Jaya Prakash
Dr. Elgin, Dr. Shaffer
Biology 434W
10 February 2017

Finishing of DEUG4927010

Abstract

The goal of this project was to prepare the DEUG4927010 contig—which covers the terminal 99,279 bases from the fourth chromosome of *Drosophila eugracilis*—for annotation. In the initial assembly of this project, there were 122 high quality discrepancies and 49 regions with low depth of coverage. Ninety one MNRs were present in low depth of coverage regions, and 67 MNRs were found near highly discrepant positions. There were no low consensus quality regions. Overall, 29 edits were made to the consensus sequence by using 'base counting' and 'Illumina prioritization' strategies, which are defined and explained in the High Quality Discrepancies (HQDs) section. Two potential SNPs were identified and tagged at positions 98,587 as well as 98,764. Finally, a third highly discrepant position at 99,103 was found within a repeat element of the contig. Primers for this site were designed to determine whether variation in the base was the result of deviation in the repeat, a SNP, or a misassembly. This project is now ready for annotation.

Introduction

Eukaryotic genomes are packaged into heterochromatic or euchromatic regions of the chromosome; this project examines the F element (dot chromosome) of the *Drosophila eugracilis*, which is predominantly heterochromatic. Heterochromatin refers to silenced regions of the genome that are highly condensed in a manner that suppresses transcription. Conversely, euchromatin refers to available regions of the genome that are less densely packed and can be accessed by transcription complexes. As a result, genes found in euchromatic regions typically

are more efficiently expressed than the same genes placed in heterochromatic regions. The smallest chromosome in the genome of *Drosophila* organisms, also known as the 'dot' chromosome or the 'F element,' is interesting because it is mainly composed of heterochromatin, enriched for silencing markers. However, genes found in these environments are still expressed as necessary for the survival of the organism. Therefore, the goal of this study is to map the genomes of fly species related to *Drosophila melanogaster*—such as *D. eugracilis*—to identify conserved regulatory motifs that might play a role in F element gene expression. The analysis could potentially allow for a better understanding of the gene regulation at work in the F element of *Drosophila* species. The *D. eugracilis* is valuable for this analysis because—based on its evolutionary history—it has fairly recently diverged from *D. melanogaster.* This allows for a more accurate representation of regulatory sequences shared between the species, as with less drift given the short evolutionary time. On the other hand, a more closely related species than *D. eugracilis* may not have had sufficient time to diverge, which means that shared elements may not necessarily be the result of conservation. Therefore, *D. eugracilis* hits a sweet spot.

The initial assembly from the DEUG4927010 contig compiled reads from 454 sequencing and Illumina technologies. The 454 sequencing reads have longer sequences, but they include multiple errors in regions with mononucleotide runs (MNRs). The 454 sequencing technologies are particularly prone to sequencing errors in MNRs that are longer than 5 bp. 454 sequencing data is generated by detecting the intensity of luminescence that results with the incorporation of each nucleotide to the read. For MNRs, multiple bases can be incorporated simultaneously. In cases that exceed five bp in length, the intensity of light does not effectively reflect the specific number of bases that have been added. Conversely, monitoring the color and intensity of fluorescence as a result of base incorporation generates Illumina data. However, in Illumina sequencing, the dye is added to a reversible terminator nucleotide so that only one base

can be added at a time. The color detected corresponds to the addition of only one nucleotide, so Illumina technologies are more accurate in correctly determining the length of MNRs. Due to 454 sequencing errors wtih MNRs, Illumina sequencing is used to more accurately determine the length of the MNR. With that said, reads from Illumina sequencing tend to be shorter than 454 reads, which can result in misassembly and the misplacement of the Illumina reads. Therefore, finishing the contig—after combining Illumina and 454 sequencing reads—involves locating as well as analyzing errors in gaps, highly discrepant, low coverage, and low consensus quality regions.

Initial Assembly

The initial Assembly View (Figure 1) included one contig, DEUG4927010, which covered the region of the *Drosophila eugracilis* dot chromosome between positions 810,000 and 909,279. Figure 1 shows the preliminary view of this project after running *Crossmatch*, which was used to detect sequence matches. All three of the regions with discrepant forward and reverse pairs (indicated by red lines in Figure 1) were found in regions of the contig with highly repetitive sequences (indicated by the orange and black lines in Figure 1). This indicates that the discrepant forward and reverse pairs may have resulted from being mapped to the wrong repeat element. The green line above the assembly view tracks the number of reads found in that region and indicates high coverage, fairly uniform, throughout the contig. The spike in coverage seen near the final 10 kb of the contig most likely corresponds to a misassembly at the terminal region of the genomic scaffold that this contig was derived from (Figure 2). Overall, there were 122 high quality discrepancies and 49 regions with low depth of coverage. There were no gaps, vectors, or low consensus quality regions detected within this contig.

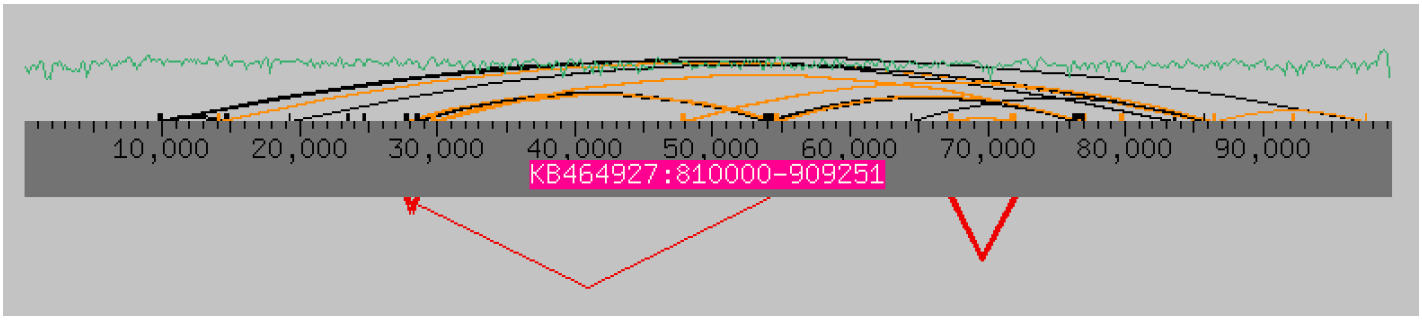

Figure 1: ***Initial Assembly View of Contig DEUG4927010***

The black and orange lines above the contig correspond to repeated regions. The green line at the top of the contig represents the number of reads found throughout the contig. The red lines below represent forward/reverse mate pair discrepancies. All of the forward and reverse pair discrepancies (red lines) in the DEUG4927010 contig were found in regions of repetitive elements (black and orange lines), which suggests that repeat elements may have resulted in mismapping of these reads.

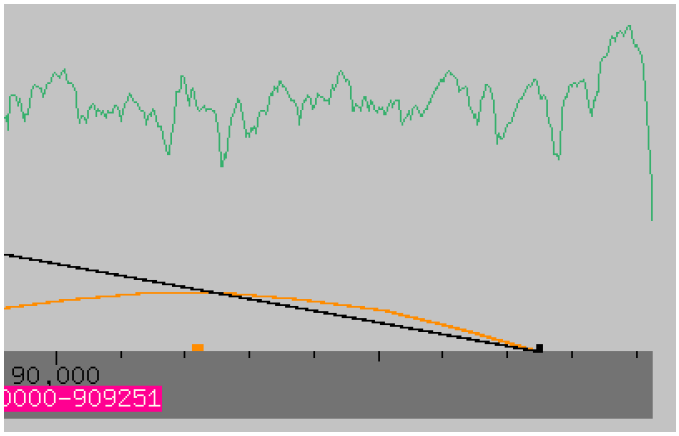

Figure 2: ***Spike in Coverage***

The green line tracks the number of reads found throughout the contig. The spike at approximately the final 10 kb of the contig suggests a misassembly, which is typical for contigs that cover the terminal region of a genomic scaffold.

High Quality Discrepancies (HQDs)

The contig was first evaluated for High Quality Discrepancies (HQDs), which refer to regions of the sequence where three or more high quality reads—each with a Phred score greater than or equal to 30—disagree with the consensus sequence. The HQDs were found using a search for 'highly discrepant positions' in Consed and then scanned for MNRs. There were 110

HQDs found through Consed, and 64 of these regions contained MNRs. Out of these 64 regions, 26 positions in the consensus sequence were edited (Appendix A and Appendix C). The length of MNRs suggested by 454 sequencing technology is often erroneous, so HQDs were resolved by evaluating MNRs where possible.

Of the 64 HQDs that were associated with MNRs, 40 regions had Illumina reads that completely or predominantly agreed with the consensus sequence; thus no edits were necessary in these regions. In these cases, the HQDs that were detected by Consed typically had multiple reads from 454 sequencing technologies that disagreed with the alignment of the consensus sequence. However, both Illumina and 454 sequencing reads generally had recorded MNRs of the same lengths. The discrepancy was typically due to only a few reads in disagreement with the consensus sequence or the misplacement of reads that agreed with the MNR length but were no longer aligned to the consensus sequence. Therefore, counting the number of bases in MNRs across all of the different reads was effective in evaluating these HQD regions. Since all—or the vast majority—of the reads agreed with the consensus sequence on the length of the MNR, there was no evidence to suggest that the consensus sequence needed to be edited. The term ‘base counting strategy’ will be used to refer to this method of counting the number of bp involved in the length of the MNR across all of the reads present at the region. An example of this strategy can be seen in Figure 3, which has an HQD spanning bases 38,844 to 38,850. Although 14 of the 454 sequencing reads disagree with the consensus, all of the Illumina reads and the vast majority of the 454 sequencing reads agree on a sequence of seven T’s. Therefore, the consensus sequence was not changed.

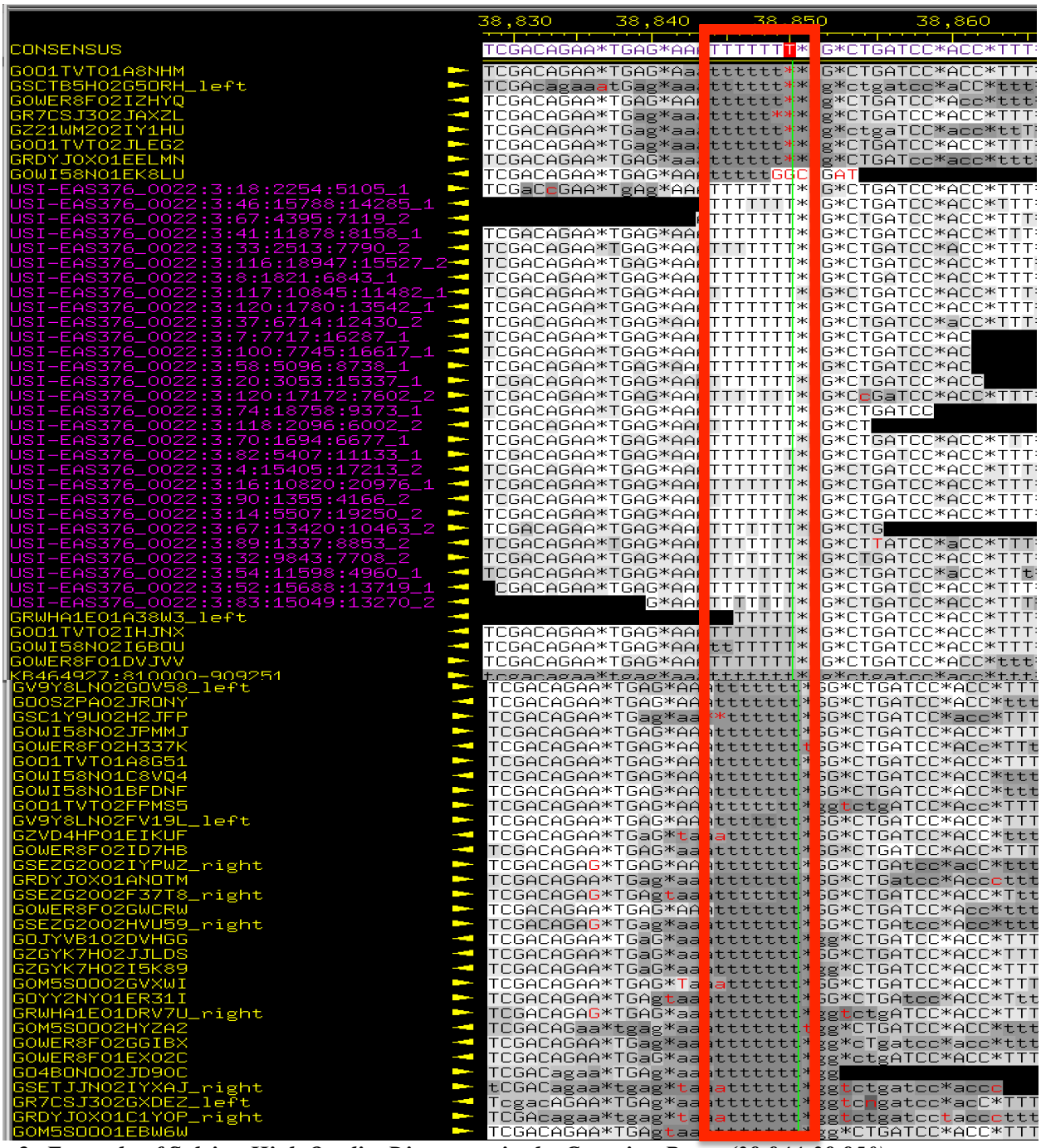

Figure 3: ***Example of Solving High Quality Discrepancies by Counting Bases (38,844-38,850)***

All of the Illumina reads (highlighted in purple) agree that seven T's (outlined by the red rectangle) should be aligned as shown in the consensus sequence. Fourteen of the 42 reads from 454 sequencing (highlighted in yellow) disagreed with the seven T length, but 28 agreed with the consensus sequence. Since the majority of the 454 Sequencing reads and all of the Illumina reads agreed with the consensus sequence, the seven T sequence was maintained.

In some cases, the length of MNRs indicated by 454 sequencing techniques did not agree with the length of the MNRs indicated by Illumina reads. Since the Illumina reads more

accurately determine the length of MNRs, data from Illumina reads were used to determine the consensus sequence. The term 'Illumina prioritization' will be used to refer to this method, which places a higher value on Illumina reads for evaluating MNRs. An example of this strategy can be seen in Figure 4 for positions 2,503-2,509. Although the majority of 454 sequencing reads disagreed with the consensus sequence in this region, all of the Illumina reads agreed with the consensus and were ultimately used in the final decision.

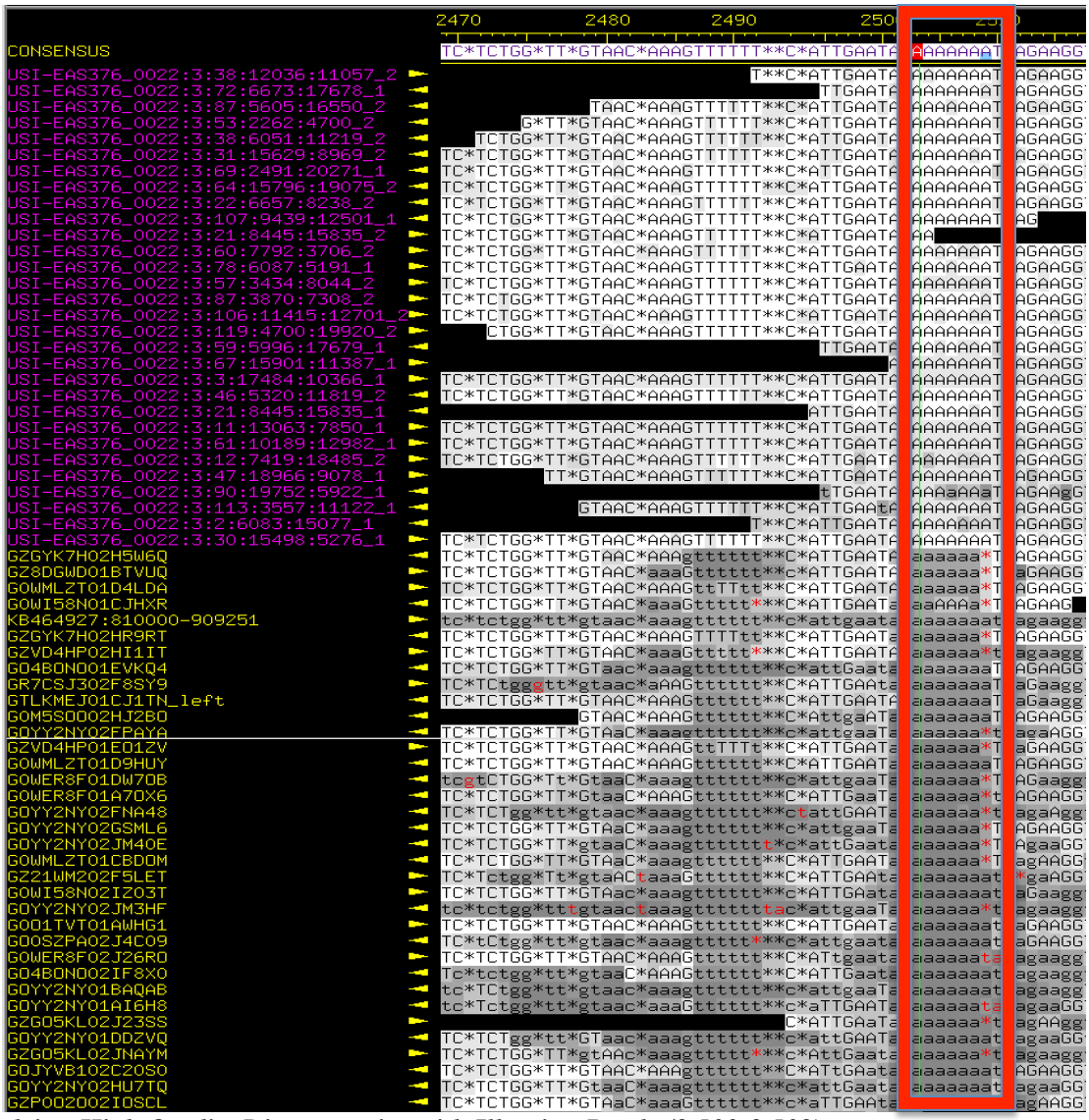

Figure 4: ***Solving High Quality Discrepancies with Illumina Reads (2,503-2,509)***

There were 22 reads from 454 sequencing (highlighted in yellow) that disagreed with the seven A length of the MNR (outlined by the red rectangle) and only 16 reads from 454 sequencing that agreed with the length for the MNR shown in the consensus sequence. Due to this disagreement, the high quality Illumina reads (highlighted in purple) were considered. Since all of the Illumina reads agreed with the consensus sequence, the seven A length of the MNR was kept in the consensus sequence.

The other 26 HQDs that were associated with MNRs required edits to be made due to disagreement between the high quality Illumina reads and the consensus sequence. In these cases, data on the lengths of MNRs from Illumina sequencing were considered more strongly than data from the reads of 454 sequencing data for reasons described above. However, many of these regions involved Illumina reads that were misaligned with the consensus sequence, which may have resulted from being incorrectly mapped to the 454 sequencing reads. Both the 'base counting' and 'Illumina prioritzation' strategies described earlier were used when determining whether a change to the consensus sequence was necessary. For example, region 23,765 to 23,775 shows 25 of the 31 Illumina sequences in agreement with a MNR of 10 A's and were aligned with the additional A under a pad (*) on the left side of the MNR. Three of the 31 Illumina sequences also agreed with an MNR of ten A's but were aligned with the additional A under an pad on the right side of the MNR. Despite issues with misalignment among the reads, 28 of the 31 Illumina sequences agreed that an MNR of ten A's should be present in the consensus sequence. This MNR length disagreed with the consensus sequence as well as with all of the 454 Sequencing reads. However, since the Illumina reads are characteristically better at determining MNR lengths, they were prioritized in determining the consensus sequence (Figure 5). Overall, 26 edits were made to the consensus sequence based on the list of HQDs determined through Consed (Appendix B).

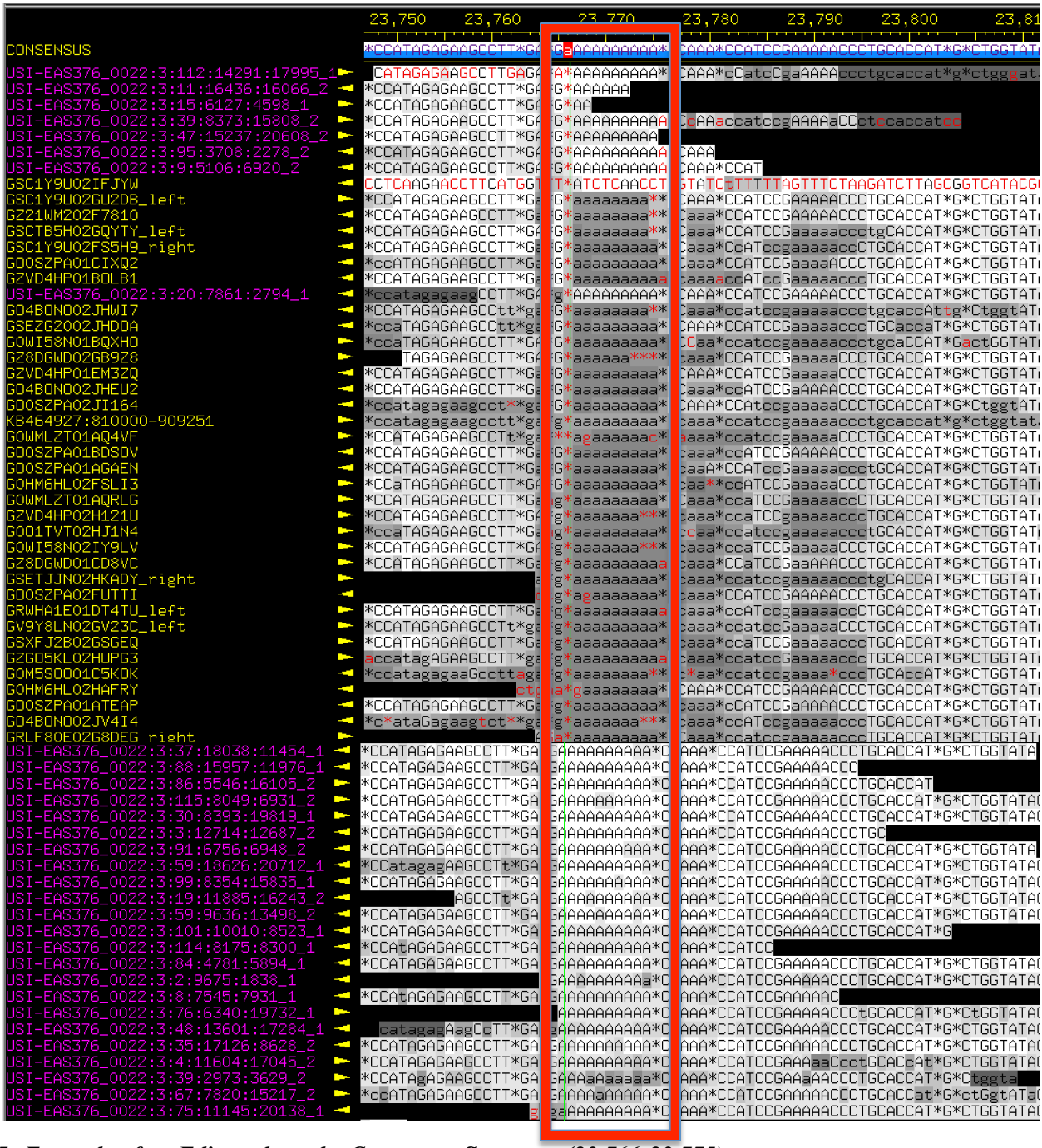

Figure 5: ***Example of an Edit made to the Consensus Sequence (23,766-23,775)***

The 'base counting' strategy was used in order to determine that the Illumina sequences (highlighted in purple) agreed on an MNR of ten A's, even though the Illumina pairs were misaligned with one another. The 'Illumina prioritization' strategy also indicated that the agreement of Illumina reads on an MNR of ten A's should override the agreement of the 454 sequencing reads on an MNR of nine A's, given the higher quality of Illumina sequencing technology with regard to MNRs. Through the application of these two methods, the consensus sequence was changed from a pad to an A at position 23,766.

High/Low Depth of Coverage

The contig was then checked for regions with low depth of coverage, which refers to sections of the assembly that have fewer than 40 reads. These regions need to be scanned for MNRs because they may not have the minimum number of discrepant reads necessary to be picked up by the 'highly discrepant positions' navigator. There were 49 regions with low depth of coverage that were listed by Consed, and these regions included 91 MNRs overall. Of the 91 MNRs shown (Appendix C), 88 of these regions effectively matched the consensus sequence or had already been edited using the 'highly discrepant positions' navigator. However, there were three MNRs in these regions with enough evidence to change the consensus sequence. For example, the 'base counting' strategy was used to determine that the MNR from position 4,626 to 4,633 had 27 of the 454 Sequencing reads in agreement with the consensus sequence. However, seven of the 454 Sequencing reads indicated changing the pad to a T at position 4,626, which was also supported by the two Illumina reads. The presence of these two Illumina reads was reason enough to warrant a change in the consensus sequence using the 'Illumina prioritization' strategy (Figure 6). However, this region falls at the very beginning of the contig and is preceded by a polyA run, so this alignment remains a hesitant prediction.

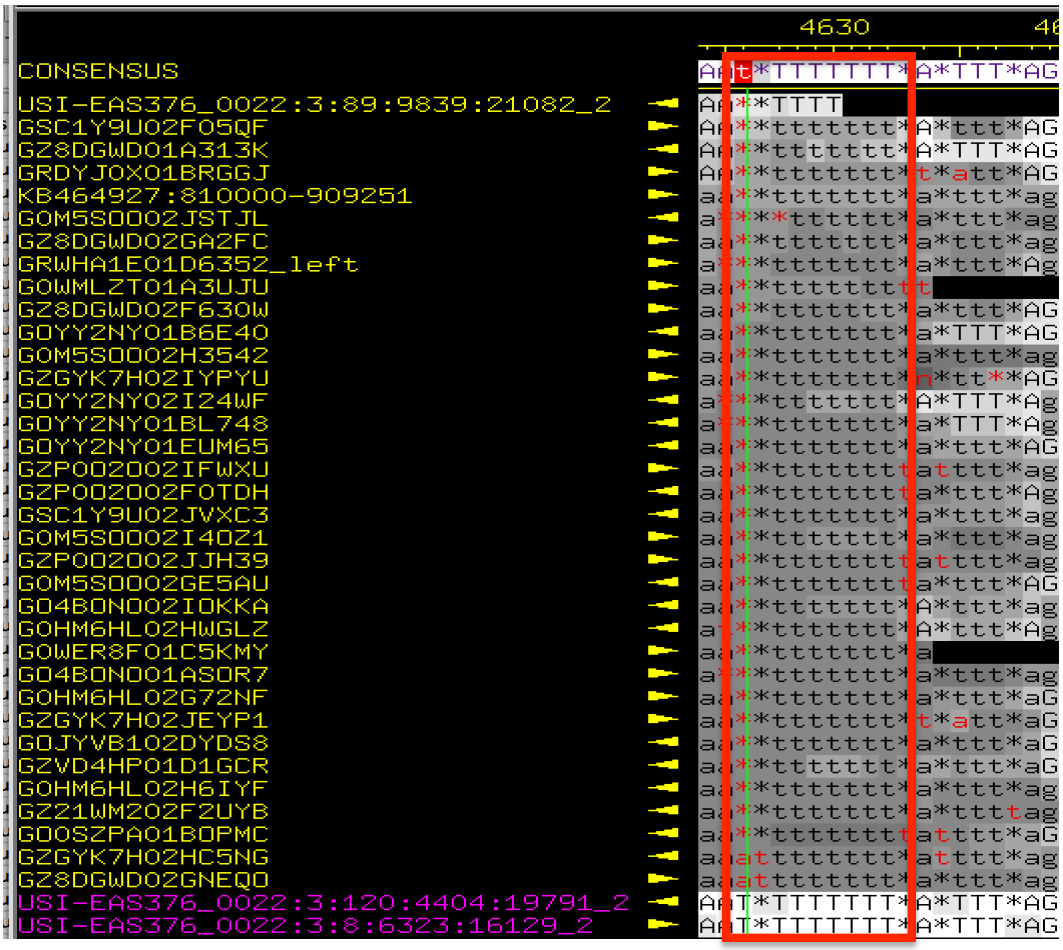

Figure 6: ***Example of an Edit Made in Region with Low Depth of Coverage (4,626-4,633)***

The low depth of coverage in this region became problematic due to the polyT run found between bases 4,626 and 4,633. Twenty seven of the 454 sequencing reads agreed with the sequence of 7 T's and 7 of the 454 sequencing reads indicated a run of 8 T's instead. There were 2 Illumina reads that also indicated a run of 8 T's. Although they were few in number, the Illumina reads provided enough evidence to support a change in the sequence.

On the other hand, an issue with high coverage was seen at the end of this contig within approximately the terminal 10 kb of the assembly. This region of high coverage corresponds to a sharp peak in the number of reads (the green line above the initial Assembly View) as seen in Figure 7. This sudden rise in the number of reads mapped to the region can be explained by the location of this contig at the end of the genomic scaffold that it was derived from. The contig approximately comprises the final 100 kb of this scaffold, so misassemblies in this final region are expected. Reads may have been inappropriately placed at the end of this contig, straddling

the end of the consensus sequence (Figure 8). To be placed at this location, paired reads need only match the base pairs on one side of the sequence, so a large number of reads that were not successfully mapped elsewhere tend to be misplaced here.

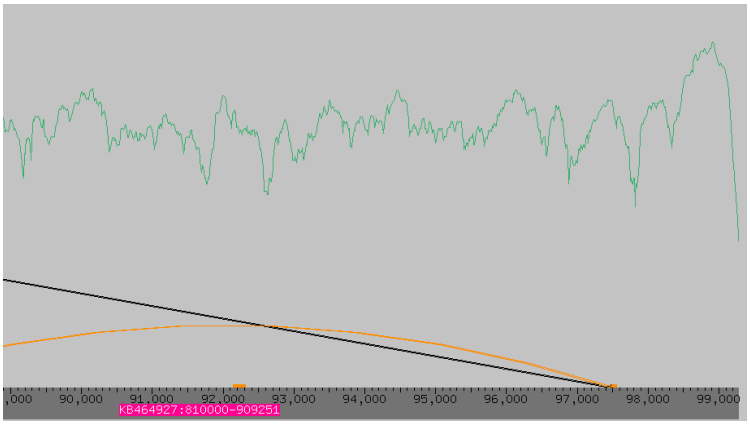

Figure 7: ***Increase in Depth of Coverage at Terminus of DEUG4927010 Contig***

The peak in depth of coverage, or the number of reads mapeed to the region, began at 98 kb and continued to the end of the contig. This may correspond to a misassembly where a large number of reads are incorrectly mapped to the end of the *D. eugracilis* dot chromosome. Figure 7 show the same data as Figure 2 albeit with a larger window.

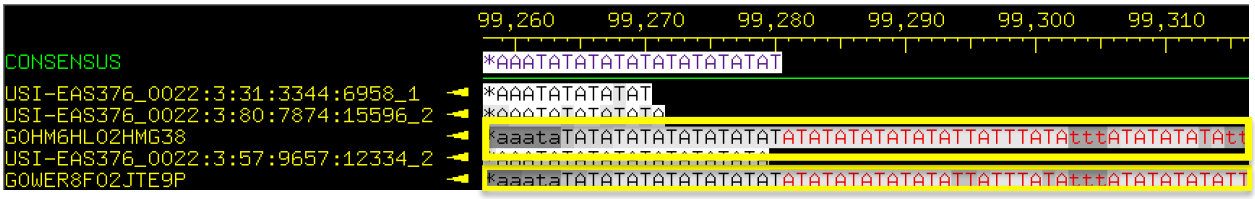

Figure 8: ***A View of Misaligned Reads Straddling the End of the Consensus Sequence***

The reads outlined in the yellow rectangles have sequences (written in red) that run past the consensus sequence. These are misplaced to the end of the contig because the first part of the sequence matches the consensus sequence. However, the consensus sequence ends before the terminus of the read, so the program accumulates indeterminable reads at this region, which ultimately results in misassembly with an increase in depth coverage.

Single Nucleotide Polymorphisms (SNPs)

The contig was checked for gaps using an 'N' search string, for vectors using an 'X' search string, and for low consensus quality regions (Phred <30) using the 'low consensus quality' navigator. The final checks for this contig confirmed that there were no gaps, vectors, or low consensus quality regions in the contig; however, two potential SNPs were found. The contig was then checked for SNPs. SNPs are sources of genetic variation between individuals of the same species based on differences in a mononucleotide, or a single base. SNPs in the assembly were identified while scanning the contig for 'highly discrepant positions' or 'low depth of coverage regions.' SNPs are defined as 40-60% the reads in agreement for one base while the other 60-40% the reads are in agreement for a different base. One potential polymorphism was found at position 98,587 where 59 reads agreed on an A at the site while 60 reads agreed on a C at the site (Figure 9). There was another potential SNP found relatively nearby at position 98,764. However, at position 98,764, 106 of the reads correspond to a G while 71 of the reads correspond to T, which is a 60:40 ratio for the second SNP. There was not complete correspondence between the two SNPs. Of the 63 reads that contained the SNPs on both sides, 85% had correspondence between the two sites. In other words, 85% of the reads had an A at 98,587 and a G at 98,764 or a C at 98,587 and a T at 98,764. However, 15% of the reads were not coordinated in this way and had an A at 98,587 but a T at 98764 or a C at 98,587 but a G at 98,764. Both SNPs at sites 98,587 and 98,764 were tagged as a 'polymorphism.'

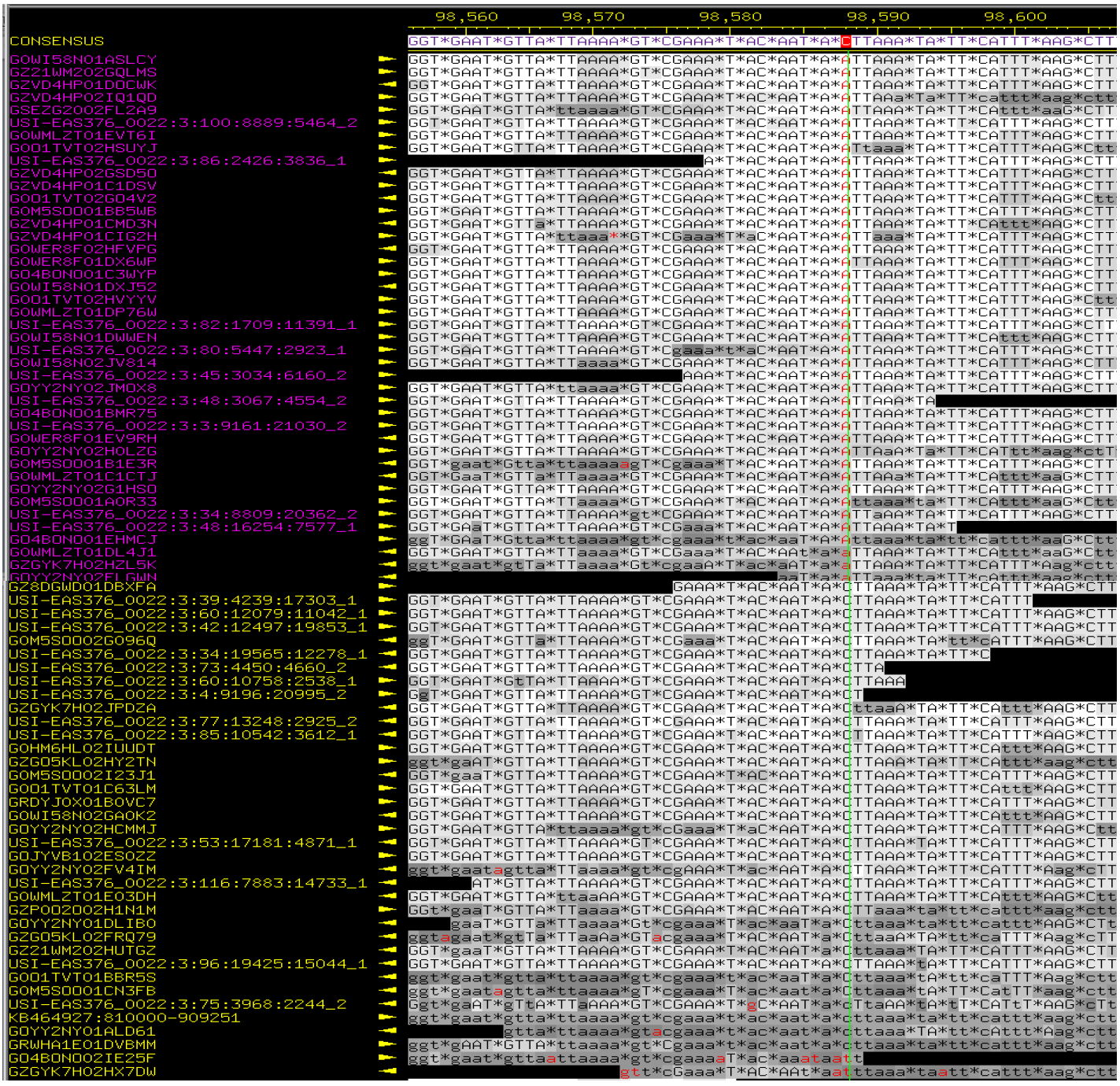

Figure 9: SNP at 98,587 with 50:50 Distribution across Reads for A or C

The purple highlighted reads correspond to reads with an A at the 98,587 site, and the yellow highlighted reads correspond to reads with a C at the 98,587 site. The placement of this position outside of any repeats, like transposable elements, as well as the even distribution across all reads for these two bases suggests that this is a SNP.

PCR Primers

One final position on this contig—site 99,103—located near the end of the contig within a repeat element, requires more analysis due to a high level of discrepancy between reads. The site potentially resembles another SNP because there are a large number of reads with the base G at that position and a large number of reads that have an A at that position. However, there are 35 reads with a G and 73 reads with an A at the 99,103 site, which roughly describes a 30:70 ratio instead of a 50:50 ratio. This ratio reduces the possibility of an SNP being associated with this site. The placement of this position within a repeat element also increases the likelihood that this trend results from variation of the repeat element. In order to determine the reason for these discrepancies, PCR primers were designed to obtain more data for this region of the contig.

PCR primers were selected around the position at 99,103 based on the PCR product size, interchangeability of the PCR primer pairs, and melting temperatures. Primers were designed so that the PCR product was not shorter than 600 bp (for more practical and efficient wet lab experiments) but no longer than 1,200 bp (because PCR products should be kept as small as possible). Four primers were chosen for the degree of interchangeability between the two pairs so that four combinations were possible in which each left hand primer could work with each right hand primer (Figure 10). These were also chosen to be unique sequences so that only the region around 99,103 would be amplified. Finally, the primers were chosen to have as small of a difference in melting points as possible to allow for efficient and simultaneous annealing during the PCR protocol. The PCR primers used for this project are shown in Figure 11. However, this involves a low priority check since the position in question falls within a repeat element.

pair #	distance between	primer1 contig	primer1 left	primer1 right	primer2 contig	primer2 left	primer2 right	melting p1	melting p2	primer1	primer2
1	120	KB464927:810000-909251	99031	99048	KB464927:810000-909251	99168	99185	55	56	cgttgatcgtttgttcgt	cggaaggtcaaattttcg
2	162	KB464927:810000-909251	99031	99048	KB464927:810000-909251	99210	99230	55	55	cgttgatcgtttgttcgt	tctccatgtatatgttgtccg
3	351	KB464927:810000-909251	98798	98817	KB464927:810000-909251	99168	99185	55	56	agacccacacatcgataaaa	cggaaggtcaaattttcg
4	393	KB464927:810000-909251	98798	98817	KB464927:810000-909251	99210	99230	55	55	agacccacacatcgataaaa	tctccatgtatatgttgtccg
5	547	KB464927:810000-909251	98600	98621	KB464927:810000-909251	99168	99185	55	56	ttaagctttctgtacttacccg	cggaaggtcaaattttcg
7	627	KB464927:810000-909251	98518	98541	KB464927:810000-909251	99168	99185	57	56	tgaatatatcaaaatcagctctgg	cggaaggtcaaattttcg
8	669	KB464927:810000-909251	98518	98541	KB464927:810000-909251	99210	99230	57	55	tgaatatatcaaaatcagctctgg	tctccatgtatatgttgtccg
9	1144	KB464927:810000-909251	98000	98024	KB464927:810000-909251	99168	99185	55	56	attttaacacaatattctcaagctc	cggaaggtcaaattttcg
10	1186	KB464927:810000-909251	98000	98024	KB464927:810000-909251	99210	99230	55	55	attttaacacaatattctcaagctc	tctccatgtatatgttgtccg
11	1315	KB464927:810000-909251	97834	97853	KB464927:810000-909251	99168	99185	55	56	cacttgaaagtcaacaaccc	cggaaggtcaaattttcg
12	1357	KB464927:810000-909251	97834	97853	KB464927:810000-909251	99210	99230	55	55	cacttgaaagtcaacaaccc	tctccatgtatatgttgtccg
13	1550	KB464927:810000-909251	97598	97618	KB464927:810000-909251	99168	99185	56	56	gacatgcatattctagagggg	cggaaggtcaaattttcg
14	1592	KB464927:810000-909251	97598	97618	KB464927:810000-909251	99210	99230	56	55	gacatgcatattctagagggg	tctccatgtatatgttgtccg

Figure 10: ***List of PCR Primers Possible for Checking Position 99,103***

The PCR primers used were determined by checking for a PCR product with a size between 600 and 1200 bases as well as melting point differences that were no more than 2 degrees apart between the primer pairs. Interchangeability between the left and right hand primers was also considered so that the 98,519 - 98,541 left hand primer could pair with both the 99,168 - 99,185 as well as the 99,210 – 99,230 right hand primers. Similarly, the 98,000 - 98,024 left hand primer was checked to be able to pair with the 99,168-99,185 and 99,210-99,230 right hand primers.

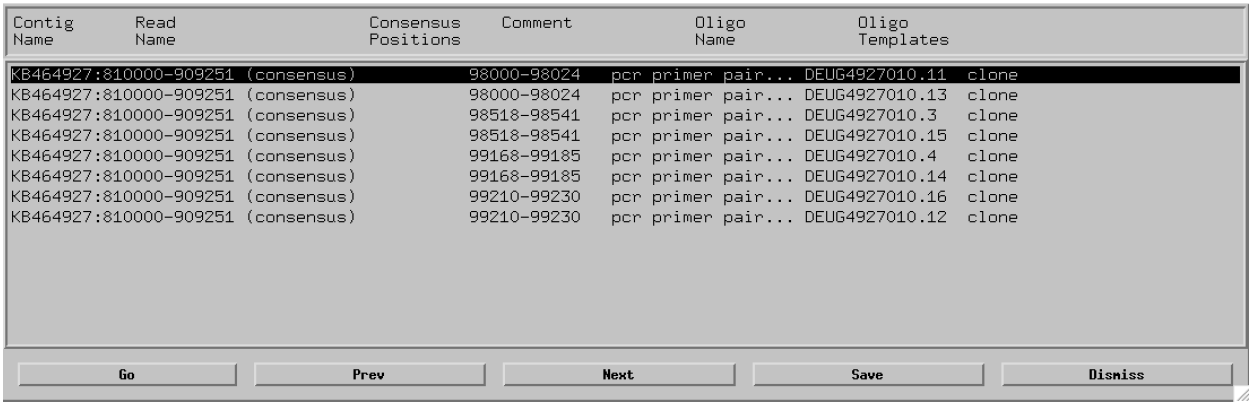

Figure 11: ***PCR Primers Chosen for the Experiment***

The PCR primers that were chosen for the experiment are shown above and are the same as pair #7, #8, #9, and #10 shown in Figure 10.

Conclusion

After finishing the DEUG4927010 contig, the final assembly (Figure 12) appears essentially the same as the initial assembly because there were neither tears performed during the finishing process nor any gaps that needed to be joined. No gaps, vectors, or low consensus quality regions were present in this project. However, all of the highly discrepant regions had been checked, which includes high quality discrepancies as well as low coverage regions. The SNPs at positions 98,587 and 98,764 had been tagged as polymorphisms. PCR primers were also

designed in order to better characterize the site at 99,103. Overall, the finishing step for DEUG4927010 has been completed and is ready for the annotation pipeline.

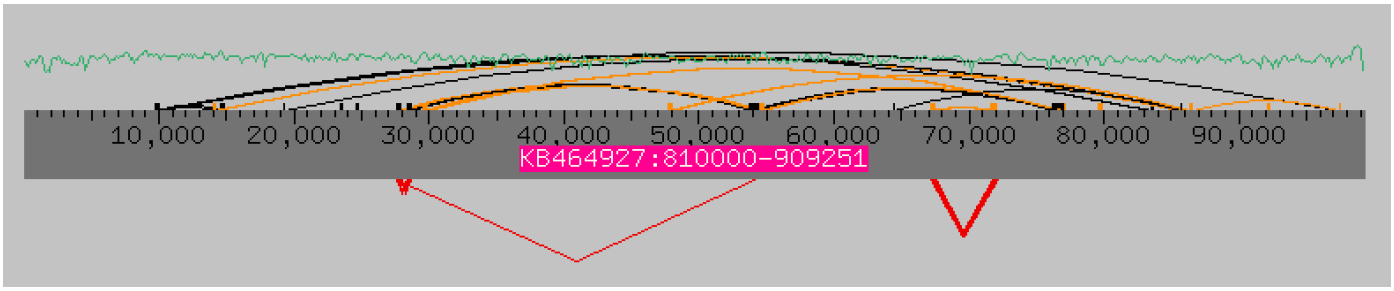

Figure 12: Final Assembly View

Acknowledgements:

I would like to sincerely thank Dr. Elgin, Dr. Shaffer, Wilson Leung, and Lee Trani for all of their help, expertise, and support with this project. I would also like to thank Dr. Bednarski for all of her help with the writing of this paper. Finally, I would like to extend my gratitude to Washington University in St. Louis as well as the Genomics Education Partnership for this opportunity.

Appendix A

	Changes to Consensus Sequence			
Contig	Position	Analysis	Action	Source of Data
DEUG4927010	3150-3156	monoT run	add T	Illumina
DEUG4927010	3617-3624	monoA run	add A	Illumina
DEUG4927010	4626-4633	monoT run	add T	Illumina
DEUG4927010	4897-4904	monoA run	add A	Illumina
DEUG4927010	6785-6793	monoA run	add A	Illumina
DEUG4927010	7877-7885	monoA run	add A	Illumina
DEUG4927010	12996-13004	monoA run	add A	Illumina
DEUG4927010	14903-14912	monoA run	add A	Illumina
DEUG4927010	17945-17952	monoT run	add T	Illumina
DEUG4927010	22551-22560	monoT run	add T	Illumina
DEUG4927010	23766-23775	monoA run	add A	Illumina
DEUG4927010	25841-25847	monoA run	add A	Illumina
DEUG4927010	29308-29316	monoA run	add A	Illumina
DEUG4927010	30998-31008	monoT run	add T	Illumina
DEUG4927010	33027-33033	monoT run	add T	Illumina
DEUG4927010	36307-36314	monoA run	add A	Illumina
DEUG4927010	41276-41283	monoA run	add A	Illumina
DEUG4927010	43597-43603	monoT run	add T	Illumina
DEUG4927010	44873-44881	monoT run	add T	Illumina
DEUG4927010	45137-45144	monoT run	add T	Illumina
DEUG4927010	48620-48628	monoA run	add A	Illumina
DEUG4927010	50014-50022	monoA run	add A	Illumina
DEUG4927010	56826-56834	monoT run	add T	Illumina
DEUG4927010	57808-57820	monoT run	add T	Illumina
DEUG4927010	69482-69490	monoT run	add T	Illumina
DEUG4927010	74987-74995	monoA run	add A	Illumina
DEUG4927010	75448-75455	monoT run	add T	Illumina
DEUG4927010	76740-76746	monoT run	add T	Illumina
DEUG4927010	87253-87261	monoT run	add T	Illumina

Appendix B:

	High Quality Discrepancies			
Contig	Position	Analysis	Action	Source of Data
DEUG4927010	677-683	monoA run	no change	Illumina
DEUG4927010	1176-1181	monoA run	no change	Illumina
DEUG4927010	2503-2509	monoA run	no change	Illumina
DEUG4927010	2525-2530	monoA run	no change	Illumina
DEUG4927010	3150-3156	monoT run	add T	Illumina
DEUG4927010	3191-3195	monoT run	no change	Illumina
DEUG4927010	3617-3624	monoA run	add A	Illumina
DEUG4927010	4558-4563	monoA run	no change	Illumina
DEUG4927010	4897-4904	monoA run	add A	Illumina
DEUG4927010	4998-5003	monoT run	no change	Illumina + 454
DEUG4927010	6617-6624	monoT run	no change	Illumina
DEUG4927010	6785-6793	monoA run	add A	Illumina
DEUG4927010	7585-7591	monoT run	no change	Illumina
DEUG4927010	7877-7885	monoA run	add A	Illumina
DEUG4927010	9365-9371	monoA run	no change	Illumina
DEUG4927010	12996-13004	monoA run	add A	Illumina
DEUG4927010	14526-14532	monoT run	no change	Illumina
DEUG4927010	14903-14912	monoA run	add A	Illumina
DEUG4927010	15654-15658	monoA run	no change	Illumina
DEUG4927010	15661-15665	monoA run	no change	Illumina
DEUG4927010	15913-15917	monoT run	no change	Illumina
DEUG4927010	17945-17952	monoT run	add T	Illumina
DEUG4927010	17968-17973	monoT run	no change	Illumina
DEUG4927010	18206-18211	monoT run	no change	Illumina
DEUG4927010	22180-22185	monoA run	no change	Illumina
DEUG4927010	22551-22560	monoT run	add T	Illumina
DEUG4927010	23765-23774	monoA run	add A	Illumina
DEUG4927010	25841-25847	monoA run	add A	Illumina
DEUG4927010	26284-26289	monoA run	no change	Illumina
DEUG4927010	28675-28771	monoA run	no change	Illumina
DEUG4927010	29308-29316	monoA run	add A	Illumina
DEUG4927010	30886-30892	monoT run	no change	Illumina
DEUG4927010	30998-31008	monoT run	add T	Illumina
DEUG4927010	33027-33033	monoT run	add T	Illumina
DEUG4927010	35514-35520	monoT run	no change	Illumina
DEUG4927010	36307-36314	monoA run	add A	Illumina

	High Quality Discrepancies			
Contig	Position	Analysis	Action	Source of Data
DEUG4927010	38845-38851	monoT run	no change	Illumina
DEUG4927010	39742-39746	monoA run	no change	Illumina
DEUG4927010	41276-41283	monoA run	add A	Illumina
DEUG4927010	42270-42276	monoA run	no change	Illumina
DEUG4927010	43597-43603	monoT run	add T	Illumina
DEUG4927010	44873-44881	monoT run	add T	Illumina
DEUG4927010	45137-45144	monoT run	add T	Illumina
DEUG4927010	48620-48628	monoA run	add A	Illumina
DEUG4927010	50014-50022	monoA run	add A	Illumina
DEUG4927010	50677-50681	monoA run	no change	Illumina
DEUG4927010	51169-51174	monoA run	no change	Illumina
DEUG4927010	56826-56834	monoT run	add T	Illumina
DEUG4927010	57808-57820	monoT run	add T	Illumina
DEUG4927010	58334-58338	monoA run	no change	Illumina + 454
DEUG4927010	58444-58450	monoT run	no change	Illumina
DEUG4927010	60523-60528	monoA run	no change	Illumina
DEUG4927010	60720-60727	monoT run	no change	Illumina
DEUG4927010	69482-69490	monoT run	add T	Illumina
DEUG4927010	74589-74593	monoA run	no change	Illumina + 454
DEUG4927010	74987-74995	monoA run	add A	Illumina
DEUG4927010	75448-75455	monoT run	add T	Illumina
DEUG4927010	76740-76746	monoT run	add T	Illumina
DEUG4927010	77469-77474	monoA run	no change	Illumina
DEUG4927010	77784-77788	monoT run	no change	Illumina
DEUG4927010	78940-78944	monoA run	no change	Illumina
DEUG4927010	80820-80825	monoA run	no change	Illumina
DEUG4927010	84539-84545	monoT run	no change	Illumina
DEUG4927010	87253-87261	monoT run	add T	Illumina
DEUG4927010	88918-88924	monoT run	no change	Illumina
DEUG4927010	89623-89629	monoA run	no change	Illumina

Appendix C:

	Low Depth of Coverage Regions			
Contig	Position	Analysis	Action	Source of Data
DEUG4927010	344-350	monoT run	no change	Illumina
DEUG4927010	360-364	monoT run	no change	Illumina + 454
DEUG4927010	394-401	monoT run	no change	Illumina
DEUG4927010	427-433	monoA run	no change	Illumina
DEUG4927010	514-518	monoT run	no change	Illumina + 454
DEUG4927010	526-530	monoA run	no change	Illumina
DEUG4927010	937-942	monoA run	no change	Illumina
DEUG4927010	970-974	monoA run	no change	Illumina
DEUG4927010	1563-1568	monoA run	no change	Illumina
DEUG4927010	1588-1595	monoA run	no change	Illumina
DEUG4927010	1803-1807	monoT run	no change	Illumina
DEUG4927010	1868-1872	monoT run	no change	Illumina
DEUG4927010	1912-1920	monoT run	no change	Illumina
DEUG4927010	2027-2031	monoT run	no change	Illumina
DEUG4927010	2073-2077	monoT run	no change	Illumina + 454
DEUG4927010	3506-3511	monoT run	no change	Illumina + 454
DEUG4927010	3533-3537	monoT run	no change	Illumina + 454
DEUG4927010	3687-3691	monoT run	no change	Illumina
DEUG4927010	3707-3711	monoT run	no change	Illumina
DEUG4927010	3716-3720	monoT run	no change	Illumina
DEUG4927010	4558-4563	monoA run	no change	Illumina
DEUG4927010	4626-4633	monoT run	add T	Illumina
DEUG4927010	4651-4655	monoT run	no change	Illumina
DEUG4927010	6436-6440	monoT run	no change	Illumina + 454
DEUG4927010	33143-33149	monoA run	no change	Illumina
DEUG4927010	33743-33747	monoA run	no change	Illumina
DEUG4927010	33749-33754	monoA run	no change	Illumina
DEUG4927010	33832-33839	monoT run	no change	Illumina
DEUG4927010	34991-34995	monoT run	no change	Illumina + 454
DEUG4927010	35163-35168	monoA run	no change	Illumina
DEUG4927010	35220-35224	monoT run	no change	Illumina + 454
DEUG4927010	36307-36314	monoA run	add A	Illumina
DEUG4927010	39141-39147	monoT run	no change	Illumina
DEUG4927010	39160-39168	monoT run	no change	Illumina
DEUG4927010	39189-39193	monoA run	no change	Illumina
DEUG4927010	39199-39203	monoT run	no change	Illumina
DEUG4927010	39234-39237	monoT run	no change	Illumina
DEUG4927010	39249-39257	monoA run	no change	Illumina
DEUG4927010	39359-39364	monoA run	no change	Illumina
DEUG4927010	39367-39376	monoT run	no change	Illumina
DEUG4927010	39405-39412	monoA run	no change	Illumina
DEUG4927010	40791-40797	monoT run	no change	Illumina
DEUG4927010	40942-40947	monoT run	no change	Illumina + 454
DEUG4927010	41151-41155	monoT run	no change	Illumina
DEUG4927010	41171-41175	monoT run	no change	Illumina + 454

	Low Depth of Coverage Regions			
Contig	Position	Analysis	Action	Source of Data
DEUG4927010	41193-41197	monoT run	no change	Illumina
DEUG4927010	41276-41283	monoA run	add A	Illumina
DEUG4927010	41421-41425	monoT run	no change	Illumina
DEUG4927010	42719-42725	monoA run	no change	Illumina
DEUG4927010	42756-42760	monoT run	no change	Illumina + 454
DEUG4927010	42771-42776	monoT run	no change	Illumina
DEUG4927010	44632-44651	monoT run	no change	Illumina
DEUG4927010	44721-44725	monoT run	no change	Illumina
DEUG4927010	45591-45595	monoA run	no change	Illumina
DEUG4927010	45597-45602	monoA run	no change	Illumina
DEUG4927010	45616-45621	monoT run	no change	Illumina
DEUG4927010	52013-52017	monoA run	no change	Illumina
DEUG4927010	52089-52093	monoA run	no change	Illumina + 454
DEUG4927010	52097-52101	monoT run	no change	Illumina
DEUG4927010	53678-53684	monoA run	no change	Illumina
DEUG4927010	56611-56616	monoA run	no change	Illumina
DEUG4927010	57221-57225	monoA run	no change	Illumina
DEUG4927010	57251-57256	monoT run	no change	Illumina
DEUG4927010	65164-65172	monoT run	no change	Illumina
DEUG4927010	70227-70231	monoA run	no change	Illumina
DEUG4927010	70257-70261	monoT run	no change	Illumina
DEUG4927010	70277-70282	monoT run	no change	Illumina
DEUG4927010	70331-70335	monoA run	no change	Illumina
DEUG4927010	70380-70387	monoA run	no change	Illumina
DEUG4927010	70394-70401	monoA run	no change	Illumina
DEUG4927010	73143-73148	monoT run	no change	Illumina
DEUG4927010	76015-76022	monoT run	no change	Illumina
DEUG4927010	76054-76058	monoA run	no change	Illumina + 454
DEUG4927010	76080-76084	monoA run	no change	Illumina
DEUG4927010	77912-77916	monoT run	no change	Illumina
DEUG4927010	78012-78018	monoA run	no change	Illumina
DEUG4927010	79896-79900	monoA run	no change	Illumina
DEUG4927010	80905-80909	monoT run	no change	Illumina
DEUG4927010	85762-85766	monoT run	no change	Illumina
DEUG4927010	92581-92589	monoA run	no change	Illumina
DEUG4927010	92633-92641	monoA run	no change	Illumina
DEUG4927010	93000-93005	monoA run	no change	Illumina
DEUG4927010	93127-93131	monoA run	no change	Illumina + 454
DEUG4927010	93143-93149	monoA run	no change	Illumina
DEUG4927010	96880-96887	monoA run	no change	Illumina
DEUG4927010	96889-96898	monoA run	no change	Illumina
DEUG4927010	96916-96200	monoT run	no change	Illumina
DEUG4927010	97754-97760	monoA run	no change	Illumina
DEUG4927010	97813-97820	monoT run	no change	Illumina
DEUG4927010	97824-97829	monoT run	no change	Illumina
DEUG4927010	99233-99237	monoT run	no change	Illumina