Daniel Cui Zhou
Bio 434w
Dr. Elgin
February 27, 2015

Finishing of DELE8596006

**Abstract**

DELE 8596006 is the first 100 kb genomic region of the Muller F element of *Drosophila elegans*. The difficulties finishing this assembly included three gaps, three low coverage regions that need Sanger sequencing, and multiple regions that contained different copies of a helitron transposon that could not be resolved. This region has no vector sequences and all mononucleotide runs have been checked. The project is not ideal for annotation yet, since the gaps, low coverage regions, and discrepant transposon copies still need to be resolved through PCR and Sanger sequencing.

**Introduction**

In eukaryotes, chromatin can be largely classified as euchromatin or heterochromatin. Euchromatin is packaged more loosely than heterochromatin and is associated with active transcription, as the DNA is more accessible to the transcription machinery. Heterochromatin, on the other hand, is more tightly packaged and is associated with silencing. Centromeric and telomeric regions of chromosomes, as well as a large number of repetitious sequences and pseudogenes that should not be expressed, are packaged as heterochromatin. For example, packing up transposable elements into heterochromatin can prevent them from amplifying and spreading throughout the genome. The Muller F element in *Drosophila*, also known as the dot chromosome, is mostly composed of constitutive heterochromatin, but many essential genes (such as *eyeless*, a gene essential for proper eye formation) are still actively transcribed in a euchromatic fashion. Very little is known about the mechanisms behind this unusual chromosome, such as how gene regulation and expression happens, and therefore, this project seeks to identify regulatory motifs in the F elements of *Drosophila*

*biarmipes* and *Drosophila elegans*. These data will allow comparative analysis with other *Drosophila* species and thus reveal more information on the unusual nature of the F element, contributing to our understanding of the relationship between chromatin structure and gene expression.

**Initial Assembly**

The initial assembly of DELE8596006 showed a single contig with 20 discrepant forward-reverse pairs and four regions with low coverage (Figure 1). As a first step, I analyzed the discrepant forward-reverse pairs. Forward-reverse pairs, or mate pairs, are pairs of reads with a known distance between each other and whose orientation must face towards each other. All pairs except one mapped on both ends to repetitious sequences, which is expected, due to the high number of repeats. They were not removed, though, since it could still be possible to map them using high quality discrepancies. The remaining pair that mapped around 12 kb and 75 kb was pulled out, since it mapped a unique sequence to a repetitious sequence and thus could likely be mapped to the right place. The pulled out read did not align to any part of the contig, however, suggesting that it does not belong in the assembly (Figure 2).

## Project Details for DELE8596006

### Assembly View

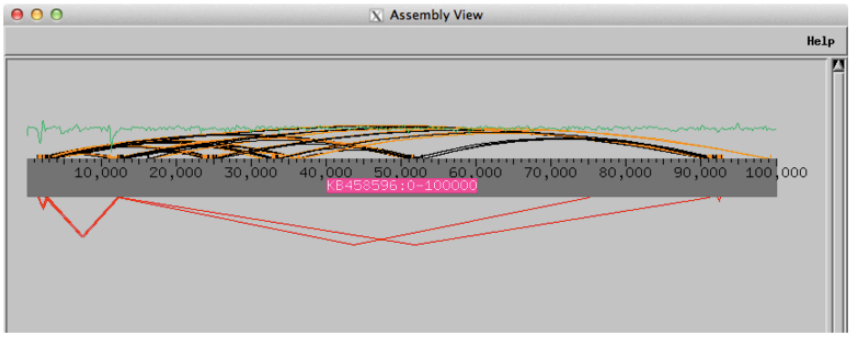


Figure 1. Initial assembly of DELE8596006. Green lines represent high quality coverage. Red lines represent discrepant forward-reverse pairs. Orange and black lines represent repetitious sequences.

### Project Status

| | |
|---|---|
| Number of low consensus quality regions: | 13 |
| Number of highly discrepant regions (at least 3 discrepant reads with quality >= 30): | 385 |
| Number of gaps: | 3 |
| Estimated total gap size: | 41 |

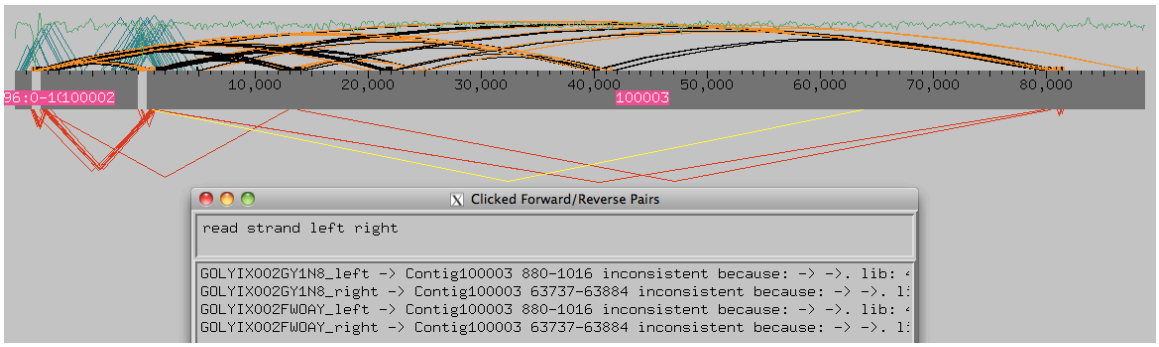


**Figure 2. Assembly View showing gaps. Paired reads spanning the gaps are shown as dark green lines. Orange and black lines show locations of repetitious sequences. The forward-reverse pair that maps a repetitious sequence to a unique sequence is marked with a yellow line. Other forward-reverse pairs are shown in dark orange.**

## Resolving gaps

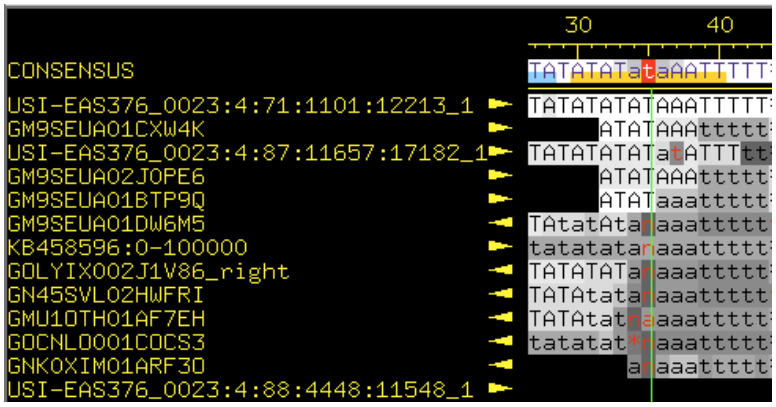


**Figure 3. 1 bp gap at site 35. The gap was filled with a T, but more data are needed to ensure that the gap should be filled by a T.**

Three gaps are present in the contig. The first gap was only 1 bp long, located at site 35 near the beginning of the contig. Since this contig holds the first 100 kb of the F element, this gap needed to be addressed due to the lack of overlap from other projects, unlike the 2.5 kb at the rightmost end of the contig. Based on five high quality reads, the gap was filled in with a T. However, the reads are highly dubious since the reads are close to their ends, so the region was marked with a "data-needed" tag (Figure 3). This gap could also have been caused by *Consed* incorrectly calculating the consensus due to the small number of reads aligned to the region. Since an upstream primer is unable to be picked, PCR primers were not designed to cover this region. In order to re-sequence this region, a primer must be designed on the scaffold originally attached to

this end of the contig, but the necessity to do so will be determined as the project moves on to annotation.

The second gap was located between sites 1908 – 1927. Comparing the sequences upstream and downstream of the gap using a search for string revealed multiple matches throughout the contig. This was expected, since the gap is flanked by repetitious sequences. The sequences on each side did not align with each other, however, so resolving the gap was not possible and PCR primers were needed (Figures 4 and 5).

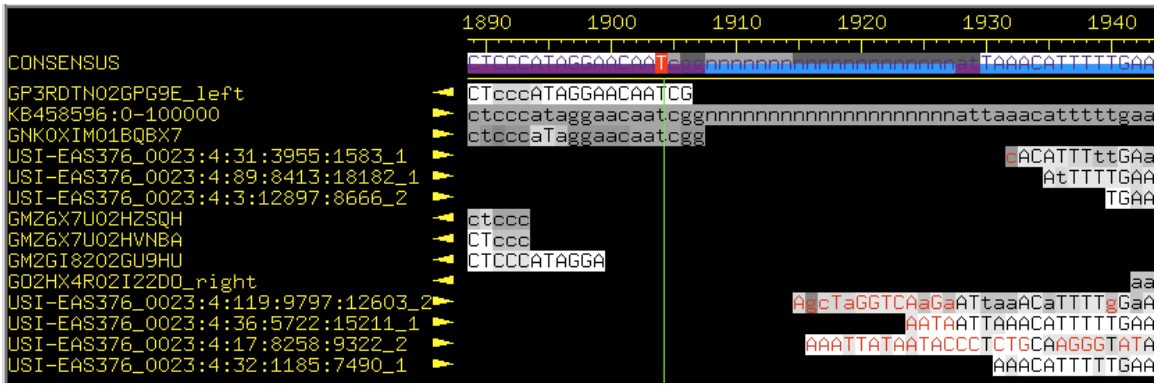


**Figure 4. 20 bp gap at 1908-1927 located between two repeats. PCR primers were designed to cover this region. This gap is spanned by paired reads (Figure 2).**

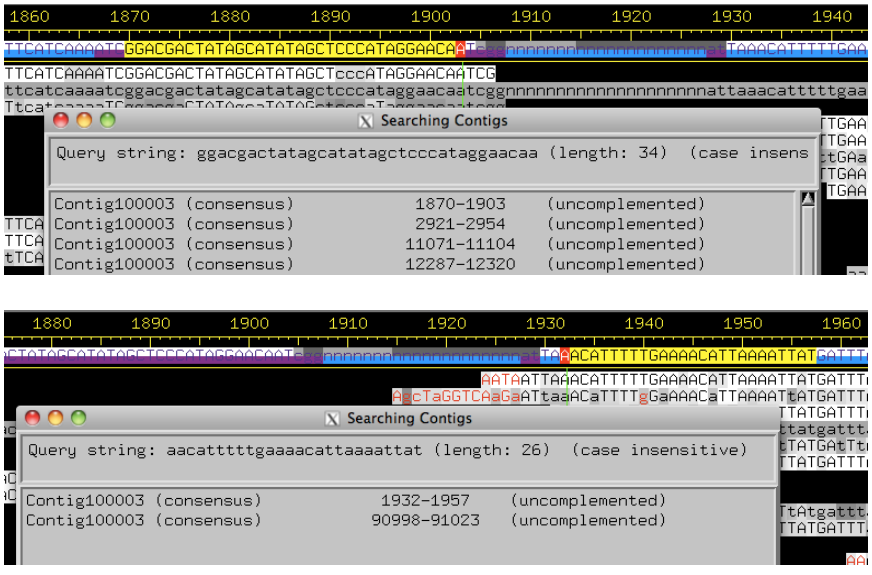


**Figure 5. Search for string upstream and downstream of the first gap. No matches were found on either side of the gap.**

The default primer picking preferences yielded PCR products with a minimum length of about 2 kb. In an attempt to reduce the size of the PCR product, several changes to the primer picking preferences were attempted. These included increasing the maximum primer size to 40, increasing the maximum melting temperature to 70 °C, increasing the maximum "match elsewhere" score value to 23, and increasing the maximum length of mononucleotide repeats to 6. However, the minimum PCR product was only reduced by 3 bp after these changes. Therefore, using the default primer picking parameters, 4 primers were picked whose combinations were approved by *Consed* (Primers 5-8, Table 1). To pick these primers, I first determined two primer pairs that shared forward and reverse oligos by eye (to be able to combine them), and then ensured that the product length was as close to 700-1000 bp as possible, and that their melting temperature did not differ by more than 1 degree Celsius.

The third gap was very similar to the second gap, aside from the lack of a repetitious sequence downstream. Searching for sequence alignments before and after the gap yielded no results, and therefore PCR primers were needed (Figures 6 and 7). Using a similar method as described above, 4 primers were picked that were approved by *Consed* (Primers 1-4, Table 1).

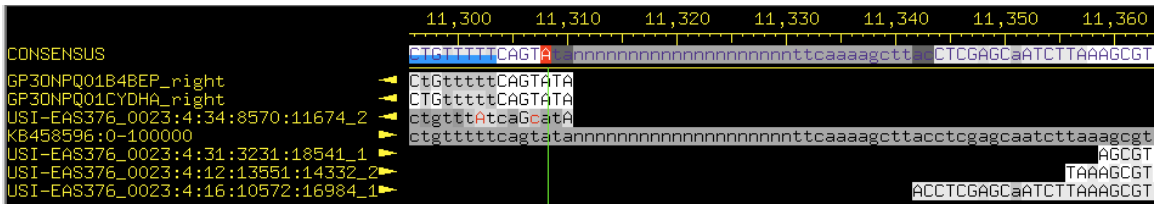


**Figure 6. 20 bp gap spanning 11305-11324. PCR primers were designed to cover this region. This gap is spanned by paired reads (Figure 2).**

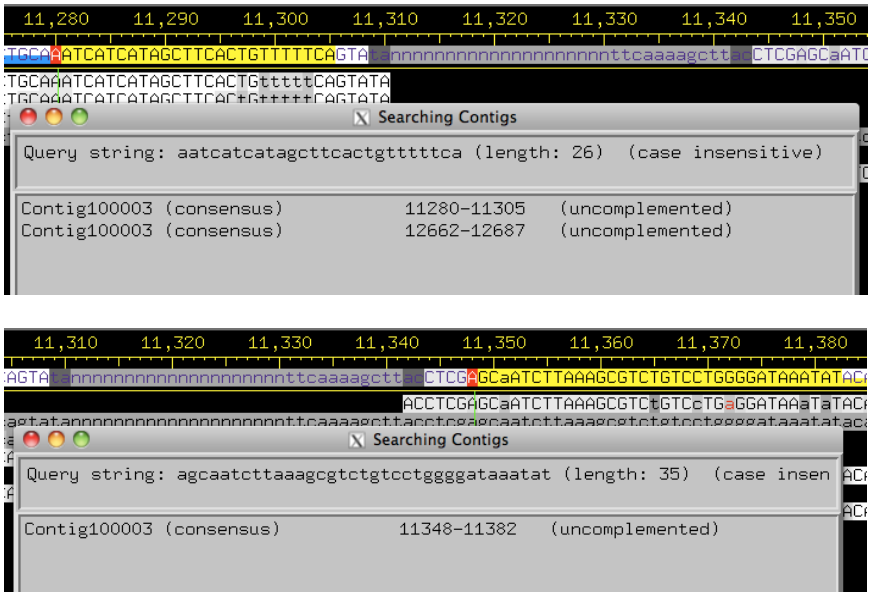


**Figure 7. Search for string upstream and downstream of the second gap. No matches were found on either side of the gap.**

| Region addressed | Primer name | Primer location | Directionality | Primer sequence |
|---|---|---|---|---|
| Gap 1: 1908-1927 Low Coverage: 2100-2550 | DELE8596006.5 | 1089-1114 | F | tttctgcttctatatttgtactaacg |
| | DELE8596006.7 | 1305-1329 | F | agaatatactgcatgcattagaaat |
| | DELE8596006.8 | 3326-3347 | R | cctcaaccaagaaagtgtgtaa |
| | DELE8596006.6 | 3527-3551 | R | aaacatttcaattatttaatgcttc |
| Gap 2: 11305-11324 | DELE8596006.1 | 10522-10540 | F | agtaagcaatcccagatcg |
| | DELE8596006.2 | 11585-11603 | R | tgggaatcaccattctgtt |
| | DELE8596006.3 | 10745-10769 | F | aattcaaattacacttgcttaagtt |
| | DELE8596006.4 | 11381-11400 | R | tcgatgagtctggcttgtat |
| Low Coverage: 11452-11468 | DELE8596006.9 | 10522-10540 | F | agtaagcaatcccagatcg |
| | DELE8596006.10 | 11771-11796 | R | caaaagtaatgttttcaattaaatgt |
| | DELE8596006.11 | 10661-10686 | F | tttaaatattcgacacatgttatttt |
| | DELE8596006.12 | 11526-11543 | R | aaaagaccgctccacaac |

**Table 1. List of primers ordered by site. In order to minimize product size to ensure successful Sanger sequencing, two different pairs of primers were designed to cover Gap 2 and the low coverage region around 11452. However, the product of primers 9 and 10 can cover both problem regions fully if the sequencing signal is long enough.**

### Mononucleotide runs

454 pyrosequencing is prone to making mistakes on mononucleotide runs due to the nature of the sequencing chemistry. However, using high quality Illumina sequencing reads, most mononucleotide runs were resolved. This contig contained relatively few mononucleotide runs of Cs and Gs, but a very large number of A and T mononucleotide runs. This fits the fact that the Muller F element is AT rich. There were 1099 regions that

needed to be checked, but only 34 mononucleotide runs (about 3%) required a change in the consensus. The vast majority of the problem areas were lacking at least one base in the mononucleotide run, as determined by counting the mononucleotide in high quality Illumina reads, and were resolved easily by manually replacing pads (*) with the required number of bases in the consensus as needed. For example, the monoA run at around 24,850 originally had 7 As in the consensus (due to the 454 reads), but manually counting the As in the Illumina reads, the actual number of As is 8 (Figure 8).

Additionally, multiple discrepancies were often caused by long mononucleotide runs that led to misalignments. This resulted in whole reads being offset by a certain number of bases. For example, in the monoA run at around 92,425, the reads look misaligned, but careful inspection of the surrounding bases and counting of the number of As (14 As) revealed that the reads do belong in that region. Thus, the consensus was easily fixed by adding 4 As, even if the reads stayed misaligned (Figure 9).

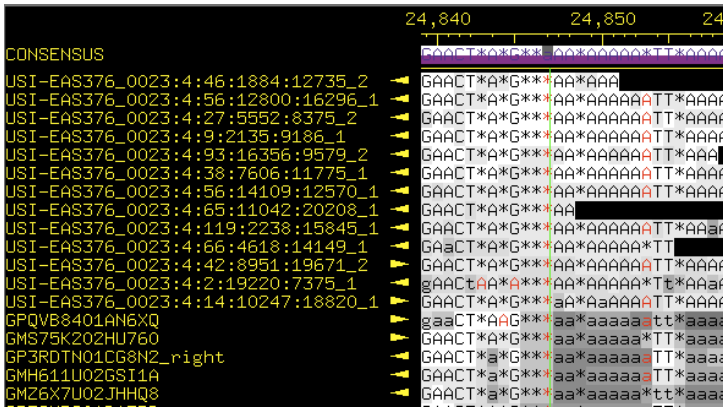


Figure 8. Region resolved by adding an A to the consensus.

Overall, there were 34 mononucleotide runs that required changes to the consensus. Seventeen of them required addition of one ore more Ts and seventeen required the addition of one or more As. No C or G runs required any change. The changes seem to be spread out evenly (17 changes in the first half of the contig and 17 changes in the second half). There also doesn't seem to be a bias for the change to occur in either a mono-A or mono-T run.

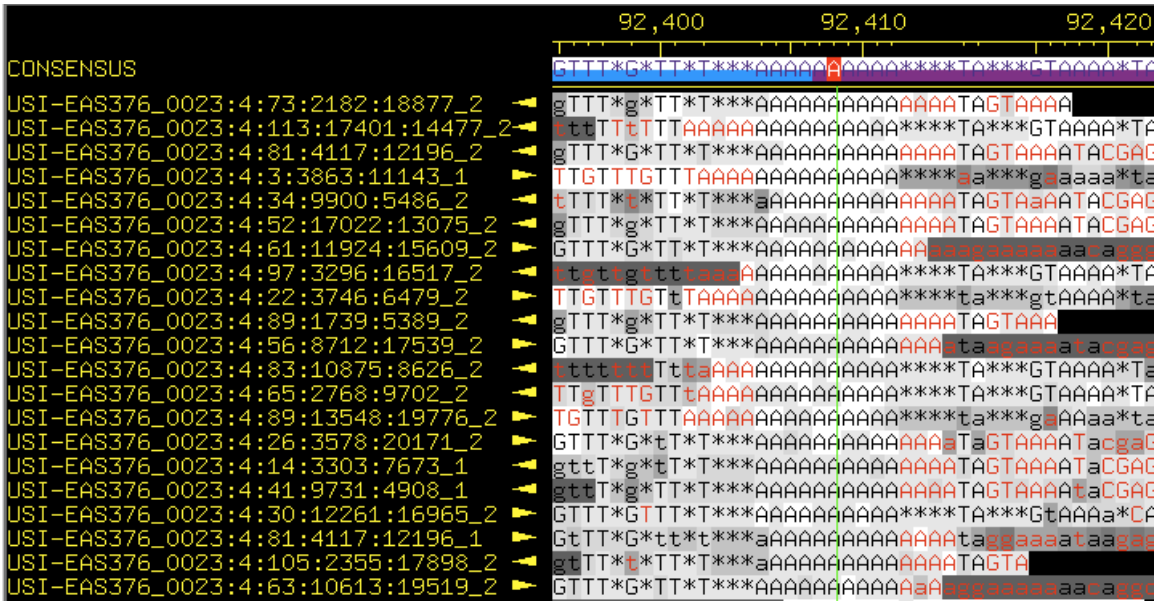


**Figure 9. Misaligned reads offset due to the Poly-A run.**

### High quality discrepancies

The high quality discrepancies throughout the entire contig could generally be categorized into two groups: discrepancies due to multiple copies of a transposon and discrepancies due to *Consed* mistakes. There were 385 high quality discrepancies, but only about 50 sites were relevant, since the rest of the discrepancies were caused by misaligned reads or manual editing of bases in mononucleotide runs. Of the relevant sites, 9 regions (each spanning about 50 – 200 bp) included multiple discrepancies caused by two different copies of helitron transposons, and thus remained unresolved (Figure 10). Unfortunately, in all of these regions, the reads where the discrepancies occurred were not long enough to contain any unique sequences past the transposon in order to ascertain which copy truly belongs to the region. Furthermore, the reads that contain both part of the repeat and some unique sequence downstream and upstream of the problem region don't have discrepancies, and thus are not helpful in determining which transposon copy is correct. I also removed the reads that represent one copy of the transposon and did a mini-assembly in order to try to realign them. Using a search

for string, I was unable to find a region where the reads align. I repeated this procedure with the other set of reads that represent the other copy of the transposon, but got similar results. This suggests that the consensus is likely a combination of both transposon copies that needs to be adderssed. These regions were therefore marked as "unresolved", since more data (such as longer reads) are needed to determine which copy of the transposon is correct.


**Figure 10. Different copies of a transposon. All the high quality Illumina reads disagree on multiple sites, but many reads contain the same discrepant bases.**

Three high quality discrepant regions necessitated changes in the consensus. Two of the three sites were *Consed* errors, where the consensus contradicts a large number of high quality reads which agree with each other (Figure 11). In both of these regions, however, *Consed* labeled the 5 bp regions (in gray) as "unknown repeats." Additionally, on both occasions, a G base was called within the same helitron repeat towards the end of the 100 kb region. The consensus error thus could have been caused by the occurrence of the unknown repeat within the helitron repeat, which led to *Consed* miscalling the bases.

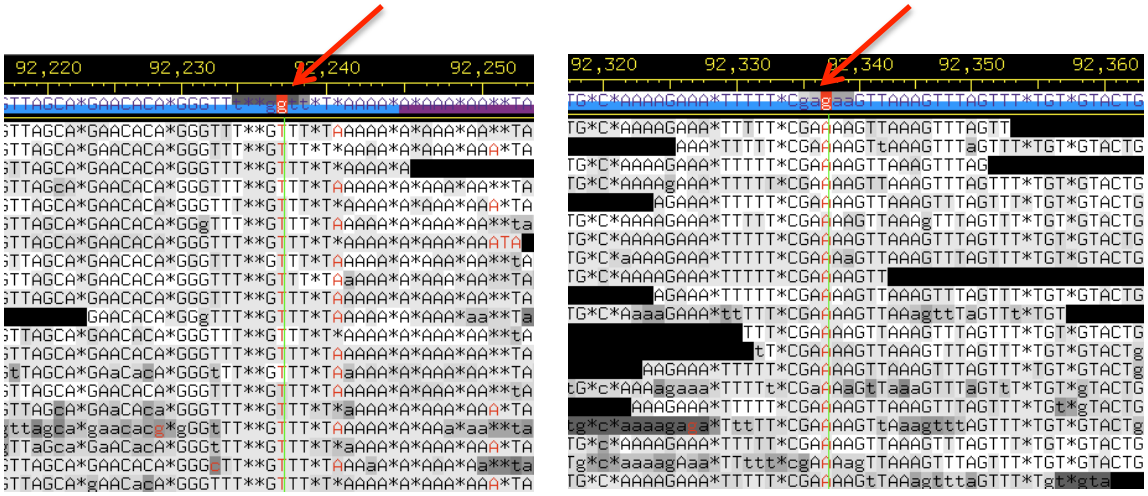

**Figure 11. Wrong base calls by *Consed* labeled as "unknown repeat." The red arrows point at the repeat.**

*Consed* was very consistent when dealing with misalignments, but the last discrepant region that required a consensus change (at site 94,866) was caused by a significant number of misaligned reads. What originally looked like two wrong bases in the consensus was actually one wrong base. However, by looking upstream of the discrepant sites, the reads that had a G at the first discrepant site also had other discrepancies upstream (for example, a discrepant G at site 94,823, among others). These reads were therefore identified as misaligned (Figure 12). This region was thus resolved by changing a T into an A in the consensus consistent with the high quality reads that showed no other discrepancies.

**Figure 12. Discrepancy caused by a large number of misaligned reads. The highlighted T was changed into an A. The highlighted reads in purple are the misaligned reads. The red arrow points at the site.**

**Regions with low depth of coverage (fewer than 40 reads)**

There were 12 large regions with mononucleotide runs that had fewer than 40 reads, and thus had to be carefully examined to determine if the reads were of high enough quality. If there were at least 2 reads with a quality of at least 20 for each base and no discrepant base calls within these reads, the consensus was considered to be correct (Figure 13).

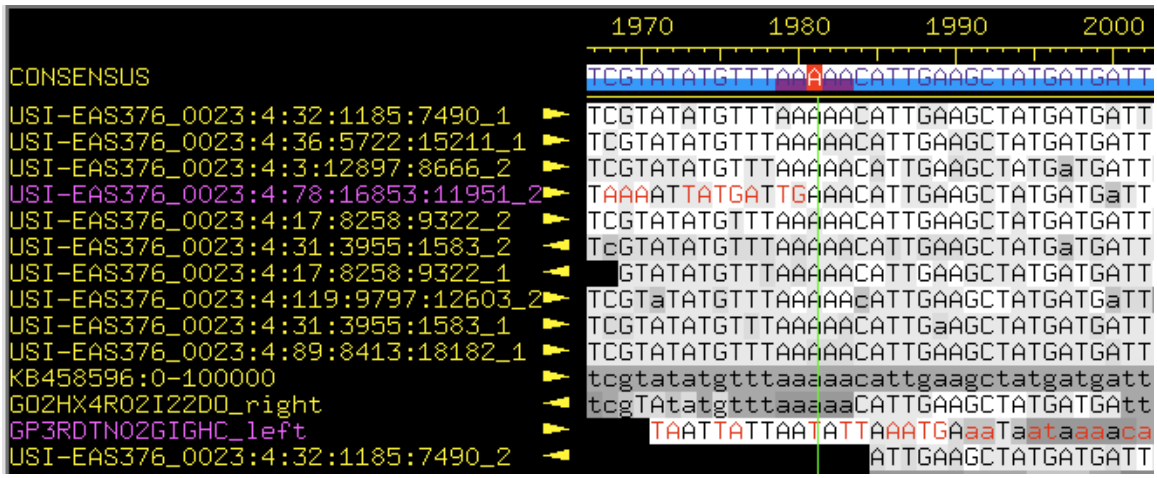


**Figure 13. Low coverage region with enough support for the consensus. The highlighted reads are misaligned.**

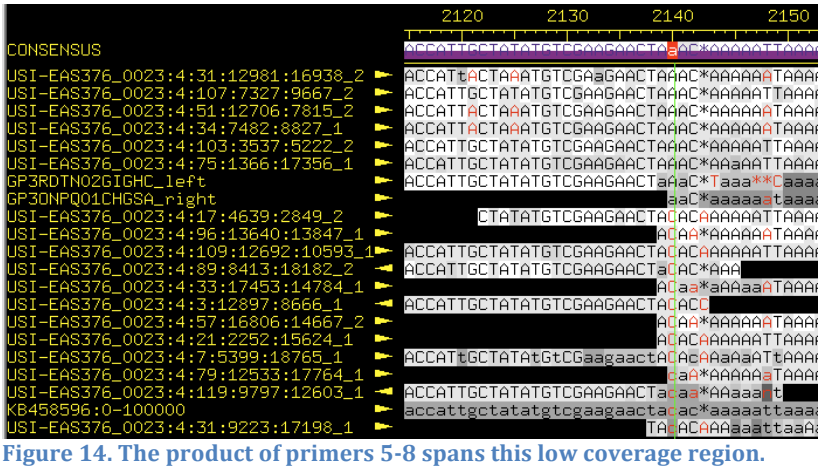


**Figure 14. The product of primers 5-8 spans this low coverage region.**

Nine of the twelve regions have enough good reads to support the consensus. The remaining three regions, however, did not have enough reads and had multiple discrepancies (Figures 14 and 15). The product of previous primer pairs ordered for a nearby gap already covered the first two regions (both within 2100 – 2500), so there was no need to design new primers.

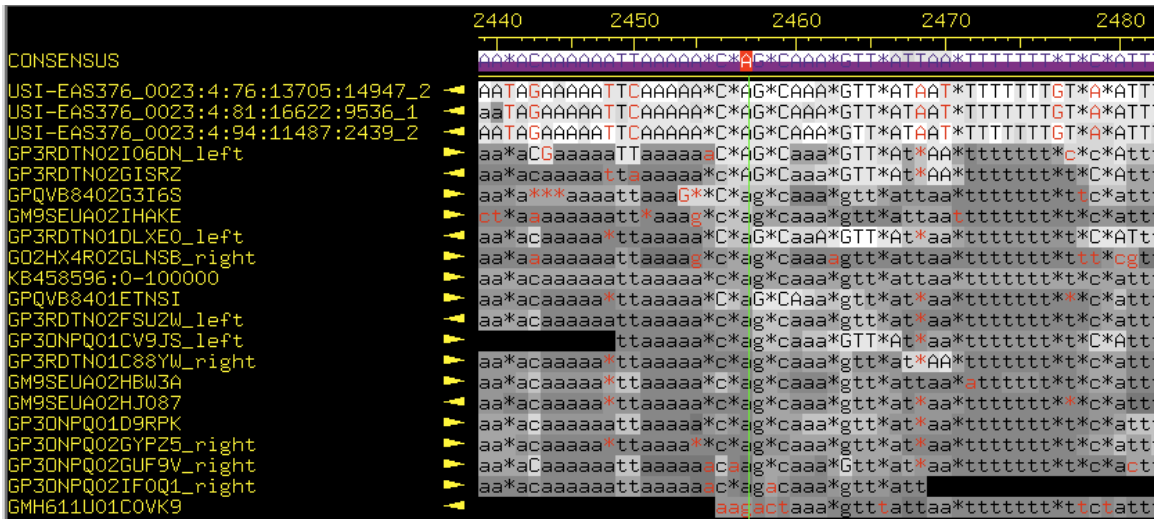


**Figure 15. The product of primers 5-8 spans this low coverage region.**


**Figure 16. Low coverage region that needs PCR.**

The third region had very low coverage and a discrepant number of As between the Illumina reads (Figure 16). Since 2 out of the 6 high quality Illumina reads disagreed on the number of bases, 4 primers were picked with the same method used for the other primers (Primers 9-12, Table 1). Primer 9 is actually the same as primer 1. Thus, the oligo tag for primer 9 was commented as “ignore this primer,” and primer 1 was used when ordering the oligos for these pairs. This problem region lies just downstream from the primer pairs designed to cover the nearby gap. These primers were thus designed to cover this region and minimize the product size to maximize sequencing success.

**Conclusions**

Region DELE8596006 is not completely finished. There are 3 gaps and 3 low coverage areas that need further sequencing in order to determine the consensus sequence. Furthermore, many repeat regions contain multiple discrepant sites due to the presence of different copies of a helitron transposon. In order to determine which copy of the transposon is correct, further sequencing is needed (ideally with longer reads), and annotation will help determine if sequencing should be done.

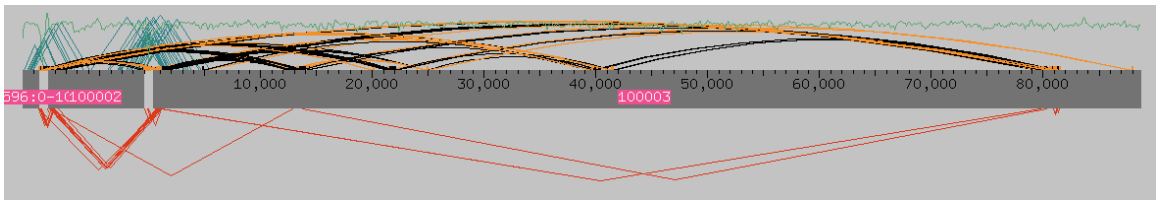


**Figure 17. Current assembly of DELE8596006.**

The final assembly of this project remained largely unchanged, but all mononucleotide runs outside of regions that need PCR are resolved (Figure 17). All remaining discrepant regions fall within the transposon copies that could not be resolved. Twelve PCR primers (primers 1 and 9 are the same) were designed in order to cover the gaps and low coverage regions, totaling 24 sequencing reactions (Figure 18). No vector sequences were present in the project. Table 2 shows a summary of the manual corrections in the consensus as well as the PCR primer pairs ordered. Additionally, DELE8596006 contains the first 100 kb of the genome, and thus there are no overlapping sequences upstream from another project for the first 2.5 kb. The beginning of the contig therefore has very low coverage and PCR primers cannot be designed to span this section of the contig. Thus, DELE8596006 is not completely finished and is not ready for annotation.

| Target Sequence | Forward Primer | Forward Primer Sequence | Reverse Primer | Reverse Primer Sequence | Reaction Name |
|---|---|---|---|---|---|
| Gap 1: 1908-1927 Low Coverage: 2100-2550 | DELE8596006.5 | tttctgcttctatatttgtactaacg | DELE8596006.6 | aaacatttcaattatttaatgcttc | selgin14DELE8596006PCR5g6_5.b1 |
| | DELE8596006.5 | tttctgcttctatatttgtactaacg | DELE8596006.6 | aaacatttcaattatttaatgcttc | selgin14DELE8596006PCR5g6_6.b1 |
| | DELE8596006.5 | tttctgcttctatatttgtactaacg | DELE8596006.8 | cctcaaccaagaaagtgtgtaa | selgin14DELE8596006PCR5g8_5.b1 |
| | DELE8596006.5 | tttctgcttctatatttgtactaacg | DELE8596006.8 | cctcaaccaagaaagtgtgtaa | selgin14DELE8596006PCR5g8_8.b1 |
| | DELE8596006.7 | agaatatactgcatgcattagaaat | DELE8596006.6 | aaacatttcaattatttaatgcttc | selgin14DELE8596006PCR7g6_6.b1 |
| | DELE8596006.7 | agaatatactgcatgcattagaaat | DELE8596006.6 | aaacatttcaattatttaatgcttc | selgin14DELE8596006PCR7g6_7.b1 |
| | DELE8596006.7 | agaatatactgcatgcattagaaat | DELE8596006.8 | cctcaaccaagaaagtgtgtaa | selgin14DELE8596006PCR7g8_7.b1 |
| | DELE8596006.7 | agaatatactgcatgcattagaaat | DELE8596006.8 | cctcaaccaagaaagtgtgtaa | selgin14DELE8596006PCR7g8_8.b1 |
| Gap 2: 11305-11324 | DELE8596006.1 | agtaagcaatcccagatcg | DELE8596006.2 | tgggaatcaccattctgtt | selgin14DELE8596006PCR1g2_1.b1 |
| | DELE8596006.1 | agtaagcaatcccagatcg | DELE8596006.2 | tgggaatcaccattctgtt | selgin14DELE8596006PCR1g2_2.b1 |
| | DELE8596006.1 | agtaagcaatcccagatcg | DELE8596006.4 | tcgatgagtctggcttgtat | selgin14DELE8596006PCR1g4_1.b1 |
| | DELE8596006.1 | agtaagcaatcccagatcg | DELE8596006.4 | tcgatgagtctggcttgtat | selgin14DELE8596006PCR1g4_4.b1 |
| | DELE8596006.3 | aattcaaattacacttgcttaagtt | DELE8596006.2 | tgggaatcaccattctgtt | selgin14DELE8596006PCR3g2_2.b1 |
| | DELE8596006.3 | aattcaaattacacttgcttaagtt | DELE8596006.2 | tgggaatcaccattctgtt | selgin14DELE8596006PCR3g2_3.b1 |
| | DELE8596006.3 | aattcaaattacacttgcttaagtt | DELE8596006.4 | tcgatgagtctggcttgtat | selgin14DELE8596006PCR3g4_3.b1 |
| | DELE8596006.3 | aattcaaattacacttgcttaagtt | DELE8596006.4 | tcgatgagtctggcttgtat | selgin14DELE8596006PCR3g4_4.b1 |
| Low Coverage: 11452-11468 | DELE8596006.1 | agtaagcaatcccagatcg | DELE8596006.10 | caaaagtaatgttttcaattaaatgt | selgin14DELE8596006PCR1g10_1.b1 |
| | DELE8596006.1 | agtaagcaatcccagatcg | DELE8596006.10 | caaaagtaatgttttcaattaaatgt | selgin14DELE8596006PCR1g10_10.b1 |
| | DELE8596006.1 | agtaagcaatcccagatcg | DELE8596006.12 | aaaagaccgctccacaac | selgin14DELE8596006PCR1g12_1.b1 |
| | DELE8596006.1 | agtaagcaatcccagatcg | DELE8596006.12 | aaaagaccgctccacaac | selgin14DELE8596006PCR1g12_12.b1 |
| | DELE8596006.11 | tttaaatattcgacacatgttatttt | DELE8596006.10 | caaaagtaatgttttcaattaaatgt | selgin14DELE8596006PCR11g10_10.b1 |
| | DELE8596006.11 | tttaaatattcgacacatgttatttt | DELE8596006.10 | caaaagtaatgttttcaattaaatgt | selgin14DELE8596006PCR11g10_11.b1 |
| | DELE8596006.11 | tttaaatattcgacacatgttatttt | DELE8596006.12 | aaaagaccgctccacaac | selgin14DELE8596006PCR11g12_11.b1 |
| | DELE8596006.11 | tttaaatattcgacacatgttatttt | DELE8596006.12 | aaaagaccgctccacaac | selgin14DELE8596006PCR11g12_12.b1 |

**Figure 18. List of sequencing reactions. All sequencing reactions will use Big Dye chemistry**

| Mononucleotide runs | | | |
|---|---|---|---|
| Contig | Location | Change | Source of data |
| Contig100003 | 7120-7133 | +T | Illumina reads |
| Contig100003 | 9449-9456 | +A | Illumina reads |
| Contig100003 | 9943-9950 | +T | Illumina reads |
| Contig100003 | 10007-10020 | +TTT | Illumina reads |
| Contig100003 | 15969-15982 | +AA | Illumina reads |
| Contig100003 | 16085-16095 | +A | Illumina reads |
| Contig100003 | 17333-17341 | +A | Illumina reads |
| Contig100003 | 19275-19285 | +T | Illumina reads |
| Contig100003 | 23263-23272 | +A | Illumina reads |
| Contig100003 | 23993-24001 | +T | Illumina reads |
| Contig100003 | 24846-24853 | +A | Illumina reads |
| Contig100003 | 24856-24861 | +A | Illumina reads |
| Contig100003 | 26528-26536 | +A | Illumina reads |
| Contig100003 | 31899-31908 | +T | Illumina reads |
| Contig100003 | 32207-32216 | +T | Illumina reads |
| Contig100003 | 32252-32260 | +A | Illumina reads |
| Contig100003 | 32979-32984 | +A | Illumina reads |
| Contig100003 | 45938-45945 | +T | Illumina reads |
| Contig100003 | 51270-51278 | +T | Illumina reads |
| Contig100003 | 60544-60552 | +T | Illumina reads |
| Contig100003 | 62344-62351 | +T | Illumina reads |
| Contig100003 | 63734-63742 | +T | Illumina reads |
| Contig100003 | 64162-64171 | +T | Illumina reads |
| Contig100003 | 72284-72294 | +T | Illumina reads |
| Contig100003 | 73812-73818 | +T | Illumina reads |
| Contig100003 | 77749-77760 | +AA | Illumina reads |
| Contig100003 | 78110-78116 | +A | Illumina reads |
| Contig100003 | 80800-80809 | +T | Illumina reads |
| Contig100003 | 81872-81881 | +A | Illumina reads |
| Contig100003 | 91933-91938 | +A | Illumina reads |
| Contig100003 | 91945-91951 | +T | Illumina reads |
| Contig100003 | 92278-92288 | +A | Illumina reads |
| Contig100003 | 92441-92454 | +AAAA | Illumina reads |
| Contig100003 | 92663-92671 | +A | Illumina reads |

| Discrepancies | | | |
|---|---|---|---|
| Contig | Location | Change | Source of data |
| Contig100003 | 92272-92276 | G -> T | Illumina reads |
| Contig100003 | 92373-92376 | G -> A | Illumina reads |
| Contig100003 | 94906-94910 | T ->A | Illumina reads |

| Recommended Sanger Sequencing (in Contig 100003) | | | |
|---|---|---|---|
| Primer pair | Target Sequence | Region Spanned | Reason |
| 5 and 6 | 1908-1927 2100-2550 | 1089-3551 | Unresolved gap and 2 low coverage regions |
| 5 and 8 | | 1089-3347 | |
| 7 and 6 | | 1305-3551 | |
| 7 and 8 | | 1305-3347 | |
| 1 and 2 | 11305-11324 | 10522-11603 | Unresolved gap |
| 1 and 4 | | 10522-11400 | |
| 3 and 2 | | 10745-11603 | |
| 3 and 4 | | 10745-11400 | |
| 1(9) and 10 | 11452-11468 | 10522-11796 | Low coverage region |
| 1(9) and 12 | | 10522-11543 | |
| 11 and 10 | | 10661-11796 | |
| 11 and 12 | | 10661-11543 | |

**Table 2. Summary of consensus changes and PCR primers.**

**Acknowledgements**

I would like to thank the Washington University in St. Louis Bio 434W faculty, Drs. Sarah Elgin, Christopher Shaffer, and Wilson Leung, for their continued guidance and support during this finishing project. I would also like to thank the professional finishers from the Genome Institute, Jennifer Hodges and Lee Trani, for their technical support. This project was funded by the Howard Hughes Medical Institute and Washington University in St. Louis.