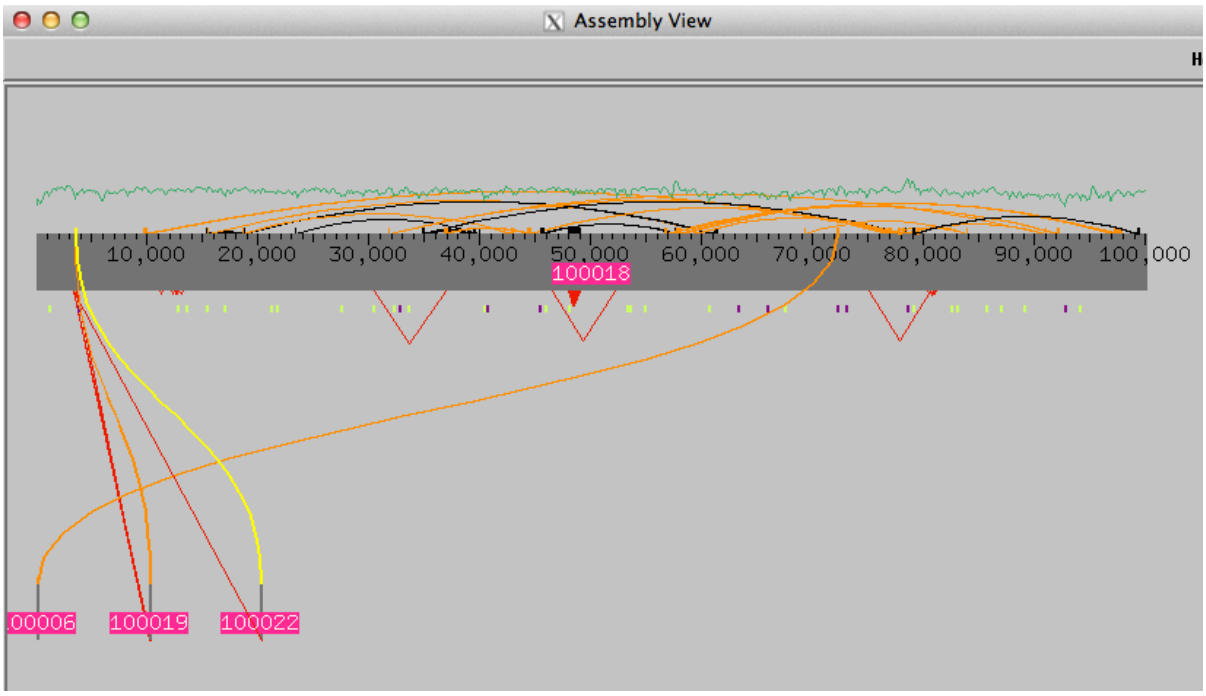

Figure 5: Assembly view of main contig with two smaller cotigs formed by pulling out distinct sets of reads from around position 3515.

Next, the consensus at the region around base 3515 was compared between each of the smaller contigs and the main contig. For both of the smaller contigs, the only discrepancy with the main contig was the addition of 2As (Figures 6 and 7).

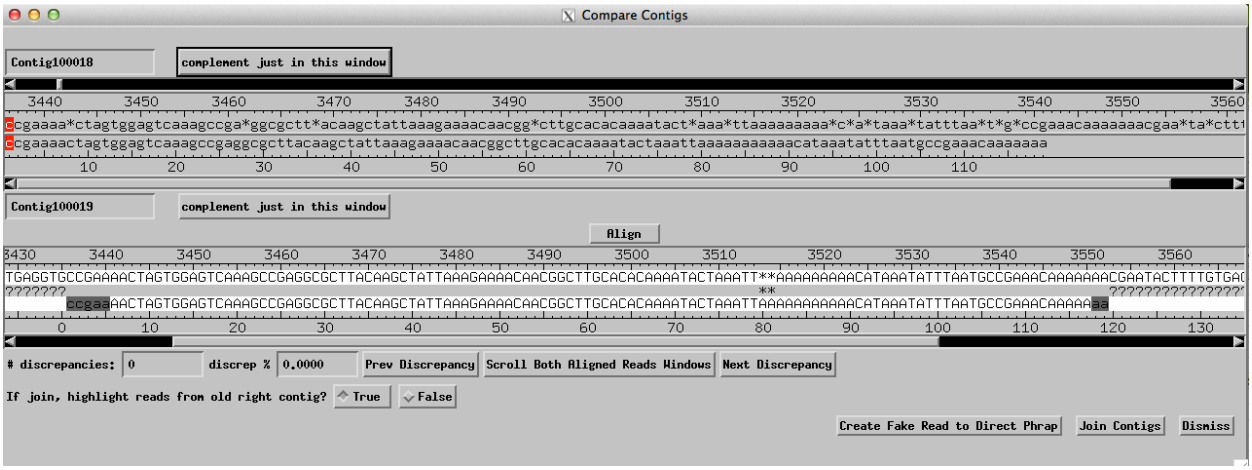

Figure 6: A comparison of contig 100019 with the main contig reveals that the monoA run at base 3520 is 2As longer in contig 100019.

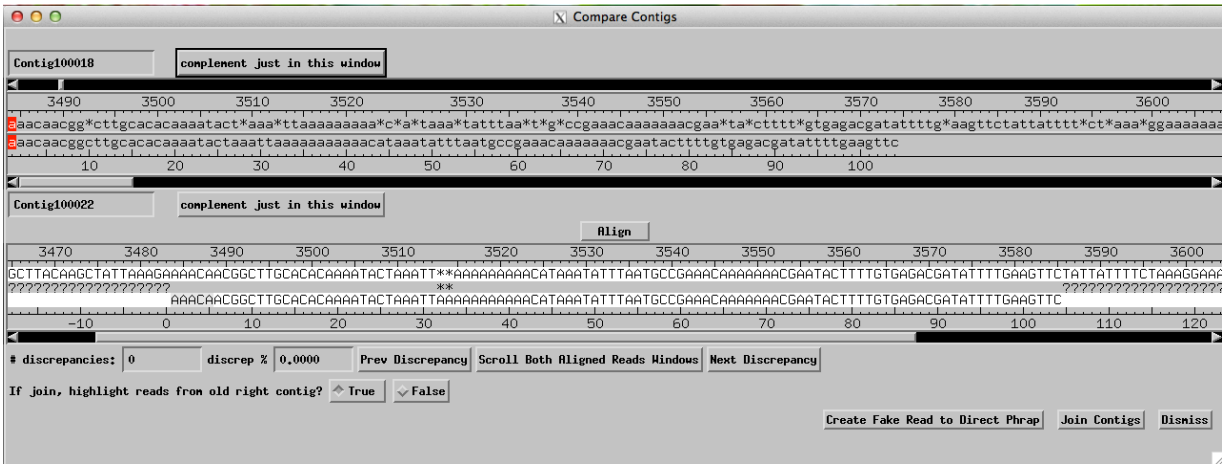

Figure 7: A comparison of contig 100022 with the main contig reveals that the monoA run at base 3520 is 2As longer in contig 100022.

Because the only discrepancy was found within the MNR, it is quite certain that both contig 100019 and contig 100022 are not mismapped. Thus, both of the contigs were rejoined to the main contig. This rejoining changed the final consensus of the main contig to reflect the additional As found in the pulled out reads (Figure 8). In this manner, misalignments can be solved, ensuring that the consensus sequence reflects the highest number of reads as possible.

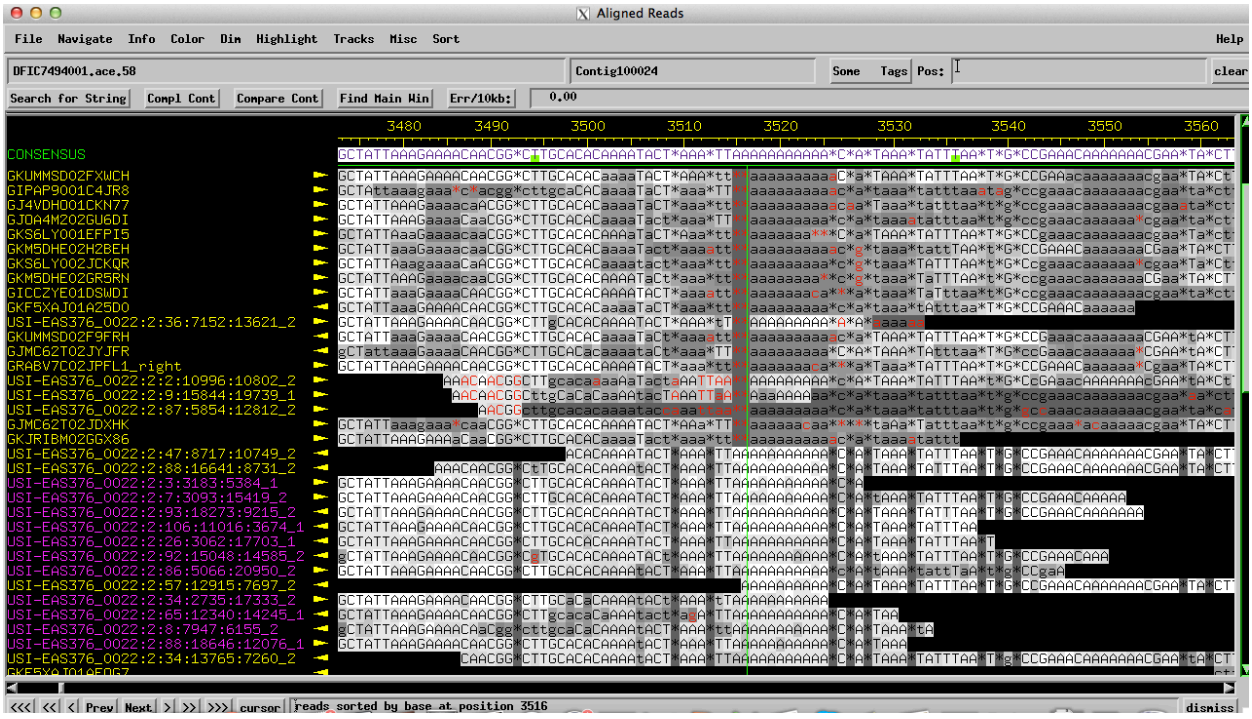

Figure 8: Changes produced by realignment of all reads. The consensus now includes the two additional As found in the reads that were initially pulled out of the main contig.

A similar method was used to pull out groups of similar discrepant reads from the monoA run at around base 95400, a repeat region (Figure 9). Four smaller contigs were formed that all mapped to the original monoT run using assembly view (Figure 10).

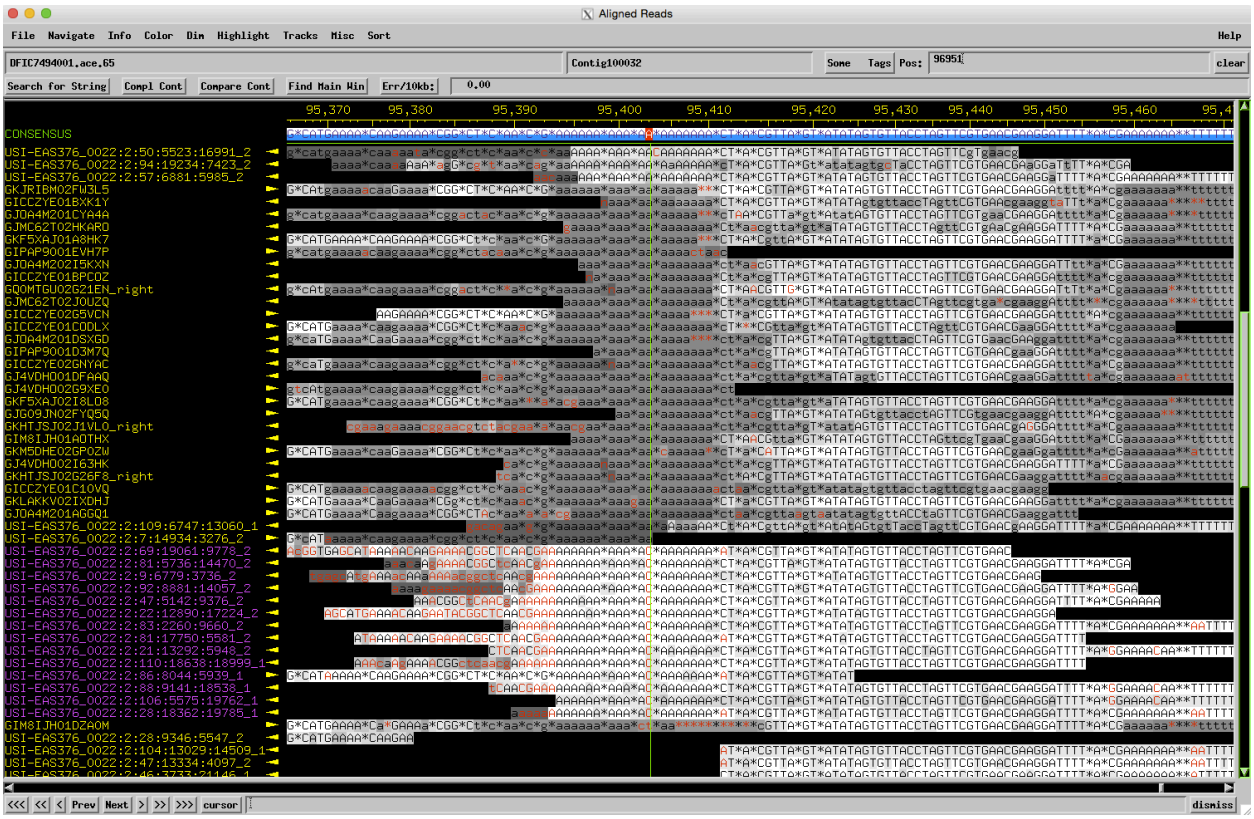

Figure 9: Original alignments of reads at the monoA run around base 95400. Several distinct sets of reads are immediately identifiable. These reads were pulled out as distinct sets and individual consensus sequences were created for each set of reads using miniassembly.

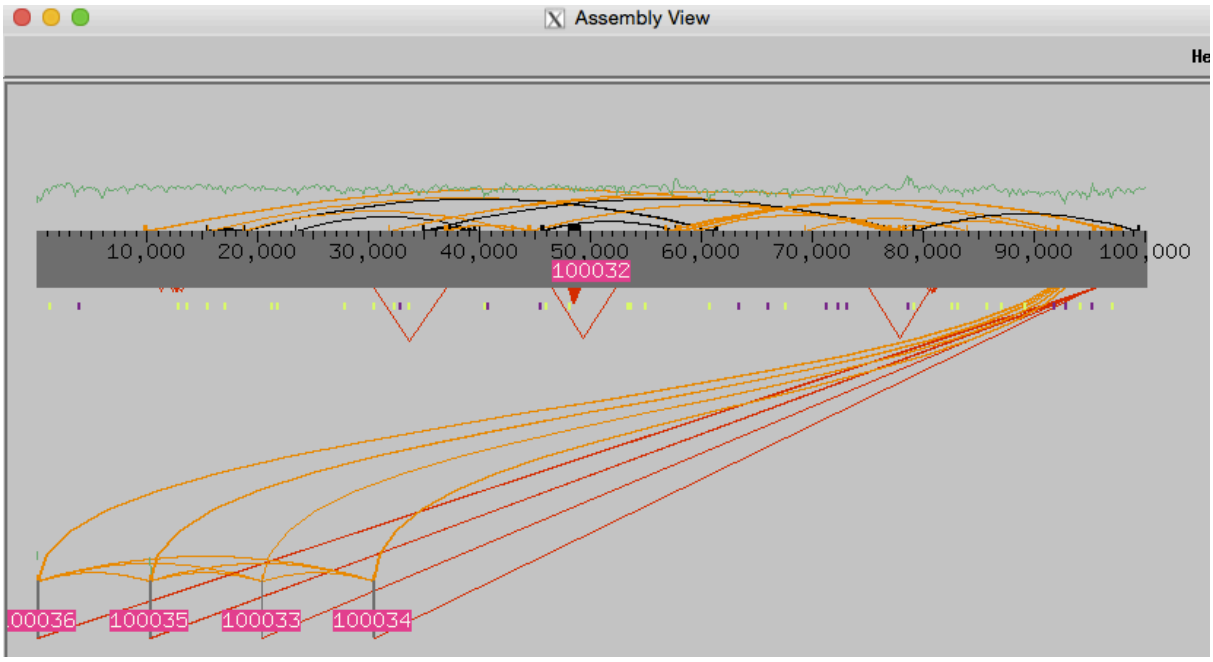

Figure 10: Assembly view of main contig with four smaller contigs formed by pulling out distinct sets of reads from around base 95400.

However, upon comparison, major discrepancies both within and beyond the MNR were discovered between the main contig and all four of the smaller contigs. For example, contig 100036 has a C in the middle of the monoA run (Figure 11). Upon inspecting the reads in contig 100036, it is clear that the C is not an error. The C is present as part of a high quality read in all of the reads contained in contig 10036 (Figure 12). However, none of the 454 reads in the main contig have a C in the middle of the monoA run. Because 454 reads are longer and thus more reliable for proper mapping, it is highly likely that contig 100036 has been mismapped to the repeat region. Similar processes reveal that contigs 100033-1000035 all contain a number of discrepancies from the consensus sequence on the main contig as determined by 454 reads. Thus none of these contigs belong to this region, and they cannot be put back into the main contig. A total of 34 reads were pulled out of the main contig in order to establish the number of nucleotides in the monoA run at base 95400.

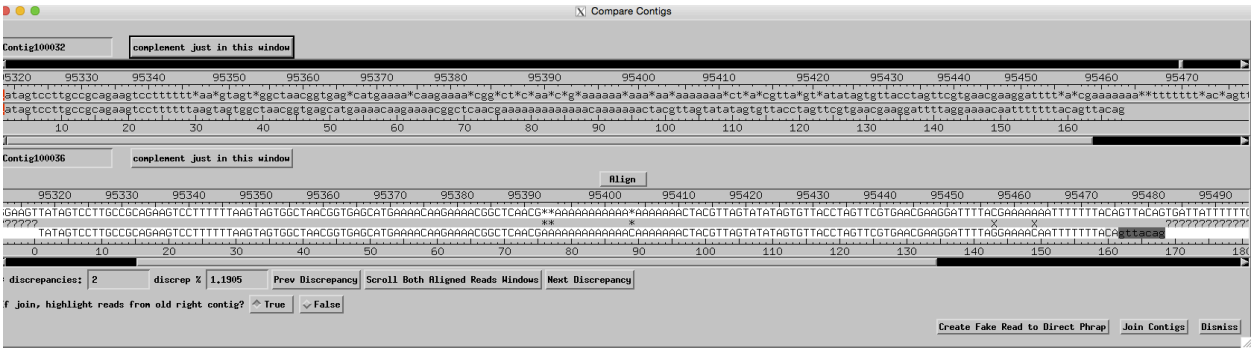

Figure 11: Comparison of contig 100036 to the main contig shows several discrepancies, marked by * or X, both within and outside of the MNR region.

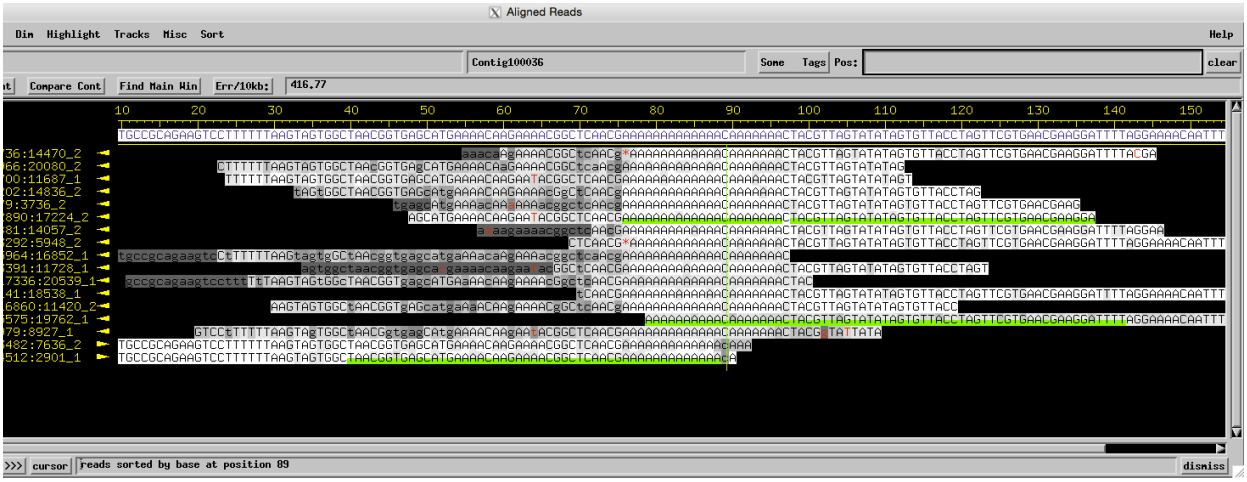

Figure 12: Miniassembly of contig 100036. The C in the middle of the monoA run in the consensus sequence of contig 100036 is established by numerous high quality reads and cannot be dismissed as an error. Contig 100036 does not belong in the main contig.

Several high quality discrepancies were caused by the presence of multiple copies of helitron transposons. In these regions, downstream and upstream reads were examined in order to determine whether any reads were obviously mismapped. For example, to investigate the high quality discrepancy at base 78718 (Figures 13 and 14), originally the last A of a monoA run, downstream reads were carefully searched for common discrepancies. This investigation revealed a high number of mismapped 454 reads, identified through high quality discrepancies starting at base 78791 (Figure 15). These reads were thus disregarded when determining the consensus at base 78718. Thus, the consensus at base 78718 was changed from an A to a T (Figures 16 and 17).

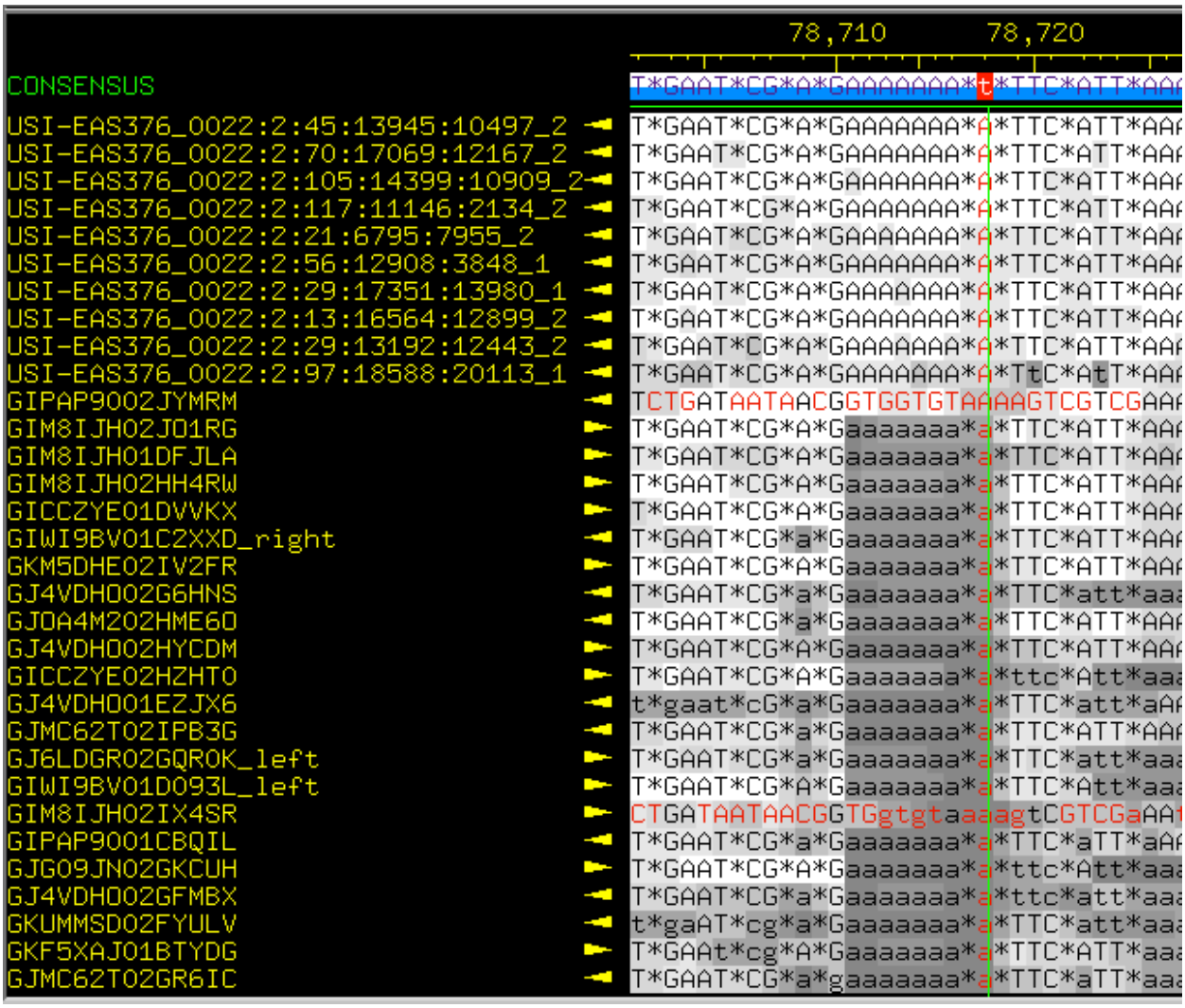

Figure 13: This group of reads has an A at base 78718.

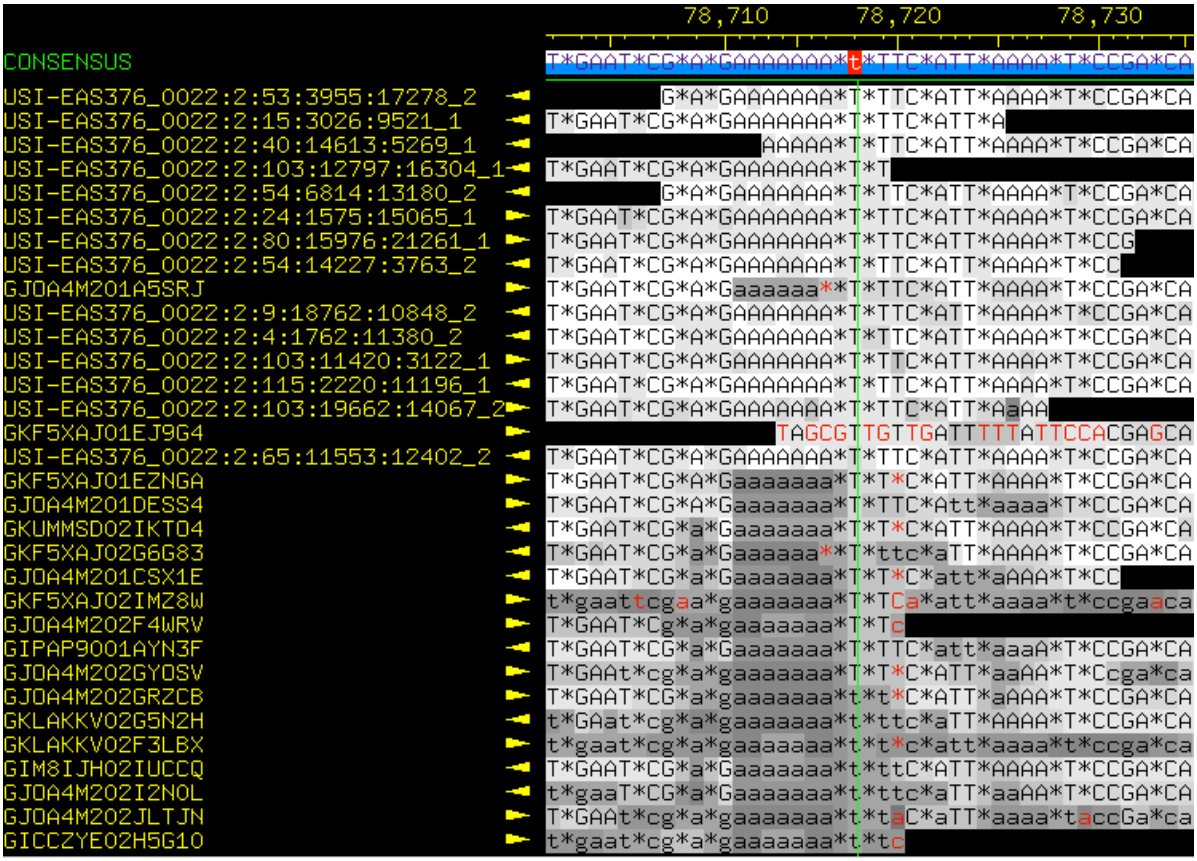

Figure 14: This group of reads has a T at base 78718.

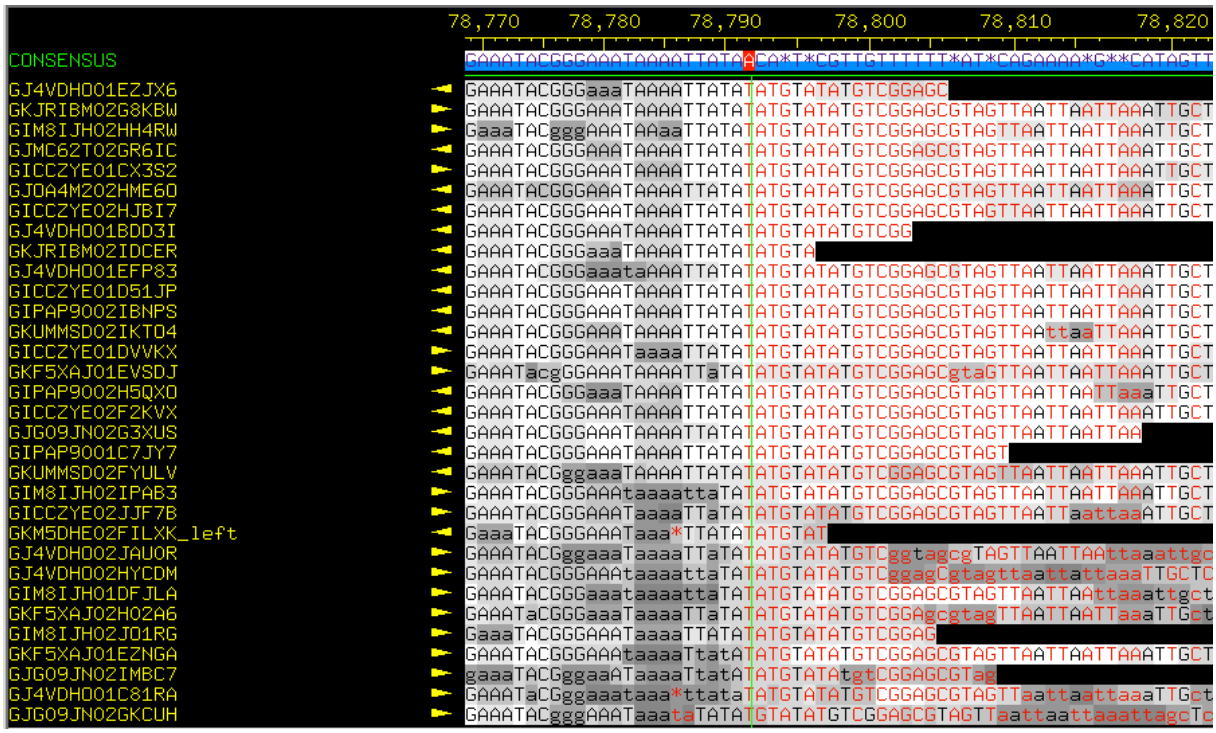

Figure 15: The high number of high quality discrepancies shows that these 454 reads are mismapped and can thus be disregarded when calling the consensus downstream.

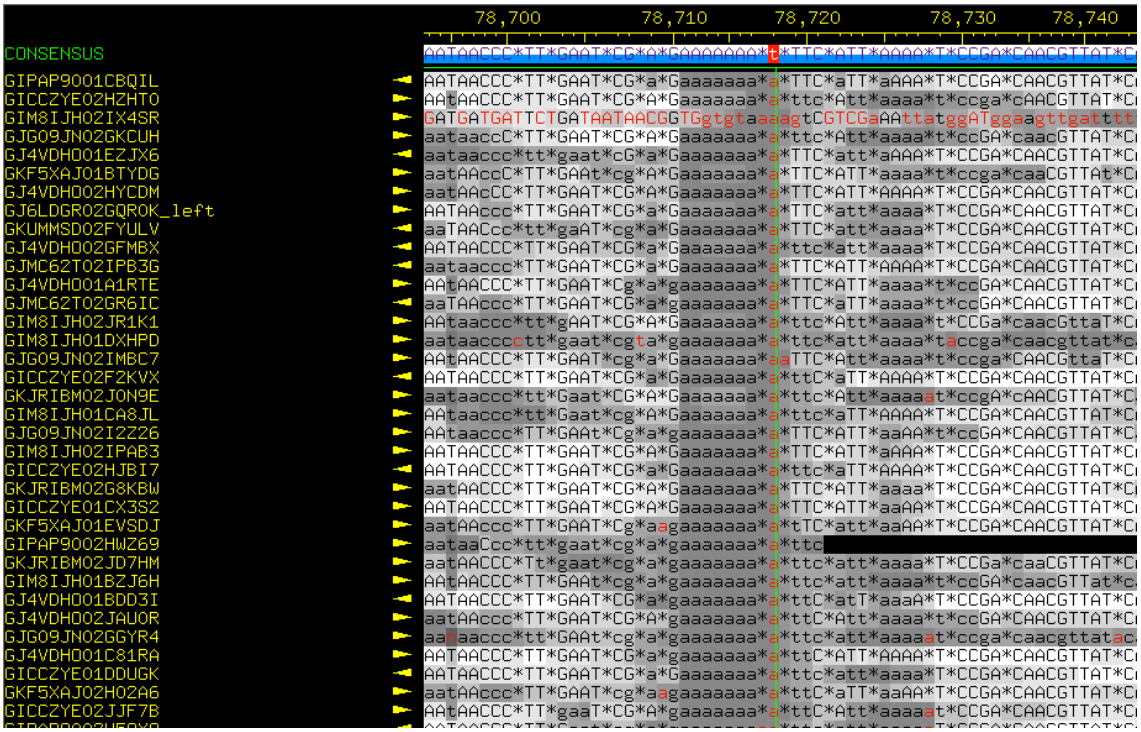

Figure 16: Properly mapped reads, as determined through an investigation of upstream discrepancies, were highlighted in purple. None of these reads, which all have an A at base 78718, are purple. These reads are likely mismapped and cannot be used to determine the consensus.

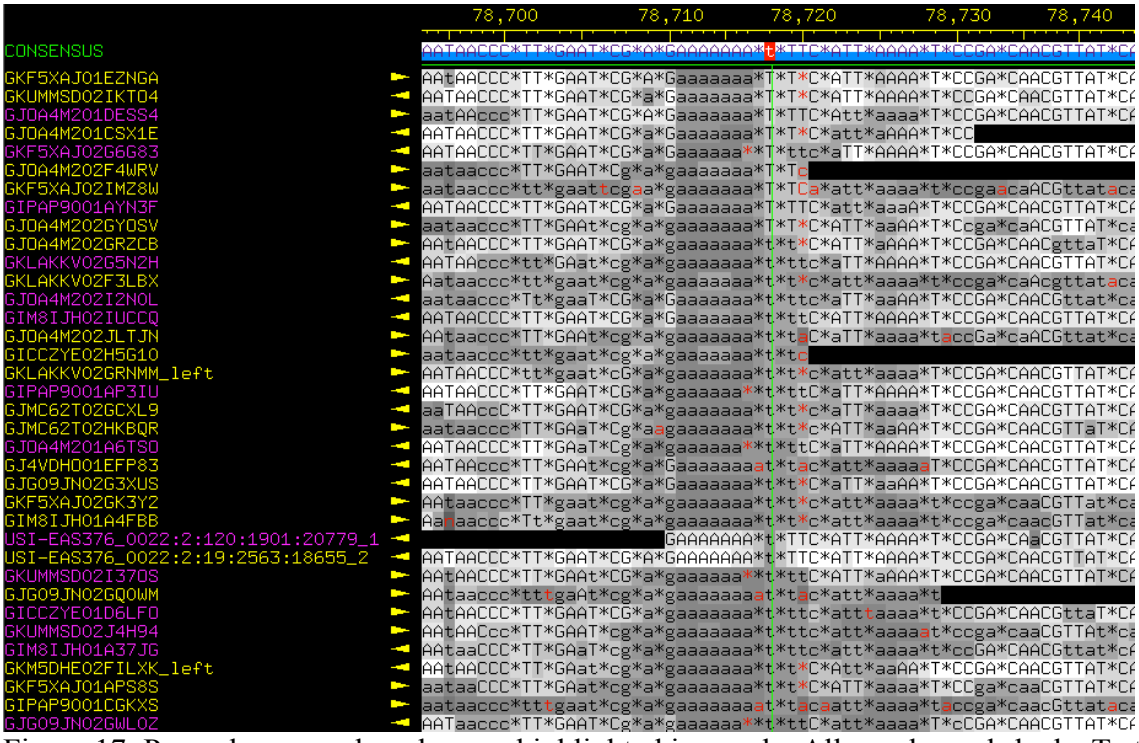

Figure 17: Properly mapped reads were highlighted in purple. All purple reads had a T at base 78718. Thus, the consensus was changed from an A to a T at base 78718 to match these results.

Two regions with high quality discrepancies associated with NMRs could not be resolved due to insufficient data. The first was a monoA run at around base 48700. Twenty-five HQ Illumina reads had fifteen As in the run while twenty-five HQ Illumina reads had sixteen As in the run. The site was unlikely to be a polymorphism because there were no other common high quality discrepancies in the area. It is unlikely that only one mutation has emerged in a region given the relatively long period of time necessary to establish a stable polymorphism. Thus, one set of these reads was likely mismapped. However, it was difficult to determine which set of reads was mismapped because the NMR is located right in the middle of a one kb repeat region. Unfortunately, there are no 454 sequences that stretch from the NMR to beyond the repeat region. Though the original consensus of sixteen As was kept, using low quality Illumina reads as a tiebreaker, additional sequencing is necessary in order to determine the true consensus. However, primers were not designed because the region is of low importance due to its location within a repeat.

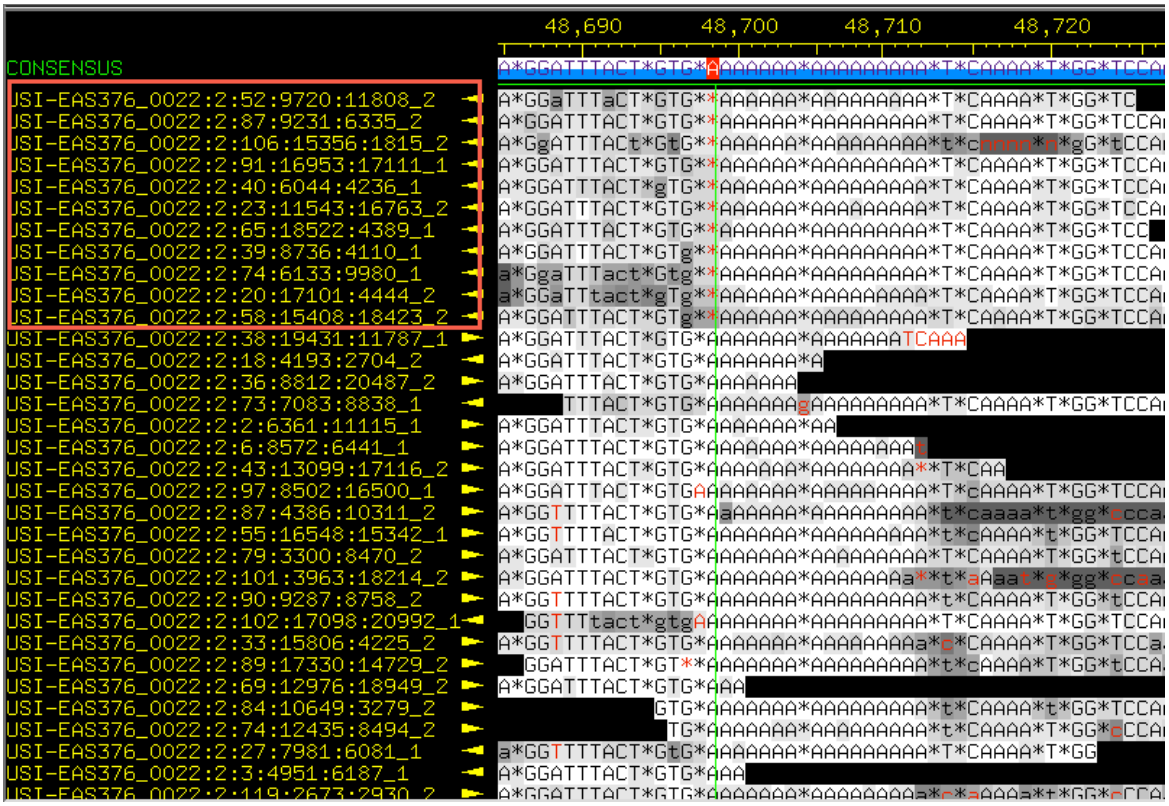

Figure 18: There was an even split of HQ Illumina reads between fifteen As and sixteen As. Reads outlined in the red box have fifteen As. Reads lower in the figure have sixteen As.

One monoT run at around base 94960 will require additional sequencing in order to determine the consensus because 4 HQ Illumina runs have 12Ts while four other HQ Illumina runs have eleven Ts. PCR primers were created using the protocol described in the next section.

Table 2: Additional PCR Sequencing for 94955-94966

Forward Primers	Oligo Sequence	Tm
94520-94549	TTTACCAAAAGTTCTTTAGTTATATATCAT	55
9435-94378	CCTTGCAGAGGGTAATATAAAA	55
Reverse Primers		
95058-95082	CAACATTACTTCTTCACATAATCAA	55
95298-95322	CTATAACTTCCACAATTCTCAAAAT	55

Overall, there were 48 MNRs identified through the high quality discrepancy search that required edits to the consensus. Twenty-seven edits involved the addition of one or more Ts and twenty edits involved the addition of one or more As. One edit involved the addition of one G.

No C runs required edits. The relatively small number of C and G runs requiring edits reflects the abundance of As and Ts on the dot chromosome, another trademark of heterochromatin.

Regions with Low Depth of Coverage (<40 reads)

There were 221 MNRs that had fewer than forty reads. These regions were all examined in order to determine whether the reads were of sufficient quality. Sufficient quality is defined as having at least 2 reads with a phred score of at least twenty for each base determined as properly mapped by the finisher. In order to have high confidence that the read is properly mapped, each read must have no more than one HQ discrepancy in the read compared to the final consensus (Figure 19).

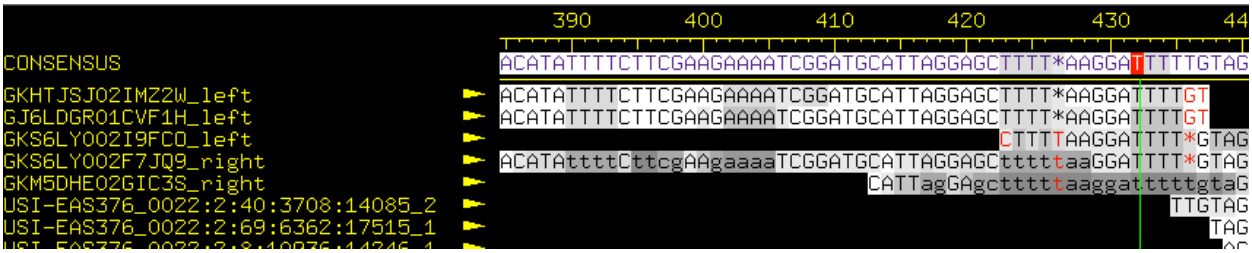

Figure 19: Low depth of coverage region dismissed after identification of misalignment. Though this region seems to display insufficient coverage, closer inspection revealed that the sequences are simply misaligned.

Two hundred ten of the 221 MNRs were of sufficient quality to support the consensus. The consensus at two MNRs was edited based on HQ Illumina reads, using the same procedures as described in the High Quality Discrepancies section. The consensus at four MNRs was edited to correct misalignment of HQ Illumina reads, using the same procedures as described in the High Quality Discrepancies section (Table 3).

Table 3: Changes in Regions Identified as Low Depth of Coverage

Location	Analysis	Edit	Source of Data
3515-3525	monoA run	+AA	HQ Illumina reads, reads misaligned
141015-11895	monoT run	+TT	HQ Illumina reads, reads misaligned
71227-71233	monoA run	+A	HQ Illumina reads, reads misaligned
91796-91814	monoT run with A in middle	+TT	HQ Illumina reads, reads misaligned

95196-95171	monoT run	+TT	HQ Illumina reads
96960-96970	monoT run	+T	HQ Illumina reads

Insufficient data was found to support or reject the consensus at fove MNRs (Figure 20). These regions were marked with a “data-needed” tag. Because three of the five MNRs were adjacent to each other, PCR primers were designed for a total of three separate regions.

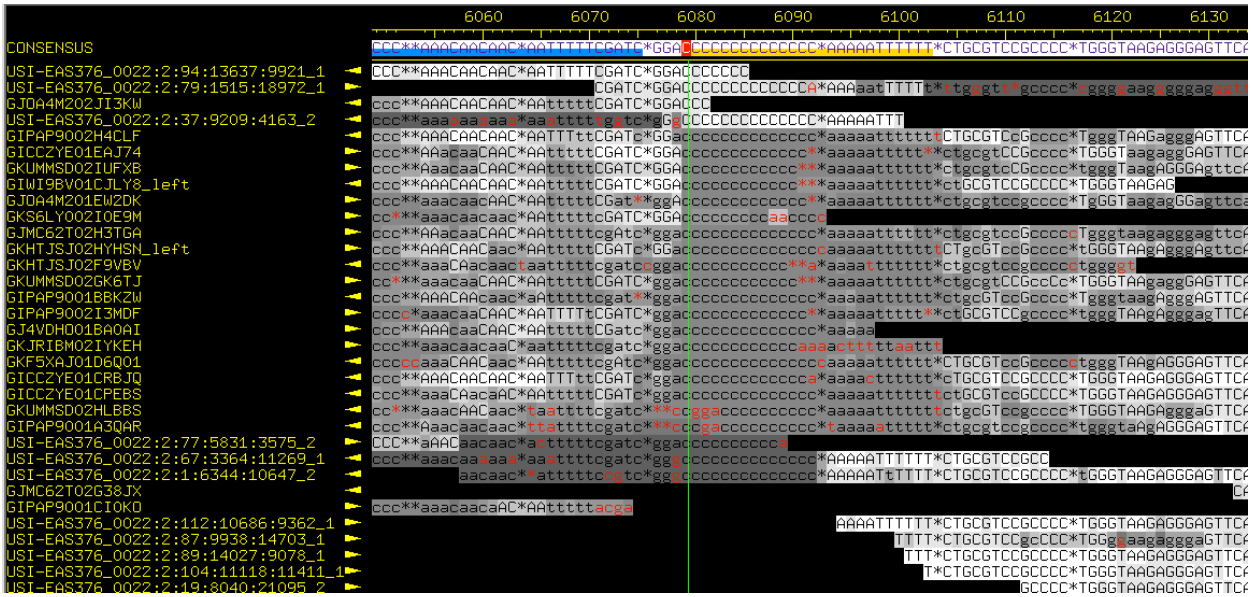

Figure 20: Insufficient data is available to support or counter the consensus at the successive mono C, A, and T runs around base 6090.

In order to select suitable primers, each end of the region of interest was marked. Consed then used an internal formula to obtain sequences for the recommended oligonucleotide primers. Two sets of mutually compatible primer pairs, taken from good quality sequences, that create amplicons as small as possible were chosen. When searching for primers to amplify the region between bases 6080 and 6103, forward primers 5990-6008 and 5801-5823 were identified as being closest to the desired amplification region. Reverse primers closest to the desired amplification region were identified as 6192-6213, 6192-6214, 6464-6486, and 6464-6487. Since 6192-6213 and 6192-6214 only differ by one base and thus cannot be considered unique primers, melting points were investigated in order to decide the best primer of the two. The melting points

of forward primers 5990-6008 and 5801-5823 were identified as 56°C and 55°C respectively. Because primer pairs with the smallest melting point differences possible are desired, reverse primer 6192-6213, with a melting point of 55°C, was selected over reverse primer 6192-6214, with a melting point of 57°C. The same procedure was used to select reverse primer 6464-6486 over 6464-6487. To summarize, forward primers 5990-6008 and 5801-5823 and reverse primers 6192-6213 and 6464-6486 were selected. The four pairwise combinations (pairs 1, 3, 5, 11 underlined in Figure 21) give the shortest possible PCR products of any four unique primers.

PCR Primer Pairs

pair #	distance between	primer1 contig	left	right	primer2 contig	left	right	melting p1	melting p2	pr
1	184	Contig100039	5990	6008	Contig100039	6192	6213	56	55	
2	184	Contig100039	5990	6008	Contig100039	6192	6214	56	57	
3	369	Contig100039	5801	5823	Contig100039	6192	6213	55	55	
4	369	Contig100039	5801	5823	Contig100039	6192	6214	55	57	
5	456	Contig100039	5990	6008	Contig100039	6464	6486	56	55	
6	456	Contig100039	5990	6008	Contig100039	6464	6487	56	57	
7	582	Contig100039	5589	5610	Contig100039	6192	6213	55	55	
8	582	Contig100039	5589	5610	Contig100039	6192	6214	55	57	
9	584	Contig100039	5990	6008	Contig100039	6592	6617	56	55	
10	584	Contig100039	5990	6008	Contig100039	6592	6618	56	56	
11	641	Contig100039	5801	5823	Contig100039	6464	6486	55	55	
12	641	Contig100039	5801	5823	Contig100039	6464	6487	55	57	
13	769	Contig100039	5801	5823	Contig100039	6592	6617	55	55	
14	769	Contig100039	5801	5823	Contig100039	6592	6618	55	56	
15	784	Contig100039	5388	5408	Contig100039	6192	6213	55	55	
16	784	Contig100039	5388	5408	Contig100039	6192	6214	55	57	
17	797	Contig100039	5990	6008	Contig100039	6805	6822	56	56	
18	854	Contig100039	5589	5610	Contig100039	6464	6486	55	55	
19	854	Contig100039	5589	5610	Contig100039	6464	6487	55	57	
20	956	Contig100039	5219	5236	Contig100039	6192	6213	57	55	
21	956	Contig100039	5219	5236	Contig100039	6192	6214	57	57	
22	982	Contig100039	5589	5610	Contig100039	6592	6617	55	55	
23	982	Contig100039	5589	5610	Contig100039	6592	6618	55	56	
24	982	Contig100039	5801	5823	Contig100039	6805	6822	55	56	
25	982	Contig100039	5800	5823	Contig100039	6805	6822	56	56	

Figure 21: List of possible primers generated by Consed for PCR sequencing of region 6080-6103

Similar procedures were followed to select two sets of primer pairs for a mono A run at base 27958 and a mono T run at around base 87945 (Table 5 and 6).

Table 4: Additional PCR Sequencing for 6080-6103

Forward Primers	Oligo Sequence	Tm (°C)
5990-6008	CCCCAAAGGAACAAATTTT	56
5801-5823	AAGGTTACAACGATACGATGTTA	55
Reverse Primers		
6192-6213	TATTATATGCTGGCACATTGAA	55
6464-6486	TTCAATGTGCCAGCATATAATA	55

Table 5: Additional PCR Sequencing for 27958-27620

Forward Primers	Oligo Sequence	Tm (°C)
27179-27207	CGTTTACAATATAGAATTAAAGTACGTTT	56
27360-27337	TGCATTTTGTTCTTGCG	56
Reverse Primers		
27694-27723	GAAAATTTAAGAGAATGAAATAATAATT	55
27895-27919	AAAGATTATATTTGACTGCTCGTAA	55

Table 6: Additional PCR Sequencing for 87945-87959

Forward Primers	Oligo Sequence	Tm (°C)
87700-87720	TCACTGCGTTACAAACTTTCT	55
87504-87521	TTGTTGCACGAGCTGTG	57
Reverse Primers		
88135-88156	CCATTTCAAGTTCGTCATAAAA	56
88358-88377	GGGTTTTAGTATTCGGGTGT	57

Low Consensus Quality

Though a low consensus quality search was conducted, all positions identified were associated with previously made manual edits.

Conclusions

The DFIC7494001 assembly began with many high quality discrepancies and low depth of coverage regions. I was able to resolve a majority of these areas. However, there are a few regions that will require future PCR sequencing in order to generate more data. A total of four sequencing reactions have been recommended. The final assembly largely resembles the initial assembly, as it was not necessary to resolve any gaps for this process.

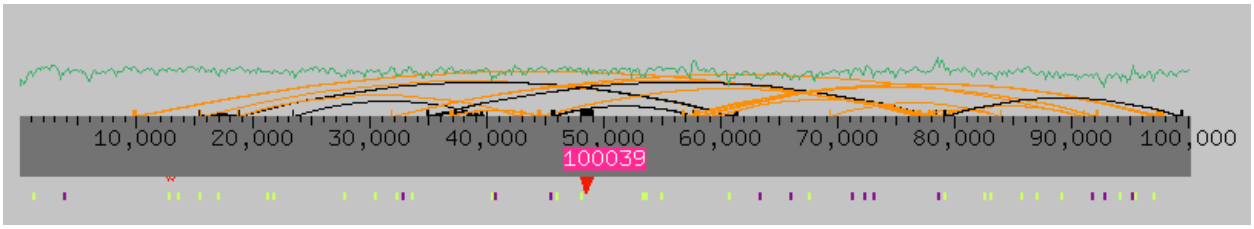

Figure 22: The final assembly largely resembles the initial assembly.

Acknowledgements

I would like to thank the BIO 434W faculty, Dr. Sarah Elgin and Dr. Christopher Shaffer for their support during the project. I would also like to express my utmost appreciation for Teacher's Assistant, Wilson Leung, for his patient guidance both in and out of class. Thank you also to Dr. Lee Trani, for his kind technical support.