Tina Zudock

Bio 434w—Research in Genomics

Dr. Elgin, Dr. Shaffer, Dr. Buhler, Dr. Bednarski

24 February 2017

Assembly of DEUG4927001: A 100 kb Contig on the *D. eugracilis* Dot Chromosome

Abstract

DEUG4927001 is the first contig in the genome map of the *D. eugracilis* dot chromosome (F element). This chromosome was sequenced using both 454 and Illumina sequencing technology. The assembly program Consed was used to manually assess the 100 kb contig for errors remaining in the consensus sequence after the initial assembly was created by the alignment software. The goal of this study was to determine the correct consensus sequence for the contig. The initial assembly for DEUG4927001 had 432 highly discrepant regions, three low consensus quality regions, 29 regions of low coverage, and no gaps. Initially, each highly discrepant region was analyzed. If the region was near a mononucleotide run, the Illumina reads for the region were used to determine the correct sequence. In the end, 40 base changes were made, making this contig ready for annotation. Forty-five of the highly discrepant regions were found to be putative single nucleotide polymorphisms and were tagged as such. This high frequency of polymorphisms compared to the other *D. eugracilis* dot chromosome contigs—which on average have 0-4 SNPs—makes DEUG4927001 an interesting region for further study.

Introduction

Most *Drosophila* species have a small chromosome known as the dot chromosome (F element). This chromosome four in *D. melanogaster* is unusual because it is believed to be entirely heterochromatic. In *D. melanogaster*, studies have shown that the 1.3 Mb arm of the chromosome has a normal gene density, but three times the density of repetitive elements on euchromatic chromosome arms, indicating that the repetitive sequences are involved in heterochromatin formation. How approximately 80 genes found on the F element are expressed and function in heterochromatin (at the same level as on euchromatin) is therefore an area of interest.

This study focuses on the dot chromosome of *Drosophila eugracilis*, a relative of *D. melanogaster*. *D. eugracilis* is an important organism to study because it is in an "evolutionary sweet spot" relative to *D. melanogaster*. *D. eugracilis* is more recently diverged from a common ancestor with *D. melanogaster* than other *Drosophila* relatives, so its dot chromosome regulatory motifs should still share enough similarities with *D. melanogaster's* to be recognizable. By comparing the dot chromosomes of many *Drosophila* species (more closely and more distantly related to *D. melanogaster*), we hope to find conserved regulatory motifs that will provide insight into how gene regulation functions on this unusual chromosome.

Initial Assembly

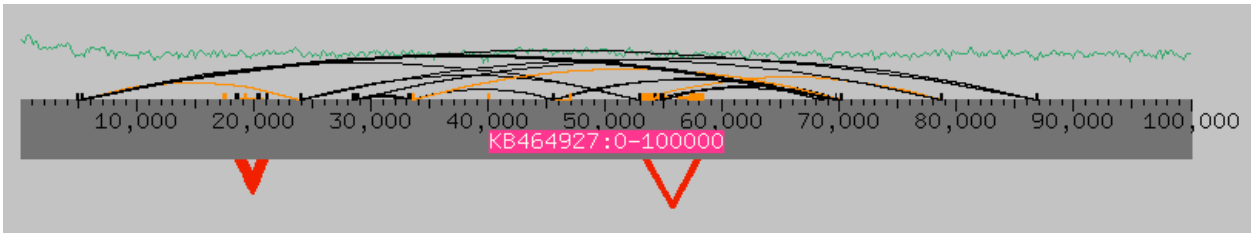

Figure 1: Initial assembly view of DEUG4927001. Red lines represent incorrect spacing for forward/reverse pairs while black and orange lines represent repetitive regions. The green line at the top indicates depth of coverage.

The initial assembly of DEUG4927001 is a ~100 kb long section of the *D. eugracilis* dot chromosome (Figure 1). This contig represents the first 100 kb of the dot chromosome assembly (Figure 2). The initial assembly contains 432 highly discrepant regions and three low consensus quality regions. This contig did not contain any gaps. Additionally, this region contained a large number of repetitive sequences, represented by the orange and black lines in Figure 1. These are not necessarily repetitive elements, but rather, multiple places where a specific sequence maps. Because this contig represents the first 100 kb of the genome map for the dot chromosome, the first 5 kb of this contig has a large number of misaligned reads, shown by the large rise and gradual drop of the green coverage line from 0 to 5 kb at the top of Figure 1.

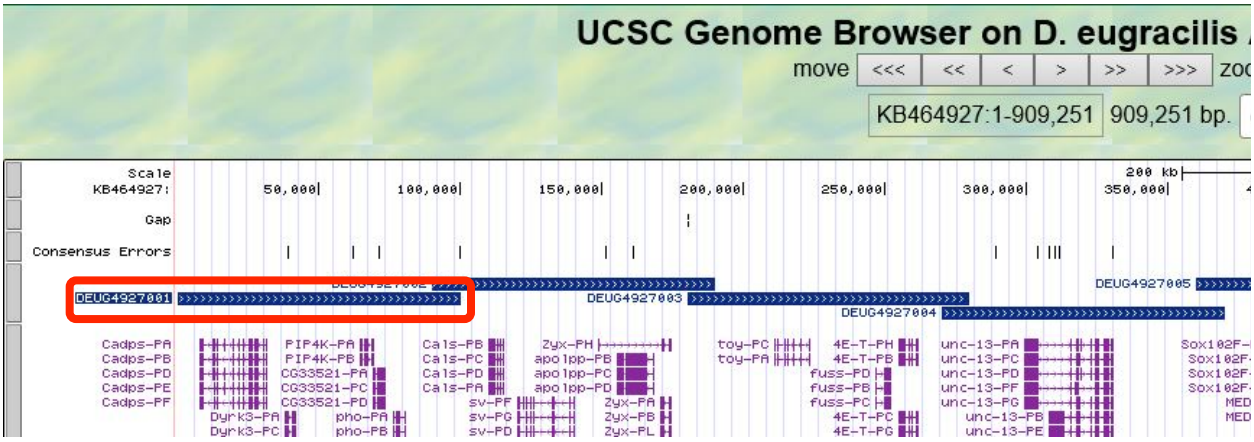

Figure 2: View of *D. eugracilis* dot chromosome genome map using UCSC Genome Browser. DEUG4927001 contig is circled in red.

High Quality Discrepencies

The most common problem with the initial assembly was high quality discrepencies. These are defined as positions at which three or more reads with quality scores above 30 differ from the consensus sequence. Each of the 432 high quality discrpencies in DEUG4927001 was examined; however, the discrepencies at mononucleotide runs were investigated first.

Mononucleotide runs with more than five of the same nucleotide in a row are especially prone to having high quality discrepencies. This is due to the nature of 454 pyrosequencing. In 454 sequencing, one type of nucleotide (either A, G, T or C) is available for addition at a time. When a base is added to the growing strand, light is emitted, so the addition can be recorded. If multiple nucleotides are added at the same time, the flash of light will be of greater intensity. However, as more and more nucleotides are added at the same time, it becomes increasingly difficult to determine the exact number of bases added. For this reason, 454 reads often differ on the number of bases in mononucleotide runs. To determine the correct sequence at mononucleotide runs with high quality discrepencies in DEUG4927001, the high quality Illumina reads were examined. Illumina sequencing does not experience problems around mononucleotide runs because in this sequencing method, only one nucleotide is added at a time. Ninety-one high quality discrepencies at mononucleotide runs in the DEUG4927001 contig were examined. Thirty-three of these high quality discrepencies were changed on the consensus strand, all to agree with Illumina sequencing data (Table 1S). An example of high quality discrepency analysis at a mononucleotide run is shown below.

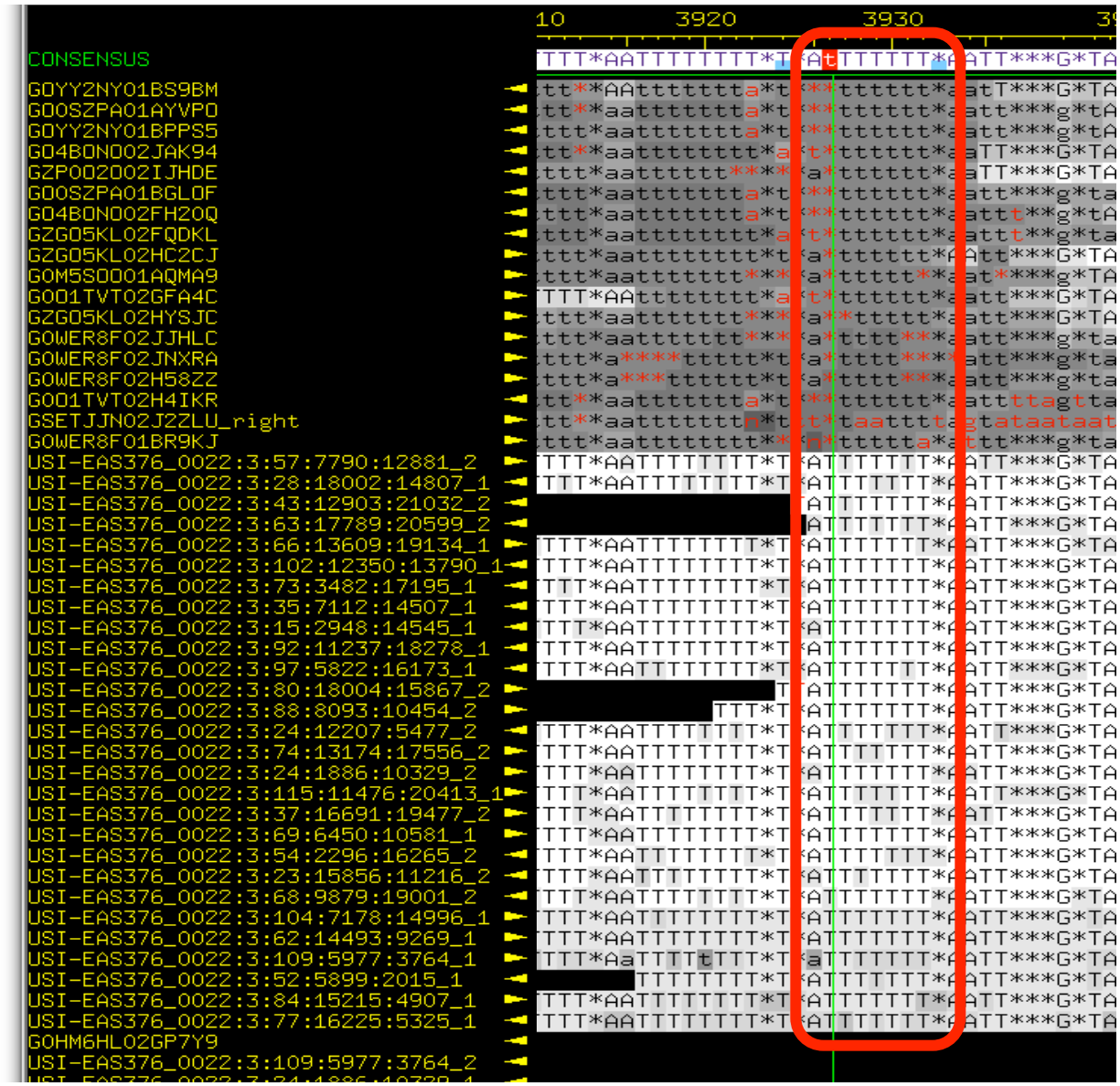

Figure 3: Mononucleotide run with 7 ‘T’s starting at 3.926 kb circled in red. Read names starting with “G” are 454 data while reads starting with “USI” are Illumina data. Uppercase nucleotides in white are of higher quality than those in lowercase and grey. Base at 3.926 kb was changed from an ‘*’ to a ‘t’.

The mononucleotide run in Figure 3 is composed of six T nucleotides. At the position at the left of the run (highlighted in red in the consensus track) there is disagreement between 454 and Illumina reads. The 454 reads (which have labels beginning with ‘G’) mark this postion as an ‘*’ which represents no base. The Illumina reads (which have labels beginning with ‘USI’), however, mark this position as a T. Due to the problems with 454 sequecing around

mononucleotide runs, the consensus was corrected to represent the data from the Illumina reads: a ‘t’ was placed at this position (this ‘t’ is lowercase because edited nucleotides are all made lowercase in Consed). Interestingly, all of the mononucleotide runs in DEUG4927001 were either ‘A’ or ‘T’. This might be due to the tendency of heterochromatin to have a high percentage of ‘A’ and ‘T’ nucleotides, though this might also just be a characteristic of the F element (Leung 2015). Additionally, A-T-rich regions have lower melting points than G-C-rich regions; this might skew the sequencing data to report more ‘A’ and ‘T’-rich regions than ‘G’ and ‘C’- rich regions.

Low Depth of Coverage Regions

In addition, all regions with low depth of coverage (a read depth between 10 and 40) were scanned for mononucleotide runs with discrepancies (Table 2S). These regions had to be scanned and examined separately from the high quality discrepancies, since discrepancies in these regions might not be flagged by Consed. This is because in low depth of coverage regions there simply might not be enough reads for Consed to mark discrepencies. A comprehensive table of all edits to the consensus sequence can be seen at Table 1S in the supplementary material section.

Low Consensus Quality

Next, the three regions of low consensus quality in the initial assembly were examined (Table 1). A base in the consensus is defined as being of low quality if it has a Phred score below 25 or a Phred score of 98. This score indicates the base has been changed manually. Only one

base with a score below 25 was changed. This was at base 237, which was changed from ‘C’ to ‘t’ (Table 1S).

Low consensus quality (<=25 or 98)

Contig Name	Read Name	Consensus Positions	
KB464927:0-100000 (consensus)		235-239	base quality below threshold
KB464927:0-100000 (consensus)		1845-1847	base quality below threshold
KB464927:0-100000 (consensus)		27051-27053	base quality below threshold

Table 1: List of low consensus quality regions in DEUG4927001 initial assembly.

Misalignment at Beginning of Contig

The first 5 kb of DEUG4297001 had a large number of reads that disagreed significantly with the consensus sequence (Figure 4). Since the consensus sequence is generated based on agreement among reads for each base, it was odd that a significant portion of reads disagreed with the consensus in this region. This disagreement can best be explained by misalignment. DEUG4927001 is the beginning of the genome map for the dot chromosome because its start did not align well, and therefore, could not be aligned to other regions of the chromosome; the alignment software could not find reads that match to its beginning. The misaligned reads could be from elsewhere in the genome. The depth of coverage for the first 5 kb of the contig is far greater than that of the rest of the contig, indicating that a number of reads were erroneously assigned to this region. Since the depth of coverage was adequate and high quality discrepancies could still be analyzed by counting types of bases on the correctly placed reads, it was not necessary to extract these misaligned reads (Figures 5 & 6).

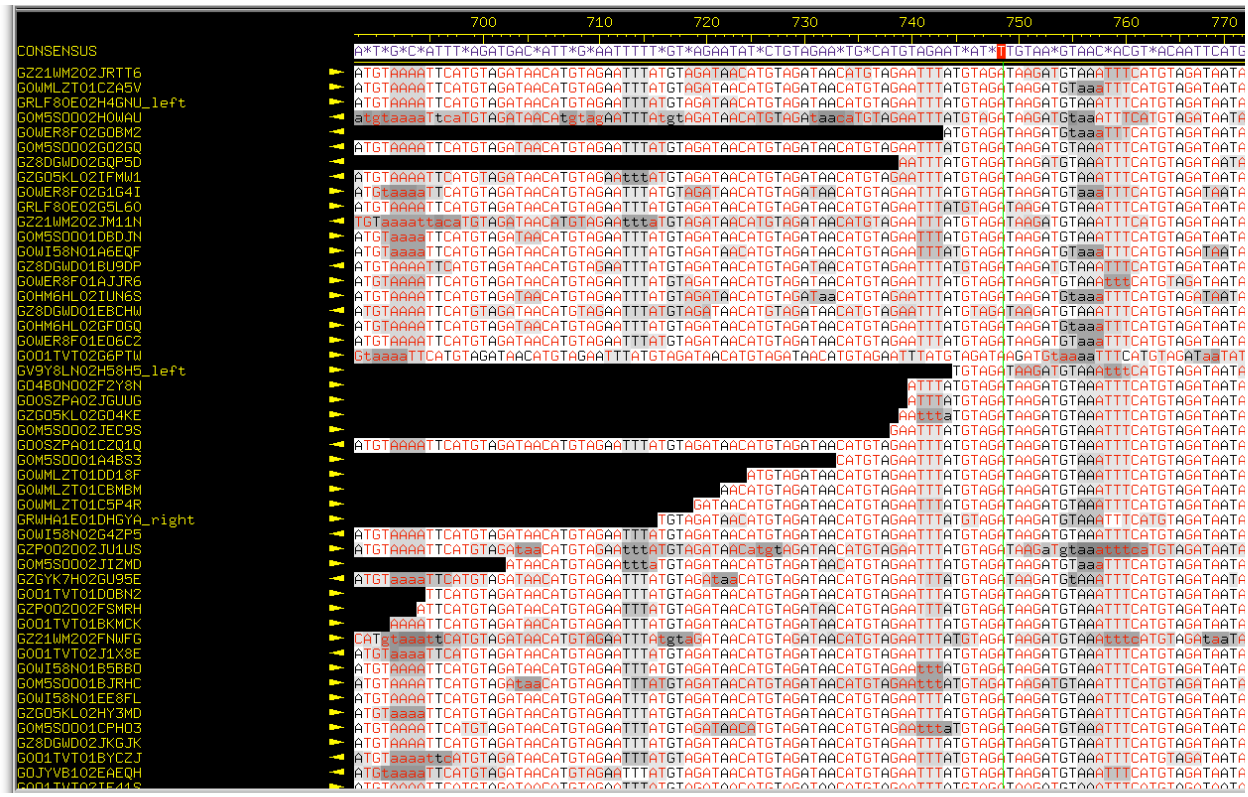

Figure 4: Misaligned reads at 0.7 kb. Red nucleotides represent disagreement with the consensus.

Figure 5: Further down the list at 0.7 kb lie the correct reads for this region.

Figure 6: Farther down the list of reads than Figure 5. The correctly assigned reads significantly outnumber the misaligned reads here.

Single Nucleotide Polymorphisms

While 91 of the high quality discrepancies were due to 454 read inaccuracies at mononucleotide runs, many of the discrepancies were not near these regions. Instead, 45 of these discrepancies had an approximately 50:50 ratio of two different possible nucleotides. This is consistent with these discrepancies being polymorphisms. An example of such a potential polymorphism is shown in Figures 7 and 8.

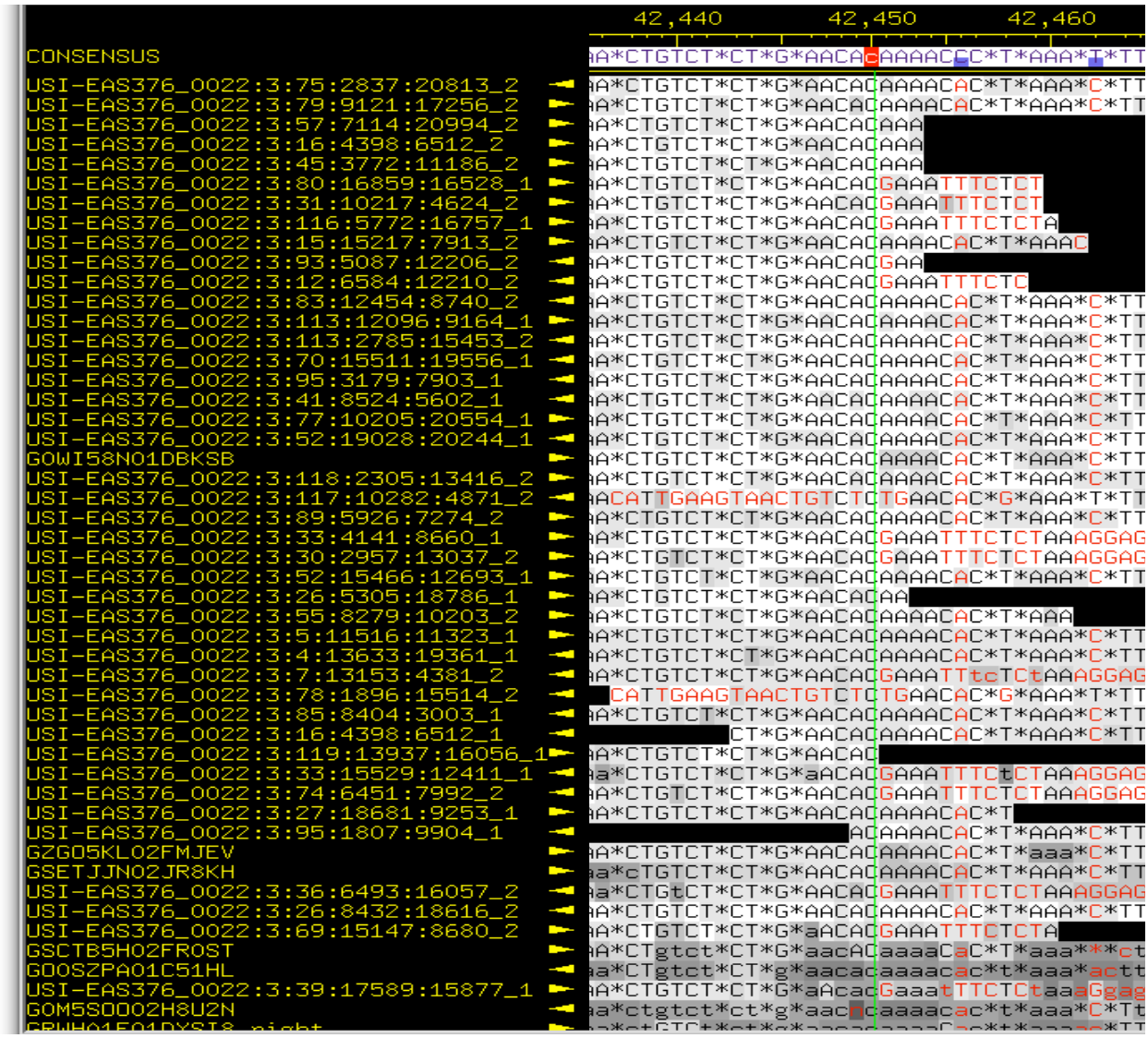

Figure 7: Position 42,450 exhibits a T-C SNP. The first half of the reads show C (highlighted by the green line).

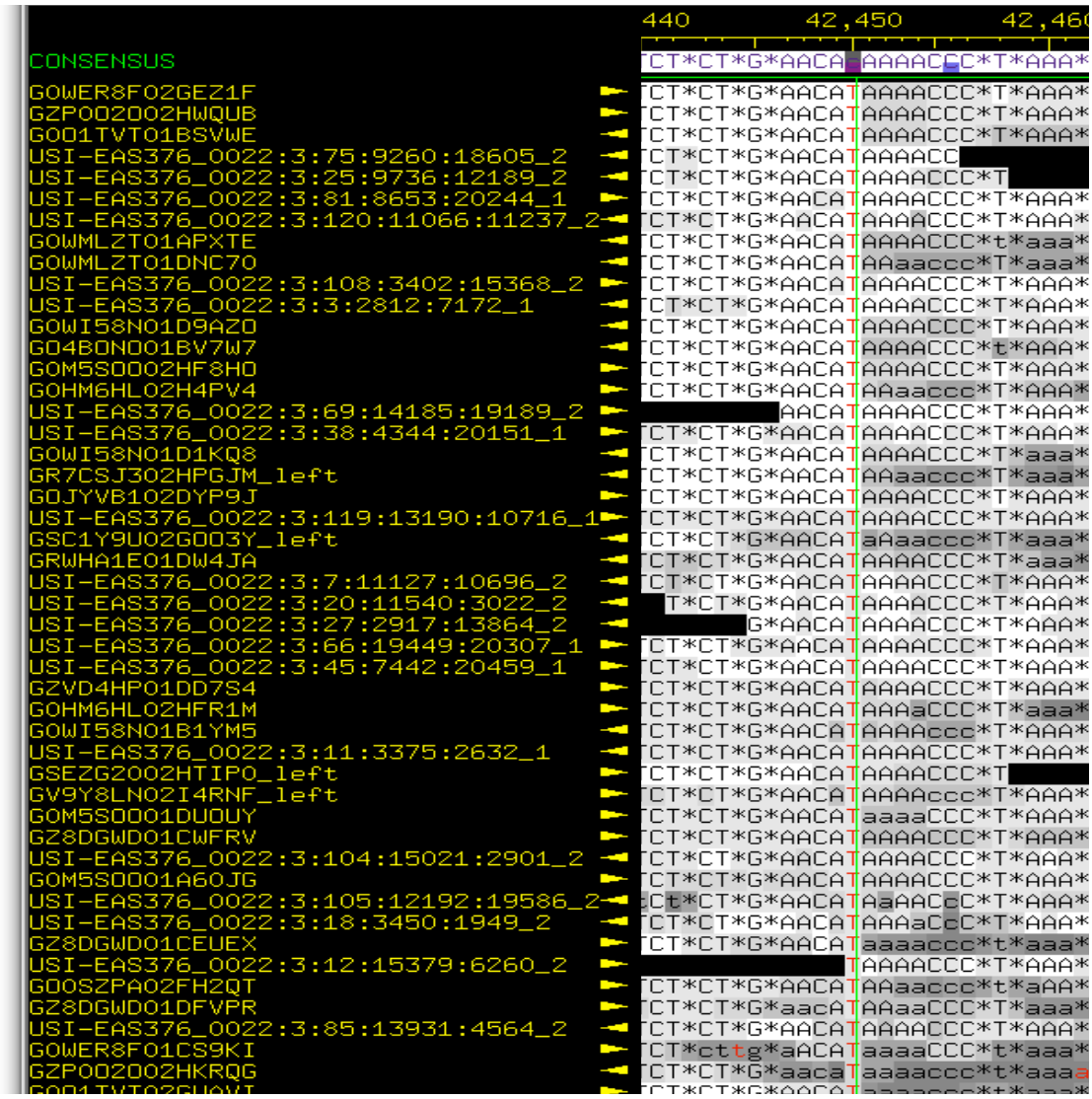

Figure 8: Position 42,450 exhibits a T-C SNP. The second half of the reads show T.

Most potential SNPs in DEUG4927001 lie between 2.7 and 73 kb or around 42 kb. No polymorphisms were called in repetitive regions. Regions that exhibit a 50:50 ratio of two nucleotides in repeats are not necessarily single nucleotide polymorphisms (SNPs), given that they are associated with a repetitive element and not a coding region of the genome. In consulting with peers, it appears that DEUG4927001 has a much greater number of SNPs than the other *D. eugracilis* contigs examined; DEUG4927001 has 45 potential SNPs while most other contigs appear to have ~0-4 SNPs. This characteristic makes this region of the

chromosome particularly interesting. A comprehensive list of all likely polymorphisms can be found in the supplemental material at Table 3S.

Repetitive Regions

Another interesting characteristic of the DEUG4927001 contig is the repetitive elements. Eight repetitive regions were identified by Consed, with three of these elements identified as Helitrons (Table 2). Helitrons are a type of transposable element that proliferates through a rolling circle mechanism. It is possible that these repetitive elements are present in this region because it is the start of the dot chromosome genome map. While it is unlikely that these repetitive regions serve any specific genomic function, Helitrons have been known to be domesticated by the genome of *D. miranda* and used for dosage compensation. It is therefore possible that these elements in DEUG4927001 are part of the regulatory pathways for the dot chromosome and help the chromosome stay heterochromatic.

repeat tags

Contig Name	Read Name	Consensus Positions	Tag Length	Comment
KB464927:0-100000 (consensus)		86010-86097		rnd-1_family-0#Unknown
KB464927:0-100000 (consensus)		86574-86636		rnd-3_family-30#RC/Helitron
KB464927:0-100000 (consensus)		86696-86819		rnd-1_family-4#Unknown
KB464927:0-100000 (consensus)		86824-86937		rnd-3_family-30#RC/Helitron
KB464927:0-100000 (consensus)		88577-88700		rnd-4_family-237#Unknown
KB464927:0-100000 (consensus)		92544-93111		rnd-3_family-30#RC/Helitron
KB464927:0-100000 (consensus)		93121-93365		rnd-4_family-237#Unknown
KB464927:0-100000 (consensus)		94696-94781		rnd-4_family-237#Unknown

Table 2: A list of repetitive elements and their positions in DEUG4927001.

Final Assembly and Conclusion

Figure 9 shows the final assembly for DEUG4927001. Since there were no gaps to correct, the final assembly looks nearly identical to the initial assembly. The purple and green rectangles at the bottom of the figure represent edited and tagged regions. In total, 40 regions were edited and 45 potential polymorphisms were identified. All discrepancies at the 29 low

coverage regions were resolved, so primer development was not needed. The high frequency of polymorphisms in DEUG4927001 compared to the other *D. eugracilis* dot chromosome contigs combined with the significant number of repetitive elements makes DEUG4927001 an interesting region for further study into the regulation of genes on the heterochromatic dot chromosome.

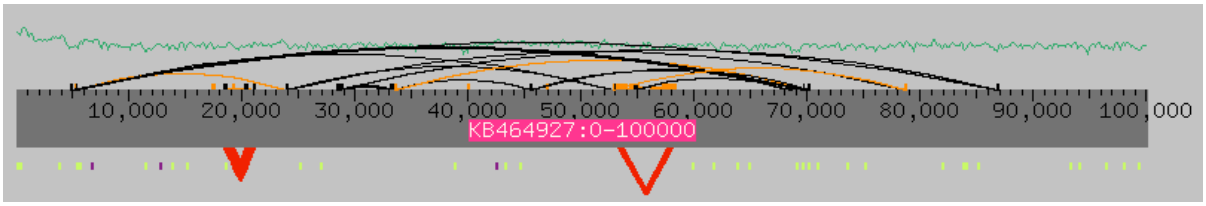

Figure 9: Final assembly for DEUG4927001. Green and purple rectangles represent edited and tagged regions.

Reference

Leung W, Elgin SRC, Shaffer CD, *et al.* (2015) *Drosophila* Muller F elements maintain a distinct set of genomic properties over 40 million years of evolution. G3: Genes, Genomes, Genetics 5(5): 719-740. doi:10.1534/g3.114.015966

Acknowledgements

Thank you to Dr. Sarah Elgin, Dr. Christopher Shaffer, Wilson Leung, and Lee Trani for their guidance and expertise throughout this finishing project. Additionally, thank you to Washington University in St. Louis and the Genomics Education Partnership for providing the facility and funds for this amazing opportunity.

Supplemental Figures

Contig Name	Read Name	Consensus Position	Change Made
KB464927:0-100000 (consensus)		237	Change C to t--low consensus quality
KB464927:0-100000 (consensus)		297	Change C to t
KB464927:0-100000 (consensus)		3926	Change * to t--MNR
KB464927:0-100000 (consensus)		5497	Change * to t--MNR
KB464927:0-100000 (consensus)		5540	Change * to a--MNR
KB464927:0-100000 (consensus)		6607	mistakenly edited--changed back to original
KB464927:0-100000 (consensus)		6631	mistakenly edited--changed back to consensus
KB464927:0-100000 (consensus)		11442	Change * to a--MNR
KB464927:0-100000 (consensus)		12751	Change T to a--MNR
KB464927:0-100000 (consensus)		12752	Change A to t--MNR
KB464927:0-100000 (consensus)		13698	Change * to a--MNR
KB464927:0-100000 (consensus)		15017	Change * to a--MNR
KB464927:0-100000 (consensus)		18462	Change * to t--MNR
KB464927:0-100000 (consensus)		25010	Change * to t--MNR
KB464927:0-100000 (consensus)		27061	Change A to t--MNR
KB464927:0-100000 (consensus)		38942	Change * to a--MNR
KB464927:0-100000 (consensus)		42405	mistakenly edited--changed back to original
KB464927:0-100000 (consensus)		42450	mistakenly edited--changed back to original
KB464927:0-100000 (consensus)		42568	Change * to a--MNR
KB464927:0-100000 (consensus)		43183	Change * to a--MNR
KB464927:0-100000 (consensus)		44550	mistakenly edited--changed back to original
KB464927:0-100000 (consensus)		59913	Change * to t--MNR
KB464927:0-100000 (consensus)		61749	Change * to t--MNR
KB464927:0-100000 (consensus)		63759	Change * to a--MNR
KB464927:0-100000 (consensus)		64865	Change * to t--MNR
KB464927:0-100000 (consensus)		69095	Change * to t--MNR
KB464927:0-100000 (consensus)		69510	Change * to a--MNR
KB464927:0-100000 (consensus)		70262	Change * to a--MNR
KB464927:0-100000 (consensus)		70833	Change C to t--MNR
KB464927:0-100000 (consensus)		73475	Change * to t--MNR
KB464927:0-100000 (consensus)		75215	Change * to t--MNR
KB464927:0-100000 (consensus)		82064	Change * to a--MNR
KB464927:0-100000 (consensus)		83942	Change * to a--MNR
KB464927:0-100000 (consensus)		84003	Change * to a--MNR
KB464927:0-100000 (consensus)		85030	Change * to a--MNR
KB464927:0-100000 (consensus)		93363	Change * to T--MNR
KB464927:0-100000 (consensus)		94002	Change * to a--MNR
KB464927:0-100000 (consensus)		96422	Change * to a--MNR
KB464927:0-100000 (consensus)		98140	Change * to t--MNR
KB464927:0-100000 (consensus)		99501	Change * to a--MNR

Table 1S: Changes made to consensus. In total, 40 bases were changed. The acronym 'MNR' stands for 'MonoNucleotide Run'.

Contig Name	Read Name	Consensus Position	Read Depth
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			
KB464927:0-1000			

Table 2S: List of areas of low coverage. Range of bps with low coverage shown in second column and corresponding read depth shown in third column.

Contig Name	Read Name	Consensus Position	Type of Polymorphism
KB464927:0-100000 (consensus)		293	polymorphism for T & G--agreed wI consensus
KB464927:0-100000 (consensus)		371	polymorphism for A & G--agreed wI consensus
KB464927:0-100000 (consensus)		748	polymorphism for T & G--agreed wI consensus
KB464927:0-100000 (consensus)		798	polymorphism for G & T--agreed with consensus
KB464927:0-100000 (consensus)		2745	polymorphism A & T--agree wI consensus
KB464927:0-100000 (consensus)		2746	agree but could be a T polymorphism
KB464927:0-100000 (consensus)		2787	polymorphism--A & G--agre wI consensus
KB464927:0-100000 (consensus)		3115	A-T polymorphism--agree wI consensus
KB464927:0-100000 (consensus)		3145	A-C polymorphism--agree wI consensus
KB464927:0-100000 (consensus)		3371	Polymorphism--but also a MNR for the T
KB464927:0-100000 (consensus)		3447	polymorphism T & A--agree wI consensus
KB464927:0-100000 (consensus)		3462	A G polymorphism--agree wI consensus
KB464927:0-100000 (consensus)		3863	A C polymorphism--agree wI consensus
KB464927:0-100000 (consensus)		3878	polymorphism--agree wI consensus
KB464927:0-100000 (consensus)		3886	polymorphism-agree wI consensus
KB464927:0-100000 (consensus)		4019	A T polymorphism--agree wI consensus
KB464927:0-100000 (consensus)		4057	polymorphism for G & A --agree wI consensus
KB464927:0-100000 (consensus)		4164	polymorphism--T/C--agree wI consensus
KB464927:0-100000 (consensus)		4167	polymorphism-- G/T--agree wI consensus
KB464927:0-100000 (consensus)		4180	polymorphism A/G --agree wI consensus
KB464927:0-100000 (consensus)		4181	A/G polymorphism--agree wI consensus
KB464927:0-100000 (consensus)		4214	C/G polymorphism--agree wI consensus
KB464927:0-100000 (consensus)		4302	C/t polymorphism--agree wI consensus
KB464927:0-100000 (consensus)		4319	G/T polymorphism--agree wI consensus
KB464927:0-100000 (consensus)		4361	A/G polymorphism--agree wI consensus
KB464927:0-100000 (consensus)		5650	G/T polymorphism--agree wI consensus
KB464927:0-100000 (consensus)		5854	G/A polymorphism--agree wI consensus
KB464927:0-100000 (consensus)		6607	A/T--polymorphism--changed A to T bc T had higher quality reads
KB464927:0-100000 (consensus)		6631	G/T polymorphism--agree wI consensus
KB464927:0-100000 (consensus)		6631	polymorphism--change from T to G
KB464927:0-100000 (consensus)		7361	agree wI consensus--polymorphism
KB464927:0-100000 (consensus)		42391	agree wI consensus--probably a polymorphism g or a
KB464927:0-100000 (consensus)		42405	change from t to c--polymorphism
KB464927:0-100000 (consensus)		42415	agree wI consensus--polymorphism for T
KB464927:0-100000 (consensus)		42450	change to c--polymorphism C /T
KB464927:0-100000 (consensus)		42450	C/T polymorphism
KB464927:0-100000 (consensus)		42456	agree wI consensus--cpolymorphism
KB464927:0-100000 (consensus)		42462	agree wI consensus--polymorphism for C
KB464927:0-100000 (consensus)		42473	agree wI consensus--polymorphism for G
KB464927:0-100000 (consensus)		42500	agree wI consensus--polymorphism for A
KB464927:0-100000 (consensus)		42525	agree wI consensus--cpolymorphism for T
KB464927:0-100000 (consensus)		42528	agree wI consensus--polymorphism for C
KB464927:0-100000 (consensus)		42529	agree wI consensus--polymorphism for C
KB464927:0-100000 (consensus)		42546	agree wI consensus--polymorphism for T
KB464927:0-100000 (consensus)		42547	agree wI consensus-polymorphism for G

Table 3S: List of likely SNPs in DEUG4027001. There are 45 total.