Harley Greene

Bio434W

Elgin

Finishing of DEUG4927002

Abstract

The entire genome of *Drosophila eugracilis* has recently been sequenced using Roche 454 pyrosequencing and Illumina paired-end reads sequencing. In this project, the contig DEUG4927002, a 100 kb genomic region of the *D. eugracilis*' dot chromosome, was finished by confirming and correcting the consensus sequence. The original assembly of the region contained 1 gap, 138 highly discrepant regions, 35 low coverage regions, and 1 region of low consensus quality. These regions were inspected for mononucleotide runs (MNRs), which often misinform the consensus sequence due to 454 pyrosequencing errors and require correcting. Here, 31 MNRs were either corrected by sequencing read inspection or confirmed. PCR primers were created to cover the unresolved gap; Sanger sequencing data will be required for this purpose. No polymorphisms were identified in DEUG4927002. Besides the one gap, this genomic region is ready for annotating.

Introduction

DNA in eukaryotic cells is found in one of two formations: euchromatin and heterochromatin. Euchromatin encompasses transcriptionally active genes which are loosely packaged in the nucleus to allow for easy access of transcription factors and RNA polymerase. Heterochromatin is typically associated with transcriptionally inactive genes and these domains are relatively compact. Heterochromatin is most often found near the centromeres and telomeres of chromosomes. Heterochromatin can be modified to allow for transcription, but it is much less accessible than euchromatin. Additionally, heterochromatin and euchromatin have unique histone and DNA modifications. For example, chromatin regions that are transcriptionally active and are associated with euchromatin often have the histone acetylation on H3K9. These marks provide a clear way to differentiate heterochromatin and euchromatin and provide insight into gene regulation mechanisms.

The species *Drosophila eugracilis* has a small dot chromosome (the F element) containing approximately 80 actively transcribed genes, but by most measures is entirely heterochromatic. The dot chromosome illustrates an unusual scenario of heterochromatic genes being expressed at the same level as euchromatic genes from other chromosomes. *D. eugracilis* has recently been sequenced but still requires finishing and annotating. Finishing and annotating *D. eugracilis* will allow for genomic comparisons with *Drosophila melanogaster*, and other evolutionarily close neighbor species, to study the mechanisms of heterochromatic gene expression.

The initial *D. eugracilis* sequenced genome was constructed using two types of sequencing data: Roche 454 pyrosequencing and Illumina paired end reads. 454 pyrosequencing, which produces long reads (~450 nts) compared to Illumina, has been known to drop off in sequencing quality at mononucleotide runs (MNRs). On the other hand, Illumina sequencing reads are more precise, but are shorter in comparison (100-150 nts). Therefore, the short Illumina reads can be used to correct errors in incorrect MNRs reported in 454 reads.

In this report, the 100 kb contig DEUG4927002 was finished by analyzing its consensus sequence and making changes to the consensus when necessary. All analysis and sequence changes were conducted on the program *Consed*.

Initial Assembly

Contig DEUG4927002 maps to bases 90,000-190,000 of the *D. eugracilis* dot chromosome (Figure 1). The initial Assembly View (Figure 2) showed a single contig of 100 kb. The contig contained one gap, 138 highly discrepant regions, 35 low coverage regions, and 1 region of low consensus quality. Additionally, there were several regions of repetitious sequences and several incorrectly matched forward/reverse read pairs. The incorrectly matched read pairs were likely marked incorrect due to the highly repetitive nature of this contig's sequence.

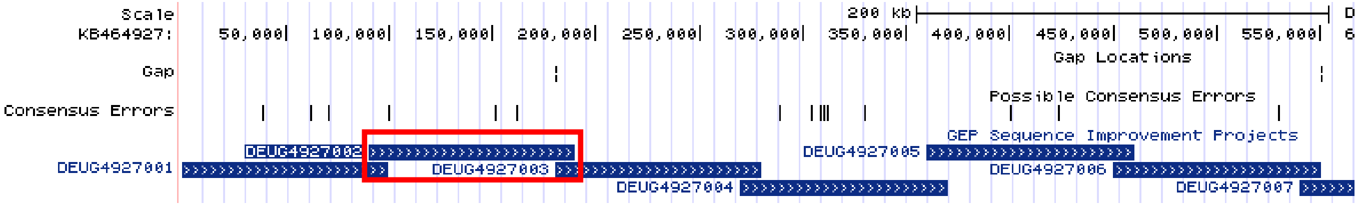


Figure 1: Position of DEUG4927002 in *D. eugracilis* dot chromosome: contig finished in this report, DEUG4927002, is highlighted by red box. It represents region 90,000 – 190,000 in the *D. eugracilis* dot chromosome.

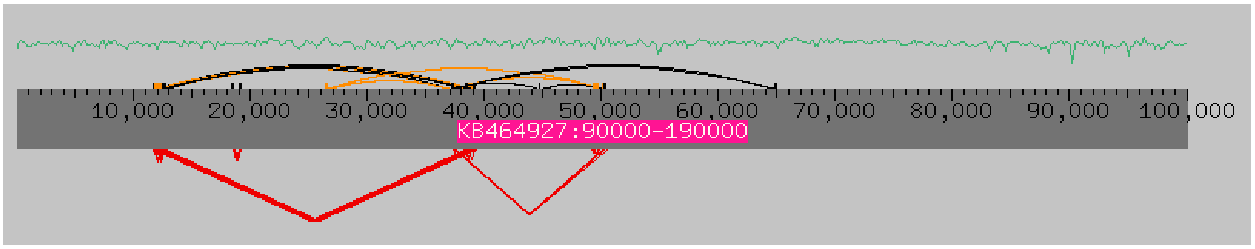


Figure 2: Initial Assembly View of DEUG4927002: 100,000 bp region visualized using assembly view on *Consed*. Green line represents depth of reads across the contig. Red lines indicate incorrect spacing or orientation between forward and reverse read pairs. Black and orange lines indicate repetitious sequences that map to multiple locations.

High Quality Discrepancies

The contig was first examined for high quality discrepancies. High quality discrepancies were identified as regions where at least three reads did not match the consensus sequence, ignoring bases with a Phred quality score below 30. In total, 138 highly discrepant regions were identified using *Consed*. These high quality discrepancies were then examined for MNRs. Of the

138 highly discrepant regions, 73 were MNRs. For each MNR, the 454 and Illumina reads were inspected to ensure the consensus was correct. At each location, the number of bases in the MNR consensus were counted and compared to the high quality Illumina reads. The consensus was confirmed if the Illumina reads matched the consensus. The consensus required editing if the reads did not match. Thirty-one MNRs required editing to correct the consensus sequence.

All MNRs identified were all mono-A or mono-T runs, which is unsurprising since heterochromatin is known to be A/T rich. Almost all of the changed bases were due to slightly misaligned reads which shortened the MNR in the consensus and were corrected by adding a single base to the beginning or end of the MNR. The addition of an A at position 42,613 illustrates this typical problem (Figure 3). The consensus sequence showed eight As with a pad (*) inserted after the fifth A. The first three Illumina reads have a ninth A at the end of the MNR,

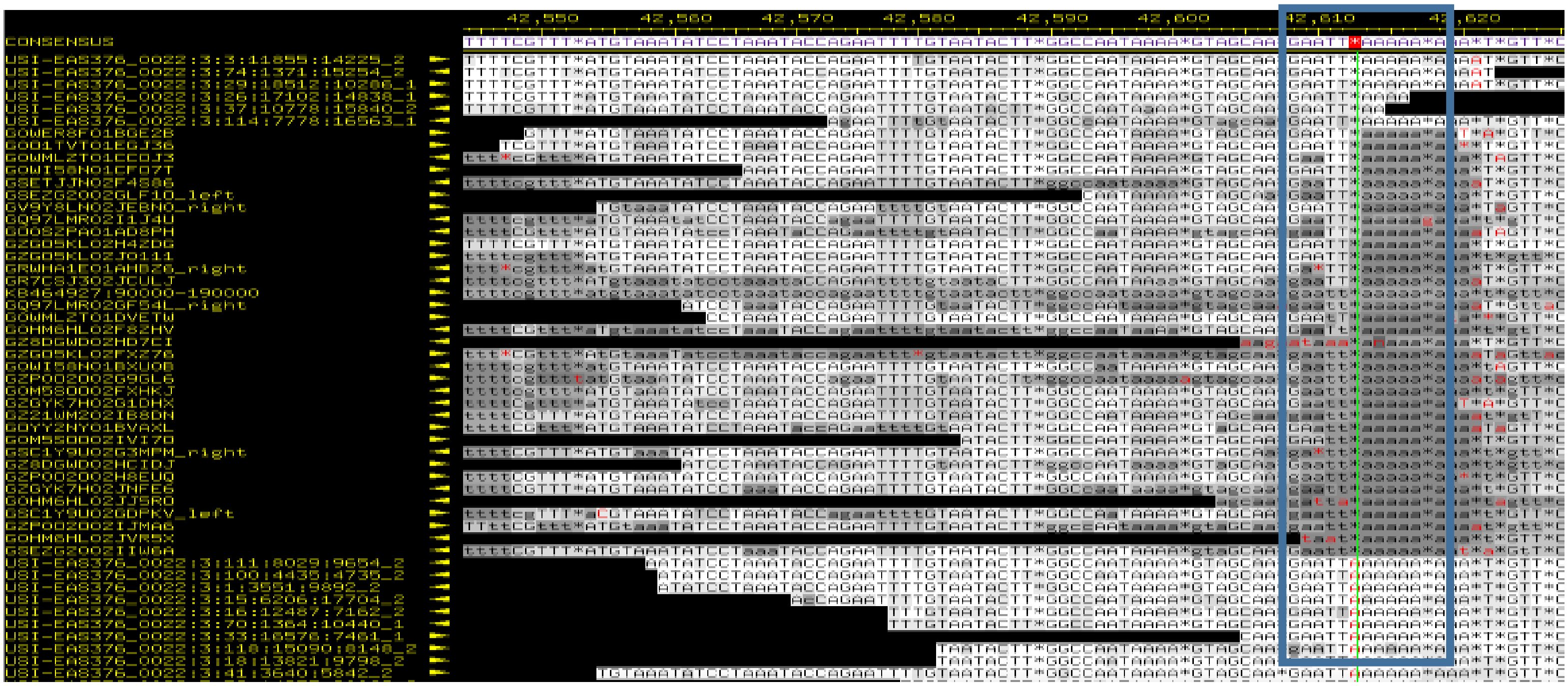


Figure 3: Common MNR correction: This is a common example of a misaligned MNR that was resolved by adding a single base. Here, a MNR of eight As was seen in the consensus but a MNR of nine As was seen in the Illumina reads (region within blue box). A single A was added at position 42,613 to resolve the area. Read names that start with "USI" refer to Illumina data and read names that start with a "G" refer to 454 data.

which is not represented in the consensus. Below the Illumina reads are 37 low quality 454 reads which provide little help for this situation. Below the 454 reads are 19 more Illumina reads that have a ninth A at the beginning of the mono-A run. Since all of the Illumina reads indicated that the MNR should have nine As instead of eight As, an A was inserted into the consensus.

Besides these easily reconcilable errors, there were a few more interesting and difficult cases. One interesting case was a highly discrepant position at 38,334 (Figure 4). This region represented an overlap between two repeat elements present on the dot chromosome. The consensus showed six Ts, with the first and last position of the MNR marked as highly discrepant. The first 14 Illumina reads only had five Ts with a pad inserted in the first position of the MNR. Additionally, there were three Illumina reads, and many mid- and low-quality 454 reads (indicated by grey boxes around letters), that also had five Ts, but with a pad inserted in the last position of the MNR. Due to the misalignment of the reads, an extra T had been added to the

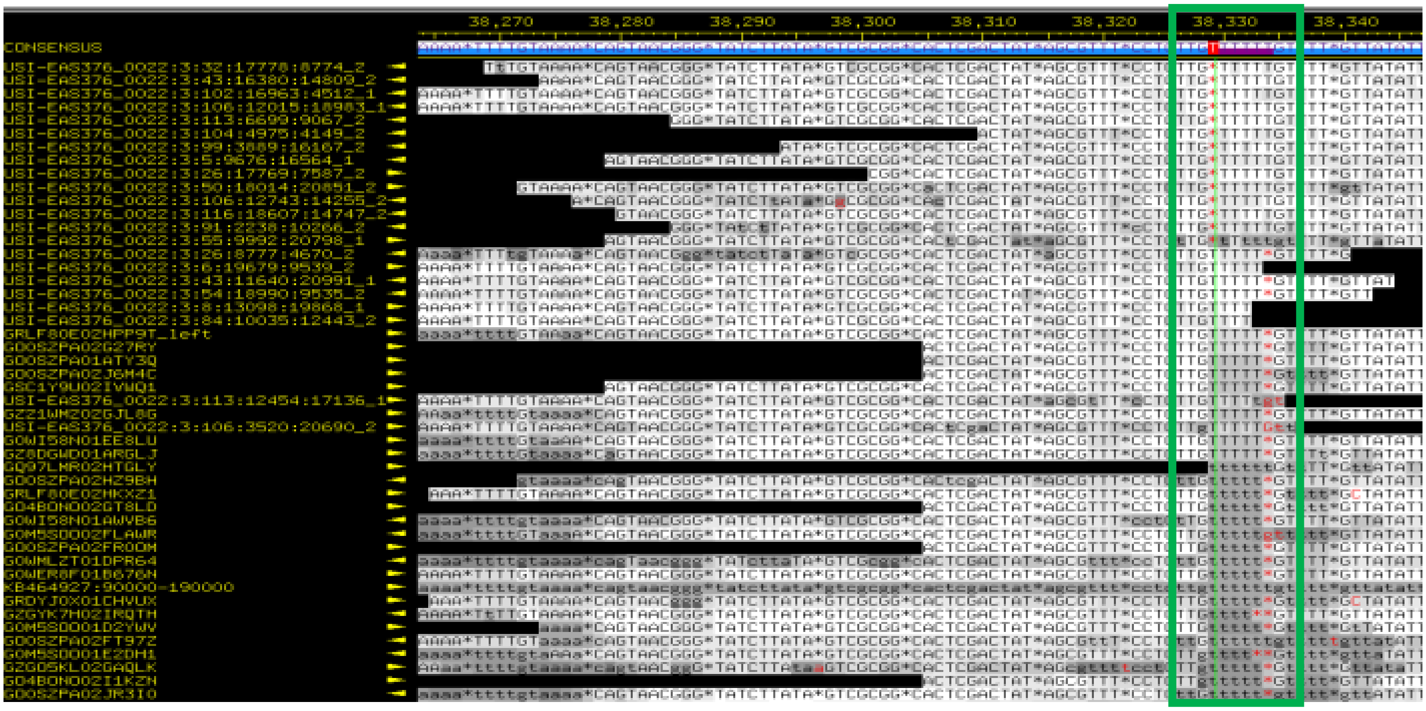
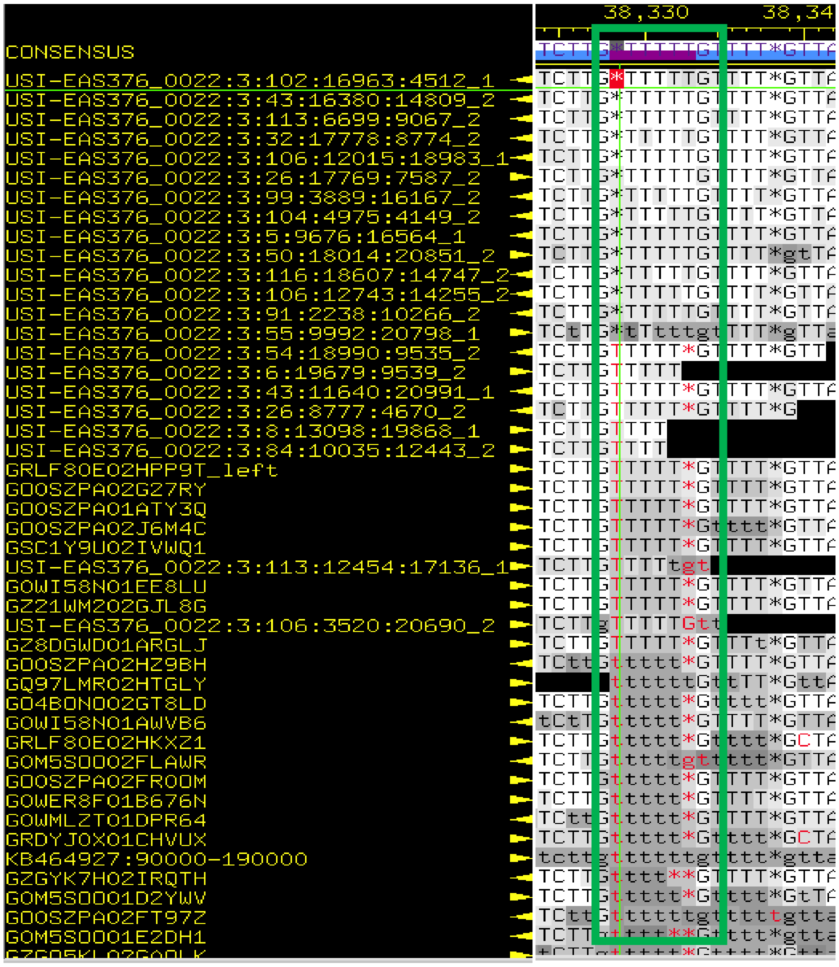


Figure 4: Highly discrepant position at 38,334: 6Ts are seen in the consensus sequence but Illumina and 454 reads have only 5Ts with a pad on one side (green box). The blue tag on the consensus corresponds to a repeat element tag for likely transposable elements. The purple tag corresponds to two overlapping repeat tags.

consensus sequence. To remedy the error, the first T in the MNR was replaced with a pad to align with the Illumina reads (Figure 5). This instance was the only time an element besides an A or a T was added to correct the consensus sequence.



```
                                             38,330     38,34
CONSENSUS                                  TCTTG*TTTTTGTTTT*GTTA
USI-EAS376_0022:3:102:16963:4512_1   ◄    TCTTG*TTTTTGTTTT*GTTA
USI-EAS376_0022:3:43:16380:14809_2   ◄    TCTTG*TTTTTGTTTT*GTTA
USI-EAS376_0022:3:113:6699:9067_2    ◄    TCTTG*TTTTTGTTTT*GTTA
USI-EAS376_0022:3:32:17778:8774_2    ◄    TCTTG*TTTTTGTTTT*GTTA
USI-EAS376_0022:3:106:12015:18983_1  ◄    TCTTG*TTTTTGTTTT*GTTA
USI-EAS376_0022:3:26:17769:7587_2    ►    TCTTG*TTTTTGTTTT*GTTA
USI-EAS376_0022:3:99:3889:16167_2    ◄    TCTTG*TTTTTGTTTT*GTTA
USI-EAS376_0022:3:104:4975:4149_2    ◄    TCTTG*TTTTTGTTTT*GTTA
USI-EAS376_0022:3:5:9676:16564_1     ◄    TCTTG*TTTTTGTTTT*GTTA
USI-EAS376_0022:3:50:18014:20851_2   ►    TCTTG*TTTTTGTTTT*gtTA
USI-EAS376_0022:3:116:18607:14747_2  ◄    TCTTG*TTTTTGTTTT*GTTA
USI-EAS376_0022:3:106:12743:14255_2  ◄    TCTTG*TTTTTGTTTT*GTTA
USI-EAS376_0022:3:91:2238:10266_2    ◄    TCTTG*TTTTTGTTTT*GTTA
USI-EAS376_0022:3:55:9992:20798_1    ►    TCtTG*tTtttgtTTT*gTTa
USI-EAS376_0022:3:54:18990:9535_2    ◄    TCTTGTTTTT*GTTTT*GTT
USI-EAS376_0022:3:6:19679:9539_2     ►    TCTTGTTTTT
USI-EAS376_0022:3:43:11640:20991_1   ◄    TCTTGTTTTT*GTTTT*GTTA
USI-EAS376_0022:3:26:8777:4670_2     ◄    TCTTGTTTTT*GTTTT*G
USI-EAS376_0022:3:8:13098:19868_1    ►    TCTTGTTTT
USI-EAS376_0022:3:84:10035:12443_2   ►    TCTTGTTTT
GRLF8OEO2HPP9T_left                  ►    TCTTGTTTTT*GTTTT*GTTA
GOOSZPAO2G27RY                       ►    TCTTGTTTTT*GTTTT*GTTA
GOOSZPAO1ATY3Q                       ►    TCTTGTTTTT*GTTTT*GTTA
GOOSZPAO2J6M4C                       ►    TCTTGTTTTT*Gtttt*GTTA
GSC1Y9UO2IVWQ1                       ►    TCTTGTTTTT*GTTTT*GTTA
USI-EAS376_0022:3:113:12454:17136_1  ►    TCTTGTTTTtgt
GOWI58NO1EE8LU                       ►    TCTTGTTTTT*GTTTT*GTTA
GZ21WM2O2GJL8G                       ►    TCTTGTTTTT*GTTTT*GTTA
USI-EAS376_0022:3:106:3520:20690_2   ►    TCTTgTTTTTGtt
GZ8DGWDO1ARGLJ                       ►    TCTTGTTTTT*GTTTt*GTTA
GOOSZPAO2HZ9BH                       ◄    TCttGttttt*GTTTT*GTTA
GQ97LMRO2HTGLY                       ►         ttttttGttTT*GttA
GO4BONOO2GT8LD                       ►    TCTTGttttt*Gtttt*GTTA
GOWI58NO1AWVB6                       ◄    tCtTGttttt*GTTTT*GTTA
GRLF8OEO2HKXZ1                       ►    TCTTGttttt*Gtttt*GCTA
GOM5SOOO2FLAWR                       ◄    TCTTGtttttgttttt*GTTA
GOOSZPAO2FROOM                       ►    TCTTGttttt*GTTTT*GTTA
GOWER8FO1B676N                       ►    TCTTGttttt*GTTTT*GTTA
GOWMLZTO1DPR64                       ◄    TCttGttttt*GTTTT*GTTA
GRDYJOXO1CHVUX                       ►    TCTTGttttt*Gtttt*GCTA
KB464927:90000-190000                ►    tcttgttttttgtttt*gtta
GZGYK7HO2IRQTH                       ◄    TCTTGtttt**GTTTT*GTTA
GOM5SOOO1D2YWV                       ►    TCTTGttttt*Gtttt*GtTA
GOOSZPAO2FT97Z                       ◄    TCttGttttttgtttttgtta
GOM5SOOO1E2DH1                       ◄    TCTTgtttt**Gtttt*gtta
```

Figure 5: Corrected consensus at position 38,334: A T at the beginning of the MNR was replaced with a pad (*) to align consensus with high quality Illumina reads (green box). MNR changed from 6Ts to 5Ts.

Another difficult case was at position 6397, where two MNRs appeared in a row: seven Ts and eight As with a pad in between them. Five Illumina reads had an A instead of a T to the right of the pad and three Illumina reads which had an extra A at the end of the mono-A run. Additionally, 17 Illumina reads had an extra A at the beginning of the mono-A run, having nine As instead of the eight As in the consensus. The 454 reads were of very low quality and were not helpful in correcting the issue. An A was added to replace the pad between the two MNRs, making the mono-A run nine As instead of eight As and confirming the mono-T run of seven Ts (Figure 6).

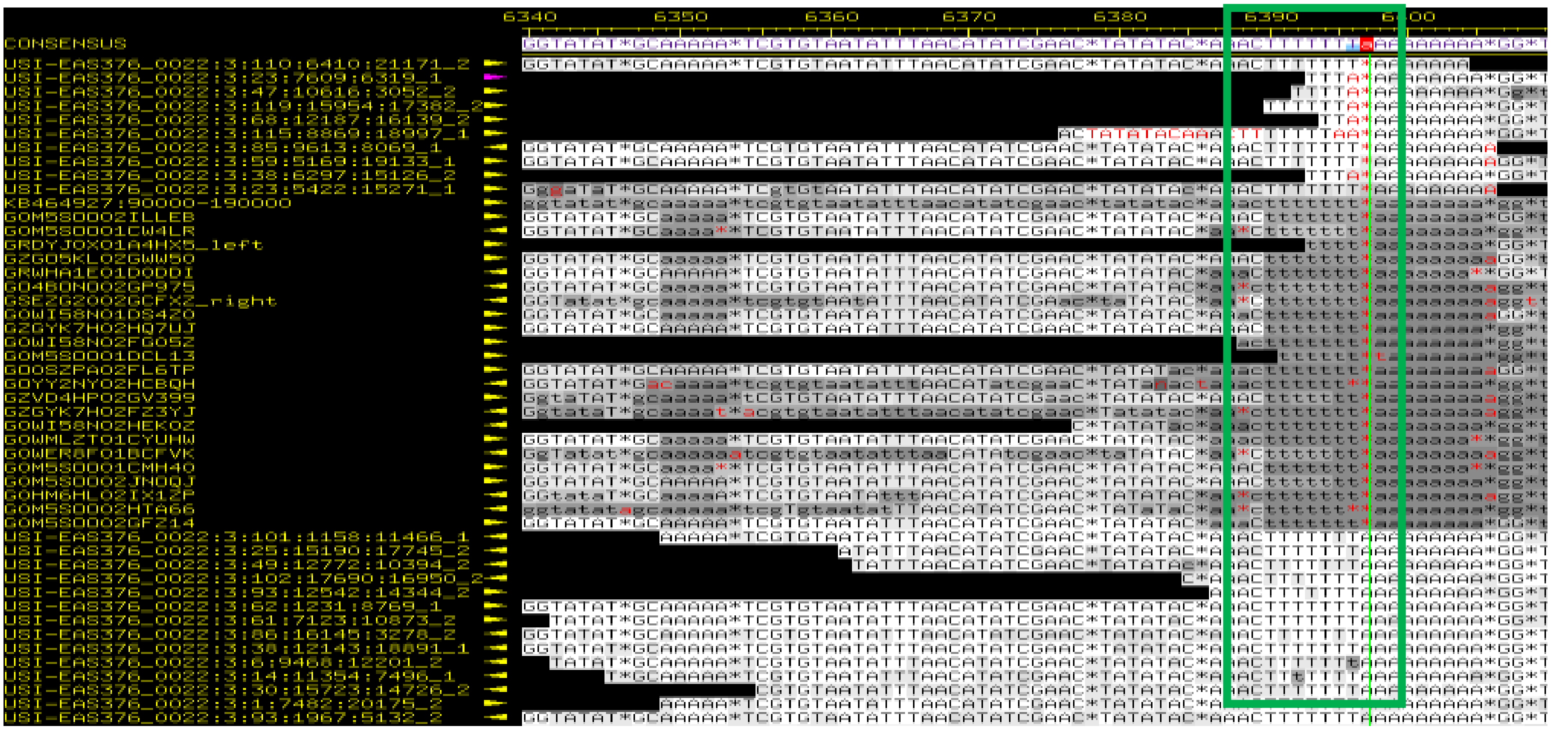
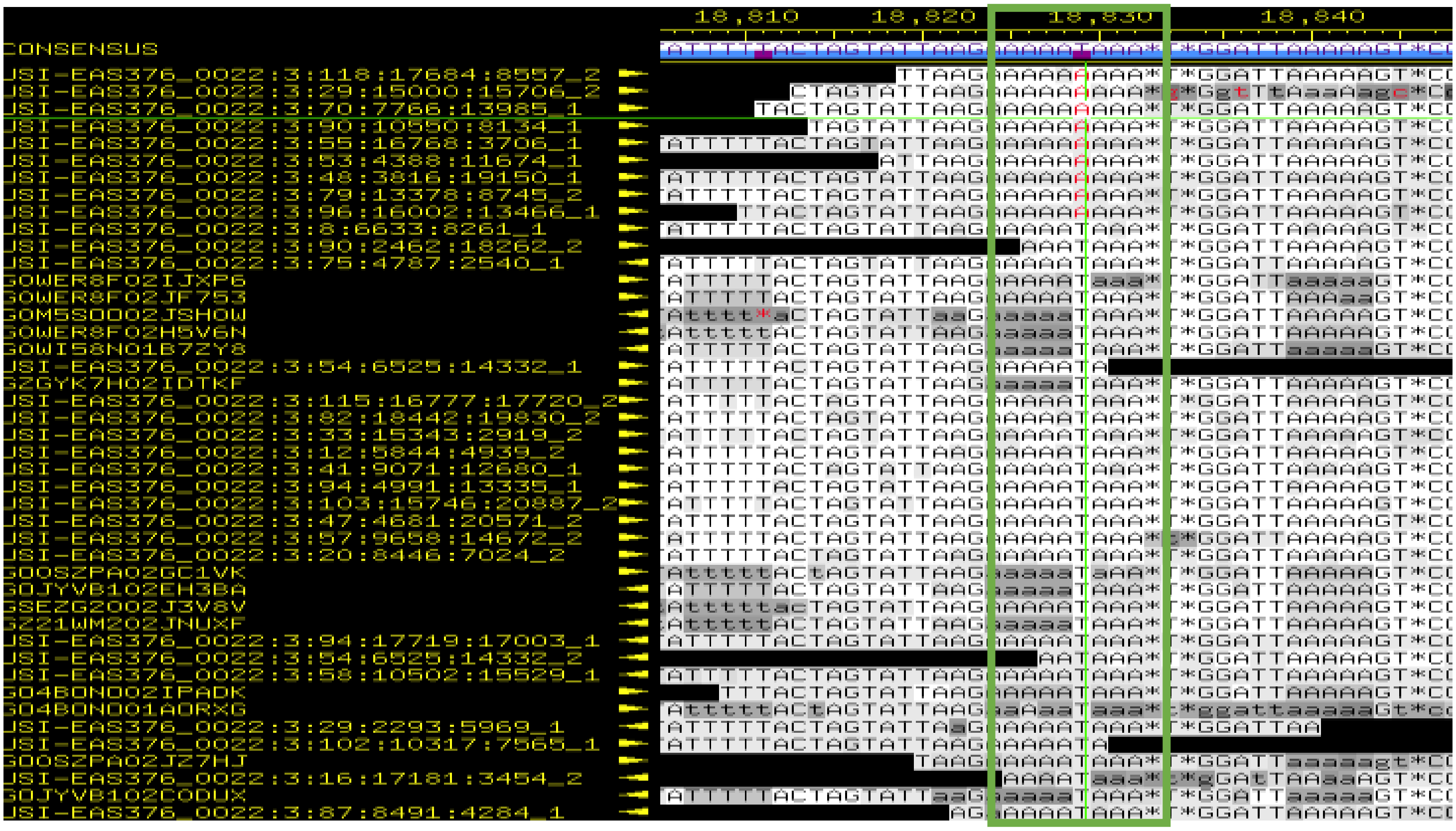


Figure 6: Corrected highly discrepant position at 6397: Consensus changed from MNRs of seven Ts and eight As to seven Ts and nine As to align with high quality Illumina reads (green box). Lowercase "a" seen in consensus marks edit made to consensus sequence. Blue tag on T corresponds to comment tag added to consensus.

In total, 73 (53%) of all identified high quality discrepancies were MNRs. Forty-two of the MNRs were confirmed and 31 bases were added to correct the remaining MNRs (Table 1). Except for the one mentioned position where a pad was added, all corrections involved the addition of an A or a T.

Polymorphisms

After all MNRs were identified, the remaining high quality discrepancies were analyzed for polymorphisms. A polymorphism corresponds to a position where the high quality reads indicate that two bases are equally likely. In other words, it is a position where there is an approximate 50/50 split among the high quality reads as to which base should be in the consensus. No polymorphisms were identified in the contig, although there was one interesting case (Figure 7). At position 18,829, there was an MNR of 5As, followed by a T and another 3As. In 9 Illumina reads, the T is replaced with an A, resulting in a 9A MNR. In the remaining Illumina and 454 reads (over 20 reads), the T is kept in that position. Based on this data, the consensus was not changed and the T was kept within the MNR. This region is particularly interesting because the incorrect Illumina reads, except for one of them, do not show other signs of being incorrectly mapped. On the other hand, this region is in the middle of a repeat tag, where incorrect mapping and misalignment is very frequent. This region was not marked as a potential polymorphism and no polymorphisms were identified in the contig.

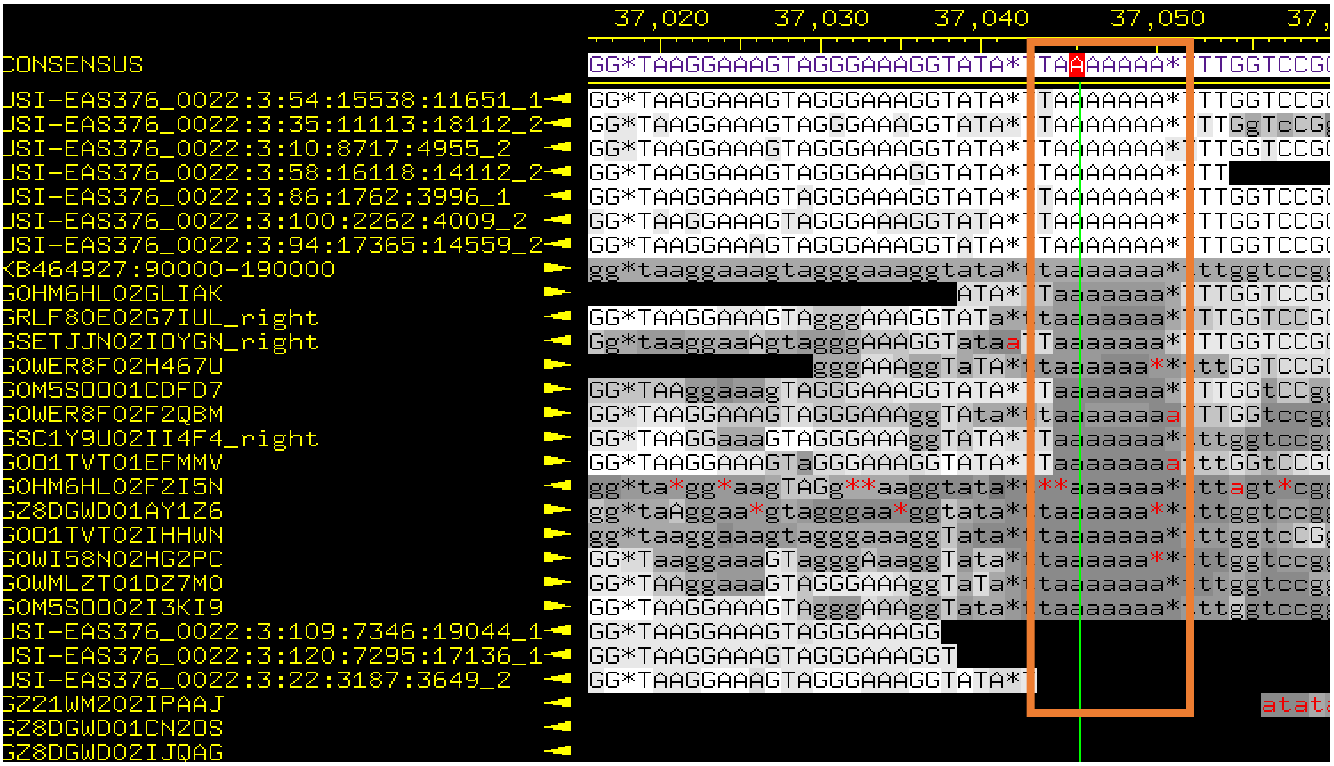


```
                                         18,810      18,820      18,830      18,840
CONSENSUS                                ATTTTTACTAGTATTAAGAAAAATAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:118:17684:8557_2                    TTAAGAAAAAAAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:29:15000:15706_2              CTAGTATTAAGAAAAAAAAA*g*GgtTtAaaaAagc*C
JSI-EAS376_0022:3:70:7766:13985_1              TACTAGTATTAAGAAAAAAAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:90:10550:8134_1                TAGTATTAAGAAAAAAAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:55:16768:3706_1        ATTTTTACTAGTATTAAGAAAAAAAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:53:4388:11674_1                    ATTAAGAAAAAAAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:48:3816:19150_1        ATTTTTACTAGTATTAAGAAAAAAAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:79:13378:8745_2        ATTTTTACTAGTATTAAGAAAAAAAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:96:16002:13466_1           TTACTAGTATTAAGAAAAAAAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:8:6633:8261_1          ATTTTTACTAGTATTAAGAAAAATAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:90:2462:18262_2                            AAATAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:75:4787:2540_1         ATTTTTACTAGTATTAAGAAAAATAAA*T*GGATTAAAAAGT*C
GOWER8F02IJXP6                           ATTTTTACTAGTATTAAGAAAAATaaa*T*GGATTaaaaaGT*C
GOWER8F02JF753                           ATTTTTACTAGTATTAAGAAAAATAAA*T*GGATTAAAaaGT*C
GOM5SOOO2JSHOW                           Atttt*aCTAGTATTaaGaaaaaTAAA*T*GGATTAAAAAGT*C
GOWER8F02H5V6N                           AtttttACTAGTATTAAGaaaaaTAAA*T*GGATTAAAAAGT*C
GOWI58NO1B7ZY8                           ATTTTTACTAGTATTAAGaaaaaTAAA*T*GGATTaaaaaGT*C
JSI-EAS376_0022:3:54:6525:14332_1        ATTTTTACTAGTATTAAGAAAAATA
GZGYK7HO2IDTKF                           ATTTTTACTAGTATTAAGaaaaaTAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:115:16777:17720_2      ATTTTTACTAGTATTAAGAAAAATAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:82:18442:19830_2       ATTTTTACTAGTATTAAGAAAAATAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:33:15343:2919_2        ATTTTTACTAGTATTAAGAAAAATAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:12:5844:4939_2         ATTTTTACTAGTATTAAGAAAAATAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:41:9071:12680_1        ATTTTTACTAGTATTAAGAAAAATAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:94:4991:13335_1        ATTTTTACTAGTATTAAGAAAAATAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:103:15746:20887_2      ATTTTTACTAGTATTAAGAAAAATAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:47:4681:20571_2        ATTTTTACTAGTATTAAGAAAAATAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:57:9658:14672_2        ATTTTTACTAGTATTAAGAAAAATAAA*t*GGATTAAAAAGT*C
JSI-EAS376_0022:3:20:8446:7024_2         ATTTTTACTAGTATTAAGAAAAATAAA*T*GGATTAAAAAGT*C
GDOSZPA02GC1VK                           AtttttACtAGTATTAAGaaaaaTaAA*T*GGATTAAAAAGT*C
GOJYVB102EH3BA                           ATTTTTACTAGTATTAAGaaaaaTAAA*T*GGATTAAAAAGT*C
GSEZG2002J3V8V                           AtttttacTAGTATTAAGAAAAATAAA*T*GGATTAAAAAGT*C
GZ21WM202JNUXF                           AtttttACTAGTATTAAGaaaaaTAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:94:17719:17003_1       ATTTTTACTAGTATTAAGAAAAATAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:54:6525:14332_2                             AATAAA*T*GGATTAAAAAGT*C
JSI-EAS376_0022:3:58:10502:15529_1       ATTTTTACTAGTATTAAGAAAAATAAA*T*GGATTAAAAAGT*C
GD4BON002IPADK                              TTTACTAGTATTAAGAAAAATAAA*T*GGATTAAAAAGT*C
GD4BON001AORXG                           AtttttACtAGTATTAAGaaAaaTaaa*t*ggattaaaaaGt*c
JSI-EAS376_0022:3:29:2293:5969_1         ATTTTTACTAGTATTAaGAAAAATAAA*T*GGATTAA
JSI-EAS376_0022:3:102:10317:7565_1       ATTTTTACTAGTATTAAGAAAAATA
GDOSZPA02JZ7HJ                                         TAAGAAAAATAAA*T*GGATTaaaaagt*C
JSI-EAS376_0022:3:16:17181:3454_2                           AAAATaaa*t*gGAtTAAaaAGT*C
GOJYVB102CODUX                           ATTTTTACTAGTATTaaGaaaaaTAAA*T*GGATTaaaaaGT*C
JSI-EAS376_0022:3:87:8491:4284_1                         AGAAAAATAAA*T*GGATTAAAAAGT*C
```

Figure 7: Interesting high quality discrepancy at position 18,829: 9 high quality Illumina reads indicate an incorrect base (A) compared to the consensus (T) (region within green box). The consensus was not changed because other Illumina and 454 reads show inclusion of the T. Additionally, the region was not marked as polymorphism because there is not a 50/50 split between reads, and the region is in a repeat tag.

Low Coverage Regions

After going through all high quality discrepancies, the next regions investigated were areas with low coverage. Areas of low coverage corresponded to regions with fewer than 40 reads covering the sequence. Thirty-five areas of low coverage were identified. One of these areas was the gap found in the contig and is discussed in the next section. In the remaining 34 low coverage regions, MNRs were identified, but no evidence was found to change the consensus at these locations. There were at least five Illumina reads in each region that matched the consensus. An example of an area of low coverage is show in Figure 8. In the end, no bases were changed in low coverage regions.



```
                                        37,020    37,030    37,040     37,050     37,
CONSENSUS                               GG*TAAGGAAAGTAGGGAAAGGTATA*TTAAAAAAA*TTTGGTCCG
USI-EAS376_0022:3:54:15538:11651_1   ◄  GG*TAAGGAAAGTAGGGAAAGGTATA*TTAAAAAAA*TTTGGTCCG
USI-EAS376_0022:3:35:11113:18112_2   ◄  GG*TAAGGAAAGTAGGGAAAGGTATA*TTAAAAAAA*TTTGgTcCGg
USI-EAS376_0022:3:10:8717:4955_2     ◄  GG*TAAGGAAAGTAGGGAAAGGTATA*TTAAAAAAA*TTTGGTCCG
USI-EAS376_0022:3:58:16118:14112_2   ◄  GG*TAAGGAAAGTAGGGAAAGGTATA*TTAAAAAAA*TTT
USI-EAS376_0022:3:86:1762:3996_1     ◄  GG*TAAGGAAAGTAGGGAAAGGTATA*TTAAAAAAA*TTTGGTCCG
USI-EAS376_0022:3:100:2262:4009_2    ◄  GG*TAAGGAAAGTAGGGAAAGGTATA*TTAAAAAAA*TTTGGTCCG
USI-EAS376_0022:3:94:17365:14559_2   ◄  GG*TAAGGAAAGTAGGGAAAGGTATA*TTAAAAAAA*TTTGGTCCG
KB464927:90000-190000                ►  gg*taaggaaagtagggaaaggtata*ttaaaaaaa*tttggtccgg
GOHM6HL02GLIAK                       ►                         ATA*TTaaaaaaa*TTTGGTCCG
GRLF80E02G7IUL_right                 ◄  GG*TAAGGAAAGTAgggAAAGGTATa*ttaaaaaaa*TTTGGTCCG
GSETJJN02IOYGN_right                 ◄  Gg*taaggaaAgtagggAAAGGTataaTTaaaaaaa*TTTGGTCCG
GOWER8F02H467U                       ►                gggAAAggTaTA*ttaaaaaa**tttGGTCCG
GOM5SO001CDFD7                       ►  GG*TAAggaaagTAGGGAAAGGTATA*TTaaaaaaa*TTTGGtCCgg
GOWER8F02F2QBM                       ►  GG*TAAGGAAAGTAGGGAAAggTAta*ttaaaaaaaaTTTGGtccgg
GSC1Y9U02II4F4_right                 ►  GG*TAAGGaaaGTAGGGAAAggTATA*TTaaaaaaa*tttggtccgg
GOO1TVTO1EFMMV                       ►  GG*TAAGGAAAGTaGGGAAAGGTATA*TTaaaaaaaatttGGtCCG
GOHM6HL02F2I5N                       ◄  gg*ta*gg*aagTAGg**aaggtata*t**aaaaaa*tttagt*cgg
GZ8DGWD01AY1Z6                       ►  gg*taAggaa*gtagggaa*ggtata*ttaaaaaa**tttggtccgg
GOO1TVTO2IHHWN                       ►  gg*taaggaaagtagggaaaggTata*ttaaaaaaa*tttggtcCGg
GOWI58NO2HG2PC                       ►  GG*TaaggaaaGTagggAaaggTata*ttaaaaaa**tttggtccgg
GOWMLZTO1DZ7MO                       ►  GG*TAAggaaaGTAGGGAAAggTaTa*ttaaaaaaa*tttggtccgg
GOM5SO002I3KI9                       ►  GG*TAAGGAAAGTAgggAAAggTata*ttaaaaaaa*tttggtccgg
USI-EAS376_0022:3:109:7346:19044_1   ◄  GG*TAAGGAAAGTAGGGAAAGG
USI-EAS376_0022:3:120:7295:17136_1   ◄  GG*TAAGGAAAGTAGGGAAAGGT
USI-EAS376_0022:3:22:3187:3649_2     ◄  GG*TAAGGAAAGTAGGGAAAGGTATA*T
GZ21WM202IPAAJ                       ◄                                                  atata
GZ8DGWD01CN2OS                       ◄
GZ8DGWD02IJQAG                       ◄
```

Figure 8: Example of low coverage region: MNR of As in area of low coverage starts at position 37,044 and is highlighted by orange box. Even though this is an area of low coverage, there is no evidence that the consensus should be changed.

Low Consensus Quality Regions

With all high quality discrepancies and areas of low coverage analyzed for MNRs and polymorphisms, the next step was to look at regions of low consensus quality. Regions with Phred quality scores less than or equal to 25 or 98 were considered low consensus quality regions. (Regions that are edited by the finisher are automatically given a quality score of 98.) In total, there were 32 low consensus quality regions, but 31 of those were edits made to the consensus sequence. Therefore, there was only one low consensus quality region to examine. This low consensus quality region was found at positions 90,392-90,415 and corresponded to a gap in the contig (Figure 9). To attempt to resolve the gap by a forced join, a unique sequence had to be found on either side of the gap. The sequence ATAAAGTGTATAAAATATATTA was identified to be slightly upstream of the gap and immediately following the gap (Figure 10). Searching for the sequence throughout the contig showed that it only appeared at these two locations. The contig-spanning read used to initially assemble the contig (KB464927:90000-190000) was removed to allow for tearing. The contig was torn at the first A of the overlapping sequence, creating two contigs (Figure 11). Next, the two contigs were compared using the overlapping sequence and were aligned to see if a forced join could be completed (Figure 12).

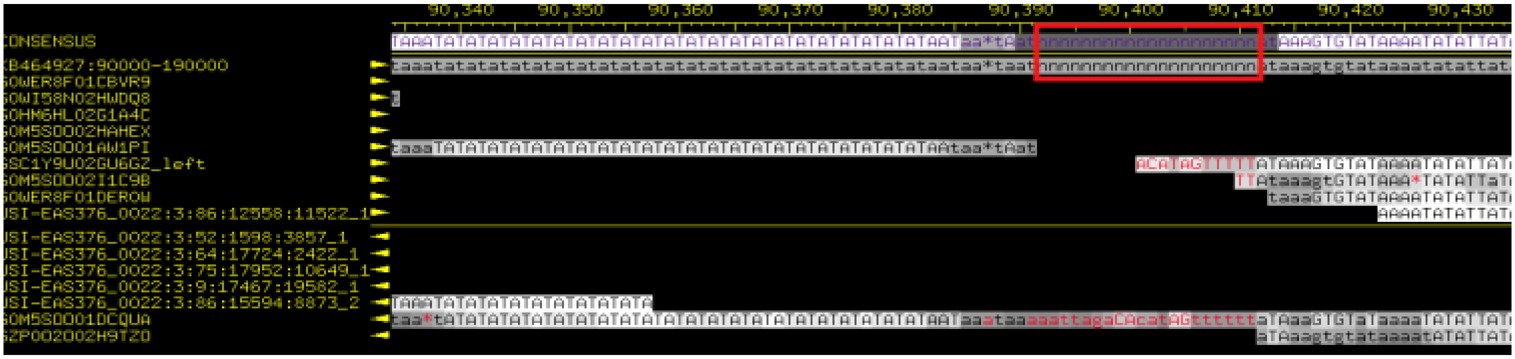


Figure 9: Picture of gap: Gap found at positions 90,392-90,415 (region within red box).

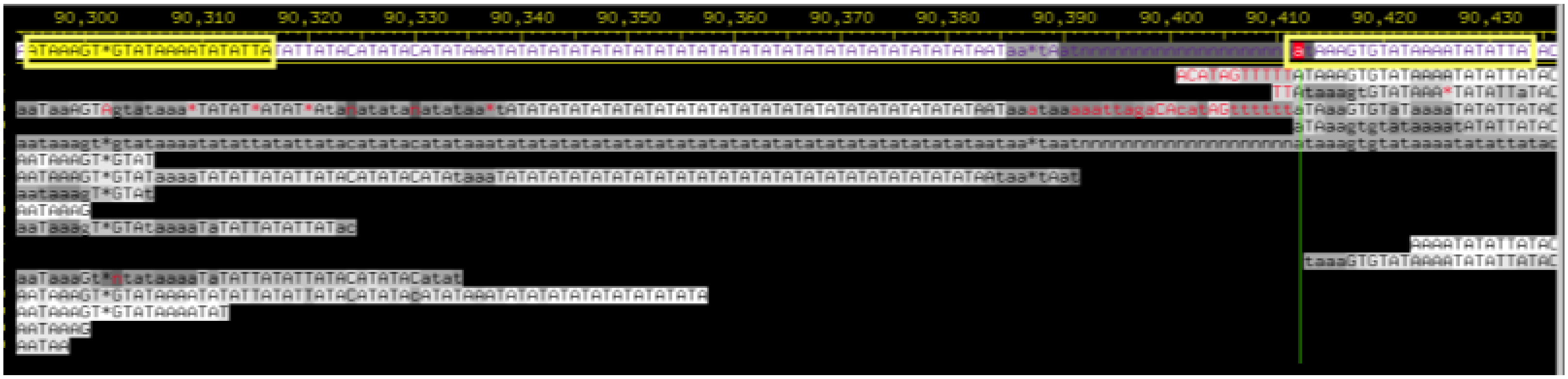


Figure 10: Common sequence found on either side of gap: The sequence ATAAAGTGTATAAAATATATTA, highlighted by yellow box, was found on either side of the gap, indicating that the gap may be resolved by matching up the sequences. This unique sequence was only found at these two locations.

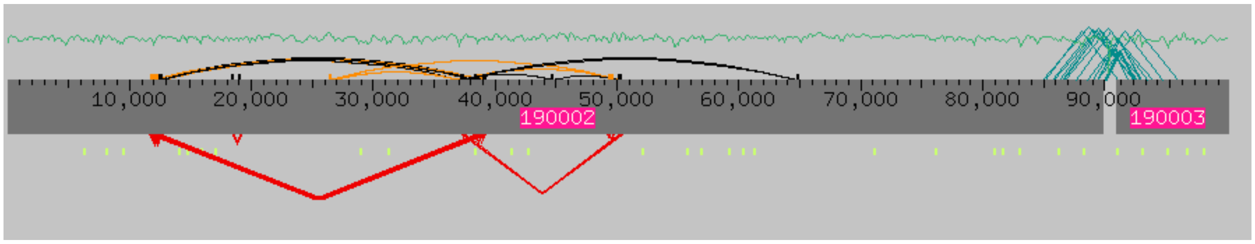
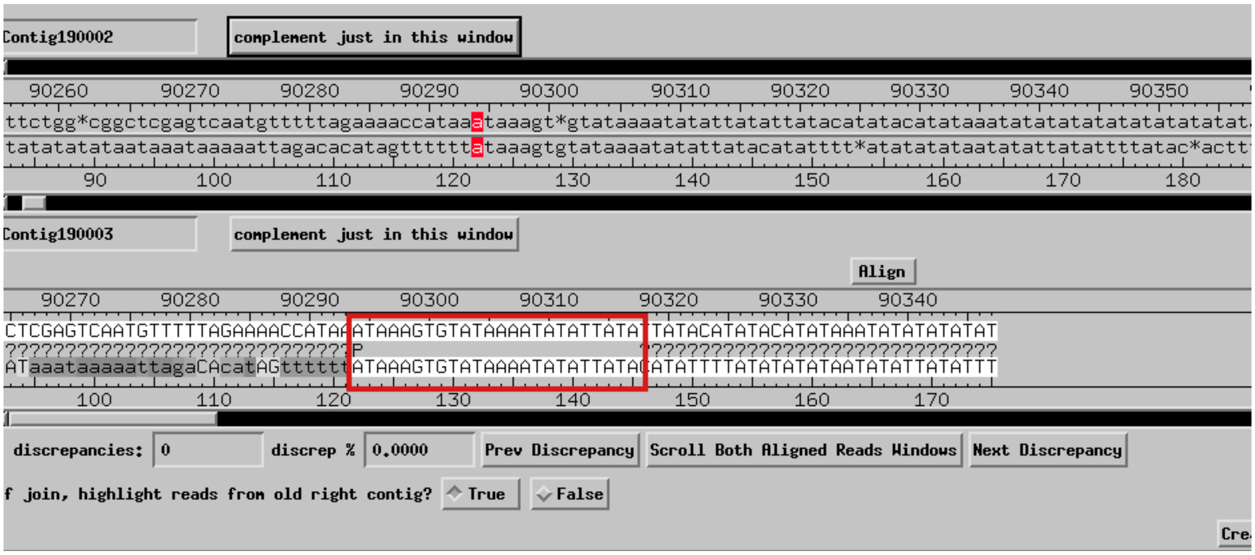


Figure 11: Assembly View post-tear: The Assembly View of the contigs after the initial contig was torn shows two sepearte contigs (190002 and 190003). As with the initial Assembly View, the green line represents depth of reads across the contig. Red lines indicate incorrect spacing or orientation between forward and reverse read pairs. Black and orange lines indicate repetitious elements that map to multiple locations. The blue/green lines near the gap indicate paired reads from the same sequencing reaction that span the gap.



```
Contig190002        complement just in this window
 90260      90270      90280      90290      90300      90310      90320      90330      90340      90350
ttctgg*cggctcgagtcaatgttttagaaaccataaataaagt*gtataaaatatattatattatacatatacatataaatatatatatatatatatat
tatatatataataaataaaaattagacacatagtttttttataaagtgtataaaatatattatacatatttt*atatatataatatattatattttatac*actt
     90        100        110        120        130        140        150         160        170        180

Contig190003        complement just in this window
                                                                                     Align
  90270      90280      90290      90300      90310      90320      90330      90340
CTCGAGTCAATGTTTTTAGAAAACCATAAATAAAGTGTATAAAATATATTATATTATACATATACATATAAATATATATATAT
?????????????????????????????P                    ??????????????????????????????
ATaaataaaaattagaCAcatAGttttttATAAAGTGTATAAAATATATTATACATATTTTATATATATAATATATTATATTT
      100        110        120        130        140        150        160        170

discrepancies: 0        discrep % 0.0000     Prev Discrepancy   Scroll Both Aligned Reads Windows   Next Discrepancy
f join, highlight reads from old right contig?  ◇ True   ◇ False
                                                                                                          Cre
```

Figure 12: Compare Contig View: Upon comparing the two contigs, the overlapping sequence matches up (red box), but unique sequences on either side prevent a forced join.

Unfortunately, even though the overlapping sequence was the same for both contigs, there were unique sequences on either side of the overlapping sequence that prevented a force join. Additionally, when looking at the post-tear Assembly View in Figure 11, Crossmatch results show that there are no repeat sequences found on both sides of the gap. Furthermore, there were many paired end reads spanning the gap, indicating that a forced join would not resolve the gap (estimated size of ~1200bp). Since the gap could not be resolved, more sequencing data is needed. PCR primers were created using *Consed* to cover the gap and the neighboring low quality areas (Table 2). Two primer pairs were chosen to order. Each pair met ideal PCR primer criteria, with the distance between them being less than 1000 bp and the primers having the same melting temperature. Two pairs were chosen to maximize the coverage of the area. Once sequencing data for this region is added, the contig will be finished and annotation may commence.

Final Assembly/Conclusion

Figure 13 shows the final assembly of contig DEUG4927002. It is not drastically different from the initial assembly of DEUG4927002, but it shows the edited bases (31 total) and the PCR primers created to cover the gap. The red lines, representing misaligned reads, were not removed and realigned to the contig because they did not contaminate the consensus sequence. Although not visible in the photo, all MNRs associated with high quality discrepancies and regions of low coverage (73 total) were investigated and either confirmed or corrected. Upon receiving sequencing data for the gap, the contig will be ready for annotation.

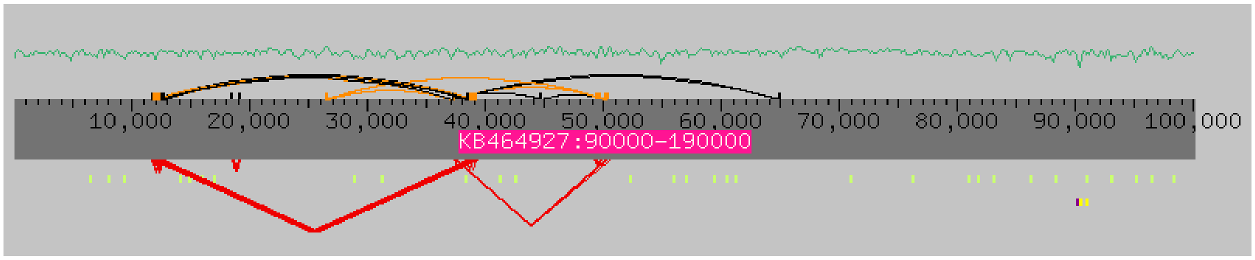


Figure 13: **Final Assembly View of DEUG4927002:** Green marks below contig represent edited bases and yellow marks below contig represent PCR primers constructed.

Acknowledgments

This research project would not have been possible without the help of the Bio434W teaching team. I would like to thank Dr. Elgin, Dr. Shaffer, Wilson Leung, Lee Trani, and Ryan Freidman for their overall guidance and wisdom, their expertise on *Consed*, and their assistance in my project, and Dr. Bednarski for improving my writing. I would also like to thank Washington University in St. Louis and the Genomics Education Partnership for making this research possible.

Table 1: List of MNRs: All MNRs identified in Contig DEUG4927002; under "Evidence," number of reads used as evidence given if sequence was changed.

Position	Analysis	Change to Consensus	Evidence
3338	MNR of Ts	+T	Illumina (21 reads)
3950	MNR of As	confirmed consensus	Illumina
3976	MNR of As	+A	Illumina (20 reads)
6252	MNR of As	confirmed consensus	Illumina
6388	MNR of Ts	confirmed consensus	Illumina
6397	MNR of As	+A	Illumina (17 reads)
6864	MNR of As	confirmed consensus	Illumina
8016	MNR of Ts	confirmed consensus	Illumina
8111	MNR of Ts	+T	Illumina (15 reads)
9473	MNR of As	+A	Illumina (20 reads)
12,593	MNR of As	confirmed consensus	Illumina
14,047	MNR of As	+A	Illumina (18 reads)
14,780	MNR of As	+A	Illumina (9 reads)
15,954	MNR of As	+A	Illumina (15 reads)
16,977	MNR of As	+A	Illumina (10 reads)
17,428	MNR of Ts	confirmed consensus	Illumina
18,804	MNR of Ts	confirmed consensus	Illumina
18,822	2 MNRs of As with T in the middle	confirmed consensus	Illumina (9 Illumina reads incorrectly have single MNR of As)
19,979	MNR of As	confirmed consensus	Illumina
20,016	MNR of As	confirmed consensus	Illumina
20,958	MNR of As	confirmed consensus	Illumina
21,498	MNR of As	confirmed consensus	Illumina
21,567-21,568	MNR of As	confirmed consensus	Illumina
21,872	MNR of As	confirmed consensus	Illumina
22,275	MNR of As	confirmed consensus	Illumina
24,432	MNR of As	confirmed consensus	Illumina
27,832	MNR of As	confirmed consensus	Illumina
28,878	MNR of As	+A	Illumina (5 reads)
30,354	MNR of As	confirmed consensus	Illumina
31,249	MNR of Ts	+T	Illumina (10 reads)
31,559	MNR of Ts	confirmed consensus	Illumina
34,946	MNR of Ts	confirmed consensus	Illumina
36,261	MNR of As	confirmed consensus	Illumina
37,842	MNR of Ts	confirmed consensus	Illumina
38,334	MNR of Ts	+ pad (*)	Ilumina (20 reads)
41,244	MNR of Ts	+T	Illumina (21 reads)
42,613	MNR of As	+A	Illumina (22 reads)

44,295	MNR of As	confirmed consensus	Illumina
46,084	MNR of Ts	confirmed consensus	Illumina
47,016	MNR of As	confirmed consensus	Illumina
52,185	MNR of Ts	+T	Illumina (12 reads)
52,486	MNR of Ts	confirmed consensus	Illumina
53,822	MNR of As	confirmed consensus	Illumina
55,540	MNR of Ts	confirmed consensus	Illumina
55,855	MNR of As	+A	Illumina (17 reads)
56,880	MNR of Ts	+T	Illumina (20 reads)
59,237	MNR of Ts	confirmed consensus	Illumina
59,270	MNR of As	+A	Illumina (15 reads)
60,396	MNR of As	+A	Illumina (15 reads)
61,252	MNR of As	+A	Illumina (19 reads)
61,426	MNR of As	confirmed consensus	Illumina
63,510	MNR of As	confirmed consensus	Illumina
68,006	MNR of As	confirmed consensus	Illumina
70,795	MNR of Ts	confirmed consensus	Illumina
70,841	MNR of Ts	confirmed consensus	Illumina
71,038	MNR of Ts	+T	Illumina (27 reads)
72,339	MNR of Ts	confirmed consensus	Illumina
73,391	MNR of Ts	confirmed consensus	Illumina
75,506	MNR of Ts	confirmed consensus	Illumina
76,163	MNR of As	+A	Illumina (19)
78, 926	MNR of Ts	confirmed consensus	Illumina
81,058	MNR of As	+A	Illumina (25)
81,609	MNR of Ts	+T	Illumina (20)
83,114	MNR of Ts	+T	Illumina (11)
86,256	MNR of As	+A	Illumina (17)
88,304	MNR of As	+A	Illumina (15)
90, 971	MNR of As	+A	Illumina (29)
93,092	MNR of Ts	+T	Illumina (6)
94,002	MNR of Ts	confirmed consensus	Illumina
94,433	MNR of Ts	confirmed consensus	Illumina
94,678-94,679	MNR of As	confirmed consensus	Illumina
94,861	MNR of As	confirmed consensus	Illumina
95,113	MNR of As	+A	Illumina (13 reads)
96,593	MNR of Ts	+T	Illumina (26 reads)
98,193	MNR of As	+A	Illumina (13 reads)

Table 2: PCR Primers: In table are two pairs of PCR primers chosen to cover gap from 90,392-90,415. P1 = primer 1; P2 = primer 2; Mp = melting temperature; Start-end = beginning and end position of primer in contig.

Distance	P1	Start-end	Mp	P2	Start-end	mp
311	gtcgaaaatatcgtatgatatcaat	90150-90174	55	tttatcaatttaaagaataaaattagacac	90487-90514	55
711	gtcgaaaatatcgtatgatatcaat	90150-90174	55	atgtgtaaacgctatacttagatgtc	90885-90910	55