THE PENNSYLVANIA STATE UNIVERSITY
MILLENNIUM SCHOLARS PROGRAM

DEPARTMENT OF Biology

GENOMIC ANALYSIS OF THE KLAMATH GENE ARRANGEMENT OF DROSOPHILA PERSIMILIS

AMIRA ELLISON
SPRING 2022

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree in Biology

Reviewed and approved* by the following:

Stephen Schaeffer
Professor of Biology
Thesis Supervisor

Benoît Dayrat
Associate Professor of Biology
Reviewer

* Signatures are on file in the Millennium Scholars Program office.

ABSTRACT

Chromosome inversions are characterized as structural mutations that result from the reversal of gene order within a chromosomal segment through two breaks that reattach. *Drosophila pseudoobscura* and its sibling species *D. persimilis* have been a model system for the study of inversion polymorphisms in nature. These inversions have been consistently studied for decades, with a specific focus on how these inversions emerge in drosophila populations. Two hypotheses have been proposed for how inversions arise at the molecular level, the repeat-mediated and staggered cuts models. Previous work in *D. pseudoobscura* has supported the repeat-mediated mechanism, but it is unclear if similar mechanisms are generating rearrangements in *D. persimilis*. This study used long read sequencing technology to determine the genomic sequence of the breakpoints of the Klamath gene arrangement on the third chromosome in *D. persimilis.* Through use of a genome assembly and breakpoint analysis, the proximal and distal breakpoints of the Klamath arrangement were identified and compared. The analysis of the breakpoint sequences of the Klamath arrangement supports the repeat-mediated hypothesis of inversion origin.

TABLE OF CONTENTS

LIST OF FIGURES iii

LIST OF TABLES iv

ACKNOWLEDGEMENTS v

CHAPTER 1 INTRODUCTION……1

CHAPTER 2 MATERIALS AND METHODS 7

Drosophila Strain 7

Cytogenetic Analysis..... 7

DNA Isolation 8

Long Read Sequencing..... 9

Whole Genome Assembly using the Galaxy Platform and Excel..... 9

MegaBLAST Analysis & Cytogenetic Map 11

Contig Assignment……12

Breakpoint Sequence Analysis……12

Data Availability……13

CHAPTER 3 RESULTS 14

Cytogenetic Analysis..... 14

Long Read Data Runs and Statistics 15

Assembly Data..... 15

Breakpoint Mapping..... 18

Breakpoint Sequence Comparison 18

CHAPTER 4 DISCUSSION & CONCLUSION..... 24

BIBLIOGRAPHY 27

LIST OF FIGURES

Figure 1. Paracentric and Pericentric Inversion Diagram

Figure 2. Chromosome Assembly Bioinformatics Workflow Diagram

Figure 3. Klamath Gene Arrangement Cytogenetic diagram

Figure 4. Arrowhead, Standard, and Klamath Chromosome Maps

Figure 5. Arrowhead and Klamath Inversion Diagram

Figure 6. Proximal Standard-Arrowhead vs. Distal Standard-Klamath Graph

Figure 7. Distal Standard-Arrowhead vs. Distal Standard-Klamath Graph

Figure 8. Distal Standard-Arrowhead vs. Distal Standard-Klamath Graph

Figure 9. Proximal Standard-Arrowhead vs. Proximal Standard-Klamath Graph

LIST OF TABLES

Table 1. Oxford Nanopore Long Read Sequencing

Table 2. Assembly Chromosome Length vs. Reference Chromosome

Length Table 3. Chromosome Transcript Matches

ACKNOWLEDGEMENTS

I would like to thank Dr. Stephen Schaeffer for advising me and providing me with the resources needed to complete my research thesis. Specifically, I would like to thank him for assisting me with my data collection, providing the cytogenetic analysis image, and guiding me in my authorship of my thesis. Additionally, I would like to thank Benoît Dayrat for reviewing my thesis. This work was supported by funds from The Pennsylvania State University to Stephen W. Schaeffer.

Chapter 1

Introduction

Inversions are characterized as structural mutations that result from the reversal of gene order within a chromosomal segment through two breaks that reattach[1]. In Drosophila, they have been shown to be mediated by ectopic exchange between repeats or through staggered cuts in the breakpoint regions.[2,3] Inversions can be categorized as one of two types: paracentric inversions and pericentric inversions[1]. A paracentric inversion is described as an inversion that is created when a chromosome breaks and rejoins on the same side of the centromere[1]. In contrast, a pericentric inversion is the result of a chromosomal break including the centromere[1], which is often detected with trypsin Giemsa banding[4]. Consequently, the effects of a paracentric inversion differs from that of a pericentric break, as it does not change the chromosome arm ratios (Figure 1), thus making detection harder using tradition mitotic cytological techniques[5].

Paracentric inversions are often detected in humans when a patient presents with a disease[5]. As for pericentric inversions, they have been found in the general population at a level of 1-2%.[6] In most cases, paracentric and pericentric inversions are not considered detrimental, but they can prove to be an issue during meiosis if an individual is heterozygous and crossing over occurs with the inverted region[6]. Consequently, this results in the formation of gametes with deletions or duplications that lead to deleterious phenotypes and are lost[6].

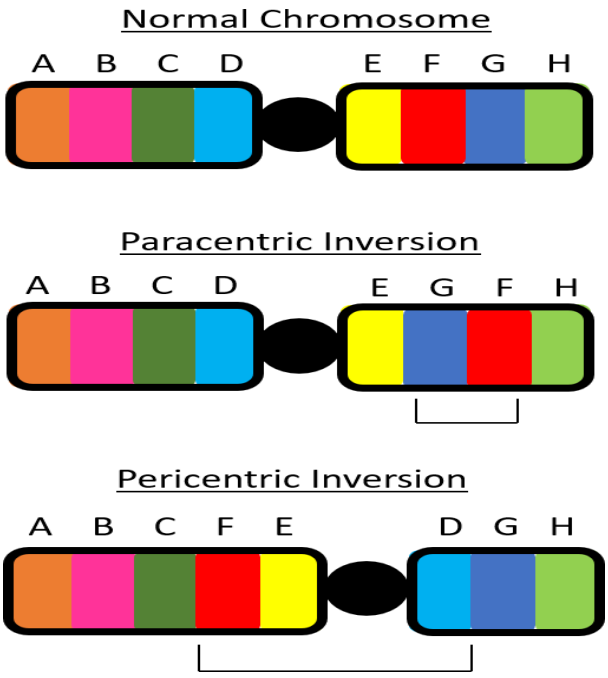


Figure 1. Diagram depicting the nature of pericentric and paracentric inversions. Colors and letters represent chromosome segments and black brackets represent inversions. The pericentric diagram shows an alternation in chromosome arm lengths and gene order after the inversion, while the paracentric inversion only shows an alteration in gene order.

Some genetic mechanisms have been identified as being responsible for inversions. Chromosomal inversions can be prompted by non-allelic homologous recombination between inverted repeats or fork stalling, template switching and non-homologous end joining, as well as mediated and staggered cut mechanisms[7]. Furthermore, there has been some evidence that chromosomal inversions have consequences as some diseases can be caused by inversions, mainly because the inversion caused the disruption of one gene.[7] These inversions can cause the disruption of coding sequences in genes or modify how genes adjacent to the breakpoints are expressed.[7]

While there is some understanding of the molecular mechanisms for the origin of inversions, there is still not a general consensus for the general evolutionary mechanisms that establish and maintain inversions within populations. Proposed hypotheses for how inversions are established in populations include random genetic drift and mutation[8], genetic hitchhiking with advantageous alleles[9], direct effects associated with changes at inversion breakpoints, and indirect effects associated with recombination suppression[10]. Indirect effects include inversions that capture a segment with low numbers of deleterious mutations[10,11] and they also include inversions that capture sets of beneficial mutations and/or adaptive alleles involved in local adaptation.[12,13]

While it has been documented that inversions are present in different species, inversions were originally identified in Drosophila. The cells of the salivary glands of Drosophila have polytene chromosomes that visually display inversions once magnified[14]. Variation in gene arrangements can be effectively observed in the salivary glands' chromosomes. Through cytogenetic analysis, homozygote and heterozygote inversions can be identified as there is a distinctive linear arrangement of chromosomal bands and puffs, which are used to label sections of salivary gland chromosomes[15]. Because of these sections, the convention in Drosophila is to divide salivary chromosomes into 100 sections numbered 1-100. In *D. pseudoobscura*, the left arm of the X chromosome (XL) is 1-17 while the right arm of the X chromosome (XR) is 18-42[16]. Moreover, the second chromosome spans from section 43 to section 62, the 3rd chromosome spans from section 63 to section 81, the fourth chromosome spans from section 82 to section 99 and, the 5th chromosome is considered a dot chromosome as it is composed of only one section, which is section 100[6]. This visual representation of their chromosomes made Drosophila flies ideal to examine the nature of chromosomal inversions.

To evaluate how inversions are generated, the relationship between *D. pseudoobscura* and *D. persimilis* gene arrangements can be used as an evaluation system. The variation in gene arrangements in *D. pseudoobscura* and *D. persimilis* is due to inversions of chromosome segments in the course of evolution as represented on a phylogeny[16,17]. These inversions can have physiological and anatomical consequences as it has been noted that these inversions in Drosophila can influence their appearance as well as their maturation rate[16]. Specifically, when comparing *D. persimilis* and *D. pseudoobscura* chromosomes, there have been additional observations in their differences. *D. pseudoobscura* has "J" shaped Y chromosomes in contrast to the "V" shaped Y chromosomes of *D. persimilis*[16]. Furthermore, *D. pseudoobscura* releases more eggs and has higher longevity in the absence of food when compared to its D. *persimilis* counterpart[16]. Also *D. pseudoobscura* has a higher number of sex combs, and has larger wings.[16] Finally, *D. pseudoobscura* has a greater number of wing beats per unit during flight[16].

On a molecular scale, there are key differences between *D. pseudoobscura* and *D. persimilis*. The left limb of the X chromosomes between the two species differs by a single inversion, as does the second chromosome[16]. The XL and 2nd chromosomes diverge by fixed paracentric inversions spanning five to six major cytological sections[20]. The XR and 3rd chromosomes are polymorphic for inversions. XR is segregating for distinct Sex ratio arrangements in both species with a single inversion existing between Standard and Sex ratio in *D. persimilis* and three non-overlapping inversions distinguishing the Standard and Sex ratio chromosomes in *D. pseudoobscura*. The 3rd chromosome is segregating for numerous arrangements in both species with a single chromosomal type (Standard) being shared between the two species[16]. Lastly, the 4th and 5th chromosomes have the same gene arrangement in *D. pseudoobscura* and *D. persimilis*.[16]

Among the five chromosomes, the 3rd chromosome is an ideal chromosome to investigate how chromosomal inversions are generated and maintained in populations as it is polymorphic for over 50 different gene arrangements[21]. There is some indication that inversions may be adaptive, as evolutionary forces can potentially determine if inversions are fixed or eliminated in populations and can also determine their geographical displacement[7]. Frequencies of arrangements in *D. pseudoobscura* align with major climatic and geographic differences[22]. Furthermore, seasonal cycling of arrangements and altitudinal gradients of inversion frequencies, further support the idea that inversions can be adaptive[22].

In total, there has been thirty-nine identified gene arrangements in the 3rd chromosome of *D. pseudoobscura* and eighteen identified arrangements in *D. persimilis*[21]. The two species have one arrangement in common designated as the "Standard". All other arrangements were named for the locality where the chromosome was first discovered[15]. The 3rd chromosome has multiple variations, therefore there was not a strong consensus for which chromosome was the ancestral arrangement[20]. The choice of a Standard arrangement in the 3rd chromosome was arbitrary[16]. The fifty gene arrangements in *D. pseudoobscura* and *D. persimilis* can easily be derived from one another through single inversion events with the exception of the transition of the *D. pseudoobscura* Standard to the Santa Cruz arrangement that requires two inversion steps through a Hypothetical arrangement that has never been discovered in nature[15,21]. Phylogenetic analyses and gene adjacency information have shown that the Hypothetical arrangement is the inferred ancestral arrangement[28,29,30].

Two particular gene arrangements in the 3rd chromosome that derive from the Standard are Klamath from *D. persimilis* and Arrowhead from *D. pseudoobscura*[16]. The Arrowhead gene arrangement is derived from the Standard arrangement, and it diverged from the Standard

arrangement due to one inversion[16]. The inversion happens between 70A and 70B at the proximal end and 76B and 76C at the distal end[16]. The Klamath arrangement is derived from the Standard as it differs by one inversion with breakpoints in the proximal part between 70C and 70D and between 77A and 77B at the distal end[16]. To gain a better understanding of the mechanisms that generate these inversions on the 3rd chromosome, this study mapped the breakpoints of the *D. persimilis* Klamath gene arrangement and compared the breakpoint region sequences to the *D. pseudoobscura* Arrowhead gene arrangement to test whether the repeat mediated or staggered cut model for the origin of inversions is responsible for the Klamath inversion in *D. persimilis*.

Chapter 2

Materials and Methods

Drosophila Strain

The *D. persimilis* strain used in this study is named Dper_KL_SN104-2. This strain was collected in 2017 in Mount Hood, Oregon (Latitude: 45.395598° N, Longitude: 121.561934° W) by Ryan Bracewell and Doris Bachtrog of the University of California at Berkeley. The strain had been maintained in small culture for three years prior to DNA isolation for sequencing. The gene arrangement of this strain was checked using squashes of third instar larval salivary glands (see cytogenetics below). This strain was identified as *D. persimilis* based on the karyotype of the left arm of the X chromosome and found to carry the Klamath gene arrangement on the 3rd chromosome.

Cytogenetic Analysis

Salivary gland squashes from *D. persimilis* third instar larvae were prepared to verify the gene arrangement of the strain used for long read sequencing. Third instar larvae were collected from strains after two weeks of egg laying and larval development. Third instar larvae are characterized by wandering where the larvae crawl out of the culture medium on the sides of the vial. Larvae are washed in Drosophila Ringers (0.128 M NaCl, 5mM KCl, 1.5 mM $CaCl_2$) solution for 5 minutes to remove any residual culture medium. Salivary glands were dissected from the larvae in 45 % acetic acid within a concave depression slide and transferred to a slide with lacto-aceto-orcein stain. A cover slip was added to the slide and a dissection probe was

used to disrupt the salivary glands with gentle circles on the cover slip. The slide was placed within absorbent paper and a wooden dowel was rolled over the location of the cover slip being careful not to allow the cover slip to move side to side. The slide was examined under an Olympus BH-2 binocular microscope at 1000 x magnification and oil immersion. Digital images were obtained with CellSens Entry version 1.11 software using an Olympus SC-30 camera attached to the microscope.

DNA Isolation

High molecular weight genomic DNA was isolated using the Oxford Nanopore protocol. Two batches of 0.15 g of adult flies were homogenized in 10 ml of nuclear isolation buffer[31]. The homogenate was filtered through 200 μm nylon mesh to remove fly body parts. The homogenate was centrifuged at 3500 xg for 15 minutes at 4° C. Nuclei were resuspended in G2 buffer from the Qiagen Genomic DNA Buffer set (Catalog No. 19060) along with Proteinase K. The nuclei were incubated at 50° C for 45 minutes with mixing of the tube every 5 minutes to ensure that nuclei were lysed. High molecular weight DNA was purified using Qiagen Genomic-tip columns 100/g (Catalog No. 10243) using the Qiagen Genomic DNA Buffer set (Catalog No. 19060). DNA was quantified using a Nanovue Spectrophotometer and the quality of DNA was checked on a 0.6% agarose gel with 1x TBE buffer (0.089 M Tris; 0.089 M Boric Acid; 0.002 M EDTA).

Long Read Sequencing

Oxford Nanopore sequencing determines the order of nucleotides in DNA by passing the DNA through a nanopore. As the DNA passes through the pore, the bases are read based on interactions with the pore. A total of 400 ng of high molecular weight DNA was used in the Oxford Nanopore sequencing using the Rapid Sequencing Kit (Catalog No. SQK-RAD004). The kit uses a transposome complex to randomly cleave DNA and add adaptors to the ends of sequences. These adaptors are necessary to guide the DNA to the nanopore for sequencing. DNA with adapters were sequenced on a MinION device with a SpotOn Flow Cell according to the manufacturer's recommendations. The MinION device was attached to a MinIT device, which allows collection of data without dedicating a lap top computer for data acquisition. The MinIT device can store up to 450 gigabytes worth of data, but a typical long read run will generate 100 gigabytes of data. Once the sequencing run is complete, data is offloaded for downstream processing.

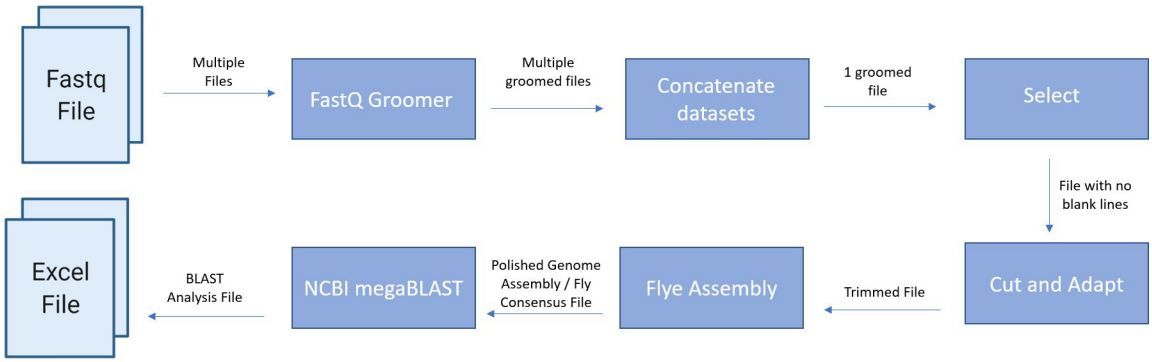
Whole Genome Assembly using the Galaxy Platform and Excel

Five Klamath FASTQ files generated during Oxford Nanopore long read sequencing runs were imported and groomed using the FASTQ Groomer within the Galaxy genome analysis platform (https://usegalaxy.eu) (Figure 2). To compile the datasets into one assembly, the FASTQ files from the five runs were concatenated using the "Concatenate datasets" tool. Blank lines between reads of each run in the concatenated data set were removed using the Select tool with the parameters set to "Not Matching" and the pattern used was ^$ denoting blank lines. The file generated from the Select tool was used as the input for the "Cut and Adapt" tool two times, to remove the adaptor sequences from the first and last 150 base pairs of the long read. The output from the second "Cut and Adapt" tool was put into Flye Assembly to generate a data consensus file[23,24]. One must specify the expected genome size and number of polishing iterations in the

Flye Assembler tool (Genome Size = 160 Mb and Polishing Iterations =3). Oxford nanopore sequencing runs incorporate errors at a rate of 12%. Flye uses the overlapping sequence reads to remove sequencing errors through polishing steps. Sequencing errors are not expected to occur at the same nucleotide, so the program uses the consensus sequence of overlapping reads to remove the errors. The quality of the assembly was assessed based on the N50, which is defined as the length of the contig that represents the midway point of the rank ordered contigs based on the total assembly length, in this case 80 Mb. In addition, the Benchmarking Universal Single Copy Orthologs (BUSCO) tool[26] was used to assess the quality of the assembly by scanning for a fundamental set of Dipteran protein coding sequences. Quality scores are between 0 and 100% detection of the core set of proteins.

Once the Flye data consensus file was produced, it was used as the input for a NCBI megaBLAST that was set to compare the data consensus file to the imported *D. pseudoobscura* reference mRNAs of the UCI_Dpse_MV25_SWS genome assembly at GenBank (Biosample: SAMN13616452, BioProject PRJNA622252, Assembly: GCA_09870125.2). Once the blast analysis was complete, the resulting file was imported into Microsoft Excel for further downstream processing.

Figure 2. Bioinformatics workflow to generate chromosome assembly as described above



MegaBLAST Analysis & Cytogenetic Map

Excel's vlookup function was used to determine where each transcript mapped in the reference genome by searching the protein master file for the Arrowhead gene coordinates in the UCI_Dpse_MV25_SWS reference sequence. The vlookup command searched for the chromosome, beginning chromosome coordinate, and ending chromosome coordinate from the master protein file for each transcript. To use the best quality hits, the data was sorted by percent match and the highest percent match. The Excel spreadsheet was sorted either by the Arrowhead reference genome or the Klamath reference genome to map the ST to KL and ST to AR breakpoints in both respective genomes. Once the Excel sheet was in the appropriate arrangement, the segments between each pair of break points were assigned arbitrary colors to separate the identified inverted regions. Using the cytogenetic maps by Dobzhansky and Epling as a reference, the Excel coordinates were compared to the map and the breakpoints were aligned with the inverted regions in the map. This was done by comparing the relative sizes of the color-coded regions to the sizes of the chromosome segments within the inversion.

Contig Assignments

The pivot table function in Excel was used with the Mega Blast Analysis vlookup Excel sheet to assign assembly contigs and scaffold to *D. persimilis* chromosomes. This pivot table produced a count of BLAST hits that the contigs and scaffolds had to each chromosome. Contigs and scaffolds were assigned to each chromosome based on two criterions. Contigs and scaffolds were assigned to a chromosome if the contig/scaffold had at least a count of 20 Blast hits or higher for a chromosome and contig/scaffold counts a chromosome had to be at least 95% of the total count across all chromosomes.

Breakpoint Sequence Analysis

The intergenic sequence intervals where breakpoints map were extracted from the Arrowhead reference genome and the *D. persimilis* KL genome. The proximal and distal breakpoint sequences from the Arrowhead reference genome were compared to the *D. persimilis* genome with BLAST. The set of intervals within the proximal and distal Arrowhead breakpoint sequences was used to develop a map of sequences that match multiple regions in the *D. persimilis* genome using the Fortran program "NG Breakpoint Repeat Map" written by S. W. Schaeffer. The program creates an Adobe postscript file that is converted to a pdf with Adobe Distiller. The map is a histogram built from the matching sequence intervals that define nucleotides that match multiple regions in the genome.

Dot plots within MegAlign within the Lasergene suite of DNA analysis programs (DNAStar) were used to compare Arrowhead and Klamath breakpoint sequences to one another.

The dot plots used a window size of 30, a percentage match of 50%, and a minimum quality of 1. The dot plots were filtered to include the top 2.6 to 5.0% of the matches to reduce the noise.

Data Availability:

Data and other supporting materials are available at https://scholarssphere.psu.edu (Amira Ellison Thesis Data and Supporting Material).

Chapter 3

Results

Cytogenetic Analysis

Salivary squashes of the Schaeffer Laboratory strain of *D. persimilis* found that the strain carried the Klamath gene arrangement (Figure 3 and see Figure 5 from Moore and Taylor (1986)[33]. Cytogenetic analysis of XL verified that the strain was *D. persimilis*[25]. A total of nine independent larvae were dissected and scored for their gene arrangement. All nine larvae were homozygous for the Klamath gene arrangement.

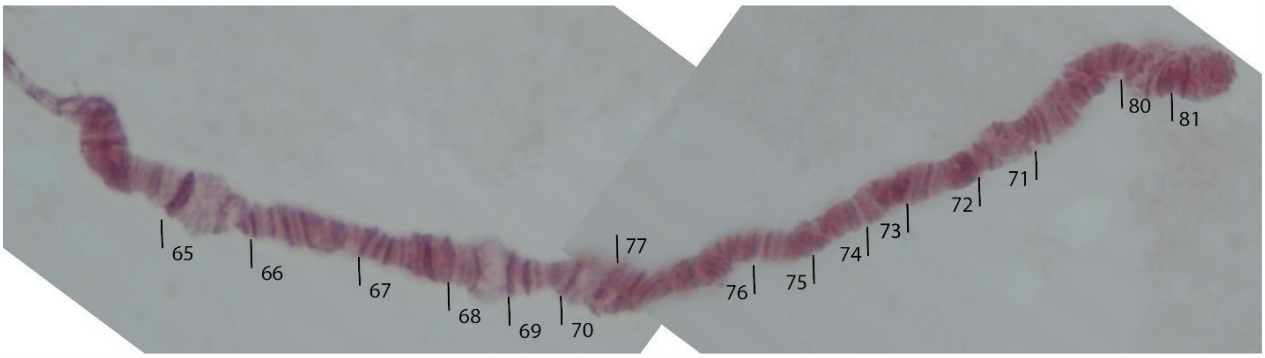


Figure 3: Cytogenetic diagram depicting chromosome sections for Klamath gene arrangement.

Long Read Data Runs and Statistics

Table 1. Statistics about the Oxford Nanopore Long Read Sequencing Run for the *D. persimilis* KL_SN104-2 genome.

Run	Date	Strain	Reads	Bases	N50	Max. Read Length	Coverage
1	2020_08_03	Dper_KL_SN104-2	631,648	2,099,569,079	13,379	153,295	13.1
2	2021_03_07	Dper_KL_SN104-2	18,800	1,695,268,481	16,015	125,332	10.6
3	2021_03_08	Dper_KL_SN104-2	84,000	817,351,729	16,475	131,623	5.1
4	2021_06_04	Dper_KL_SN104-2	568,000	4,721,151,212	14,784	139,840	29.5
5	2021_06_05	Dper_KL_SN104-2	64,000	515,177,106	14,573	112,578	3.2
Totals			**1,366,448**	**9,848,517,607**			**61.6**

Date, Date of the sequencing run; Reads, number of reads in the sequencing run; Bases, total number of bases sequenced; N50, length of the read at 50% point of the cumulative sequenced bases; Max. Read, Maximum Read Length; Cov., Coverage.

There were five sequencing runs with a total of 1,366,448 reads and 9,848,517, 607 bases. The maximum read lengths over all runs was 153,295 and the minimum per run maximum length was 112,578. The maximum reads in a run were 631,648, while the minimum reads for a run was 18,800. In terms of bases, the maximum number of bases for a run was 2,099,569,079 and the minimum for a run was 515,177,106. N50s for all sequencing runs ranged between 13,000 and 16,500 bases. The run with the maximum N50 length was run 4, while the run with the minimum N50 length was run 5. Coverage for the runs fell between 3.2x and 29.5x, with a total coverage of 61.6x across all five runs.

Assembly Data

Total contigs and scaffolds in the final Flye assembly was 710, with 8 scaffolds and 702 contigs. Additionally, the total read length generated by the Flye assembly was 173,005,162

bases and the N50 was 21,153,950. The total length of the contigs was 73,585,748 while the total length of the scaffolds was 99,419,414. The average scaffold length was 12,427,427 base pairs while the average contig length was 104,823 base pairs. The scaffold N50 was 24,585,885 base pairs long and over half of the overall scaffold length is within 2 scaffolds. The contig N50 was 3,282,263 and over half of the overall contig length was in 5 contigs. The quality of the assembly was assessed with the Benchmarking Universal Single Copy Orthologs (BUSCO) tool[26]. This tool determines if a core set of protein coding genes is present in the assembly from 0 to 100%. The dipteran gene set of 3,285 genes was used to assess the quality of the *D. persimilis* KL SN104-2 assembly. The assembly had an overall BUSCO completeness score of 88.0%. There were 2,892 complete BUSCOs and 2,861 complete/single-copy BUSCOs. Fragmented BUSCOs amounted to 244 while there were 149 missing BUSCOs. In comparison, the BUSCO scores reported in Miller *et al*. (2021) for 15 Drosophila genomes was 97%[32].

Table 2: Names, transcript matches, and contig/scaffold lengths correlated to each chromosome

Chromosome	Scaffold/Contig	Transcript Matches	Length
Chromosome 2	scaffold_1198	4734	24,585,885
	contig_871	531	3,282,263
	contig_1217	748	3,637,944
Chromosome 3	contig_813	5113	21,153,950
	contig_506	24	1,052,971
Chromosome 4	scaffold_1019	3653	22,683,725
	contig_1251	819	5,512,269
	contig_1784	152	1,130,887
Chromosome 5	contig_13	375	1,816,702
Chromosome X	scaffold_996	5333	29,783,284
	scaffold_801	3209	18,502,966
	contig_135	115	2,094,197
	contig_784	283	2,642,253
	contig_793	184	1,856,991

Chromosomes 2-5 and chromosome X were sequenced. Scaffold_1198, Contig_871, Contig_1217 were associated with chromosome 2 while contig 813 and contig_506 were

associated with chromosome 3 (Table 2). Chromosome 4 was associated with scaffold_1019, contig_1251, and contig_1784 while contig_13 was associated with chromosome 5. Lastly, chromosome X was associated with scaffold_996, scaffold_801, contig_135, contig_784, and contig_793. 5113 copies of contig_813 were associated with chromosome 3 based on the established MRNA master file, therefore it was determined that contig 813 was the primary contig associated with chromosome 3. Thus, contig 813 was further analyzed to examine the breakpoints and inversions in chromosome 3.

Table 3: Comparison of assembly chromosome lengths and reference genome (UCI_Dpse_MV25_SWS) lengths.

Chromosome	Dper KL SN104-2 Strain Assembly Chromosome Length	Reference Genome Length	Assembly Completion (%)
Chromosome 2	31,506,092	32,422,566	97.2
Chromosome 3	22,206,921	23,510,042	94.4
Chromosome 4	29,326,881	30,706,867	95.5
Chromosome 5	1,816,702	1,881,070	96.6
Chromosome X	54,879,691	68,154,638	80.5
Total Length	139,736,287	156,675,183	89.1

Compared to the established reference genome (UCI_Dpse_MV25_SWS), chromosomes 2-5 produced by the Flye Assembly were similar in size to the chromosome lengths from the reference genomes as measured as a percent of the reference chromosome length (Table 3). For chromosome 2-5, the assembly length was between 94.4% and 97.2% of the reference genome lengths. Specifically, chromosome 3 had a length equivalent to 94.4% of the reference genome's length as the assembly length was 22,206,921 and the reference length was 23,510,042. In contrast, chromosome X derived from the Flye Assembly was considerably shorter than chromosome X from the reference genome as it was only 80.5% of the reference genome's length

Breakpoint Mapping:

Based on prior analysis, contig 813 was determined to be the primary contributor to chromosome 3. Within contig 813 for the Klamath arrangement, the identified points of discontinuity were at positions 10815777, 10366469, 9406843, and 3948581. Therefore those positions were identified as the breakpoints in the Klamath arrangement (Figure 4). Within contig 813 for the Arrowhead arrangement, the points of discontinuity were determined to be position 12635778, 18110158, 18583147,19533090. Therefore, those positions were identified as the breakpoints in the Arrowhead arrangement (Figure 4).

Breakpoint Sequence Comparison

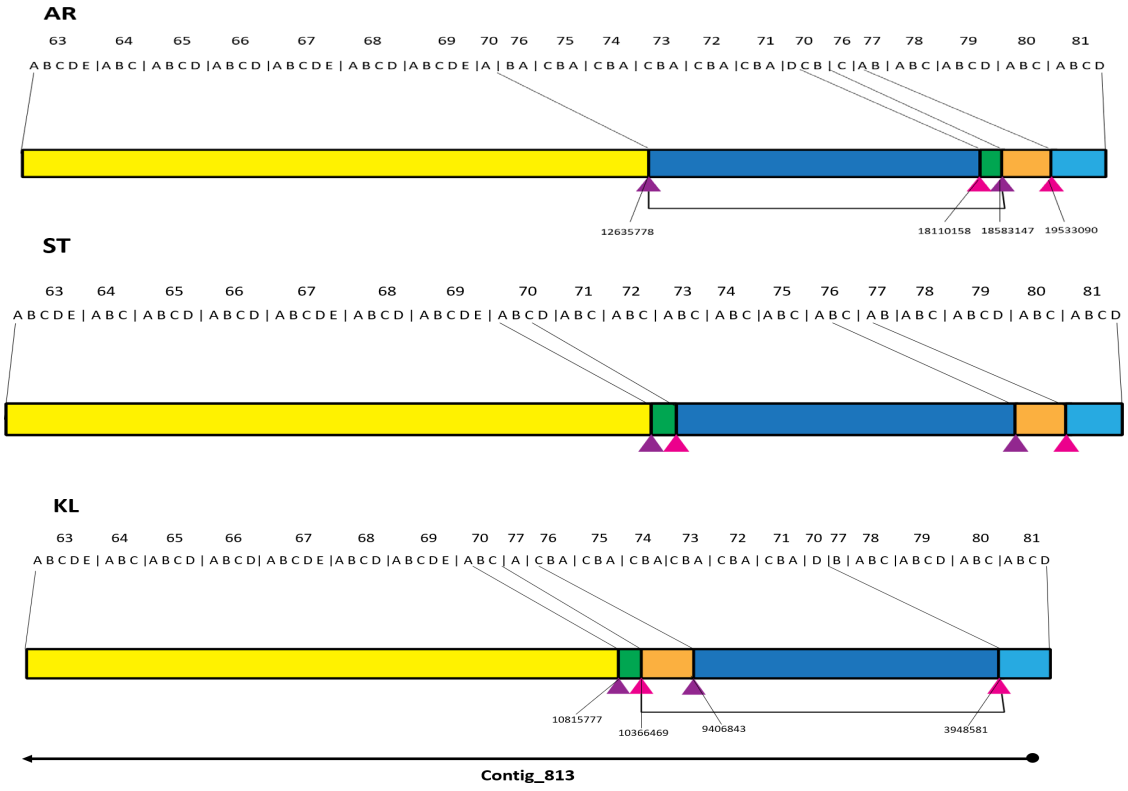


Figure 4. Chromosome maps for Arrowhead, Standard, and Klamath chromosome 3 arrangements that contain breakpoint positions. The Arrowhead genome is derived from reference genome UCI_Dpse_MV25_SWS, while the Klamath diagram is associated with contig 813. Yellow and light blue boxes represent contiguous segments. Pink arrows represent Klamath-Standard break points and purple arrows represent Arrowhead-Standard break points. Green, orange, and dark blue boxes represent inverted regions. Top numbers represent the chromosome regions in cytogenetic map and bottom numbers in the Arrowhead and Klamath diagrams represent identified chromosomal breakpoint positions.

In reference to the Dobzhansky cytogenetic maps[16], the Standard chromosomal positions for chromosome 3 are arranged numerically for sections 70-77. In contrast, for the Arrowhead gene arrangement, the order of the chromosomal positions are as follows: 76B-70D, 70C-B, 76C-77A. From the experimental data generated for the Arrowhead arrangement, chromosome position 12635778 - 18110158 correlates to sections 76B-70D. Chromosome position 18110158 - 18583147 correlates to section 70B-C. Lastly, chromosome position 18583147 - 19533090 correlates to sections 76C-77A. For the Klamath gene arrangement, the order of the chromosome positions is the following: 70B-C, 77A-76C, 76B-70D. From the experimental data generated, chromosome position 10815777 - 10366469 correlates to 70B-C, while chromosome position 10366469 - 9406843 correlates to sections 77A-76C. Lastly, chromosome position 9406843 - 3948581 correlates to sections 76B-70D.

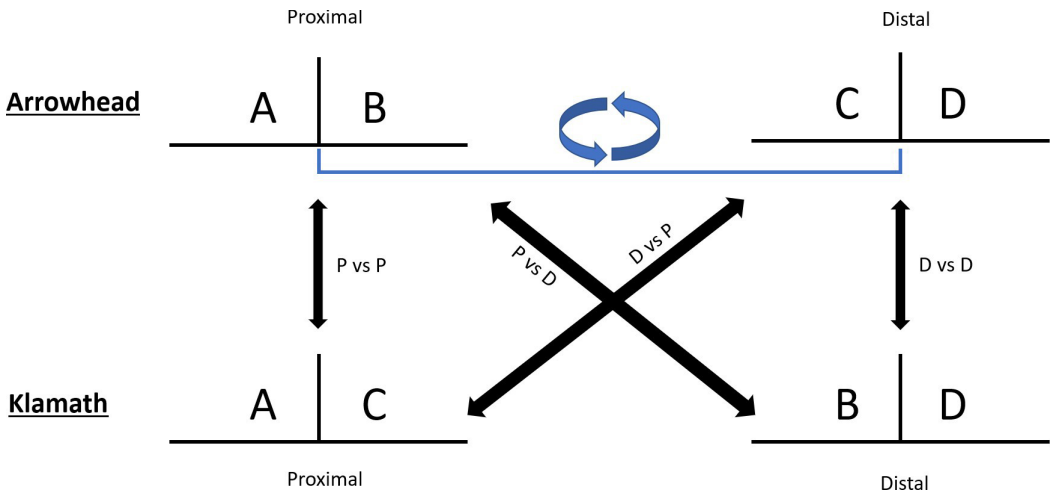


Figure 5: Diagram illustrating the inversions between the Arrowhead and Klamath gene arrangement, in terms of the proximal and distal breakpoints.

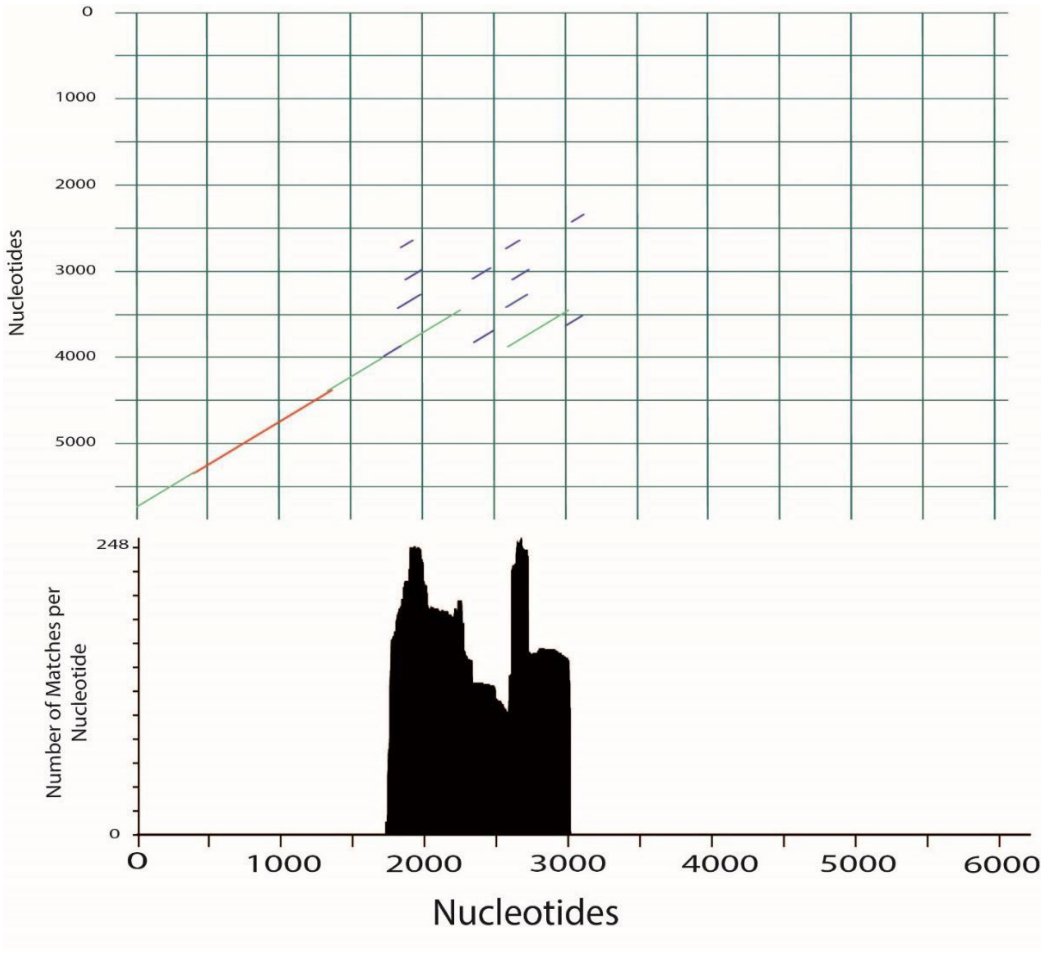


Figure 9. The top figure shows a dot plot comparison of the breakpoint sequence for the proximal Standard-Arrowhead (pSTAR) from the *D. pseudoobscura* AR reference genome MV25 (horizontal axis) and the distal Standard-Klamath (dSTKL) from *D. persimilis* KL SN104-2 genome (vertical axis). Diagonal lines indicate similar sequences between the two sequences with blue being less similar and red being more similar. The top 43 of 1256 (3.4%) matches were included in the plot. The bottom figure is a histogram that integrates the set of BLAST High-Scoring Segment Pair from the comparison of the *D. pseudoobscura* AR reference genome MV25 sequence to the *D. persimilis* KL SN104-2 genome. The elevated match numbers on the bottom graph represent regions within the breakpoint sequence that are repeated in the *D. persimilis* genome.

For the pSTKL from Arrowhead *D. pseudoobscura* genome and the dSTKL from the Klamath *D. persimilis* genome, there was linear similarity between the 5' end of the Arrowhead arrangement and 3' Klamath arrangement. There was no similarity between the 3' end of Arrowhead and 5' end of Klamath as a breakpoint populated with repeat sequences, as large as 248 nucleotides, interrupted the Klamath sequence.

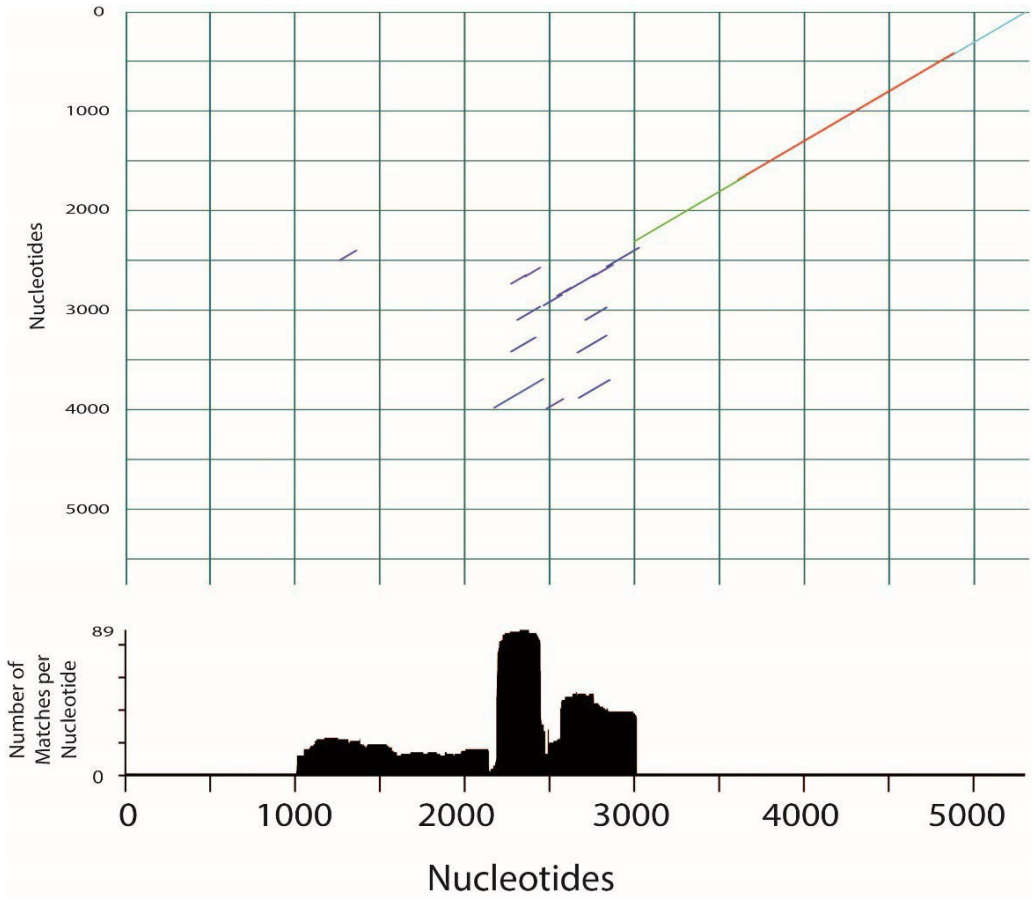


Figure 6. The top figure shows a dot plot comparison of the breakpoint sequence for the dSTAR from the *D. pseudoobscura* AR reference genome MV25 (horizontal axis) and the dSTKL from *D. persimilis* KL SN104-2 genome (vertical axis). Diagonal lines indicate similar sequences between the two sequences with blue being less similar and red being more similar. The top 36 of 953 (3.7%) matches were included in the plot. The bottom figure is a histogram that integrates the set of BLAST High-Scoring Segment Pair from the comparison of the *D. pseudoobscura* AR reference genome MV25 sequence to the *D. persimilis* KL SN104-2 genome. The elevated match numbers in the bottom graph represent regions within the breakpoint sequence that are repeated in the *D. persimilis* genome.

Between dSTKL from Arrowhead *D. pseudoobscura* genome and the dSTKL from the Klamath *D. persimilis* genome, there was similarity between the 3' end of the Arrowhead arrangement and 5' Klamath arrangement. There was no similarity between the 5' end of Arrowhead and 3' end of Klamath as a breakpoint populated with repeat sequences, as large as 248 nucleotides, interrupted the Klamath sequence.

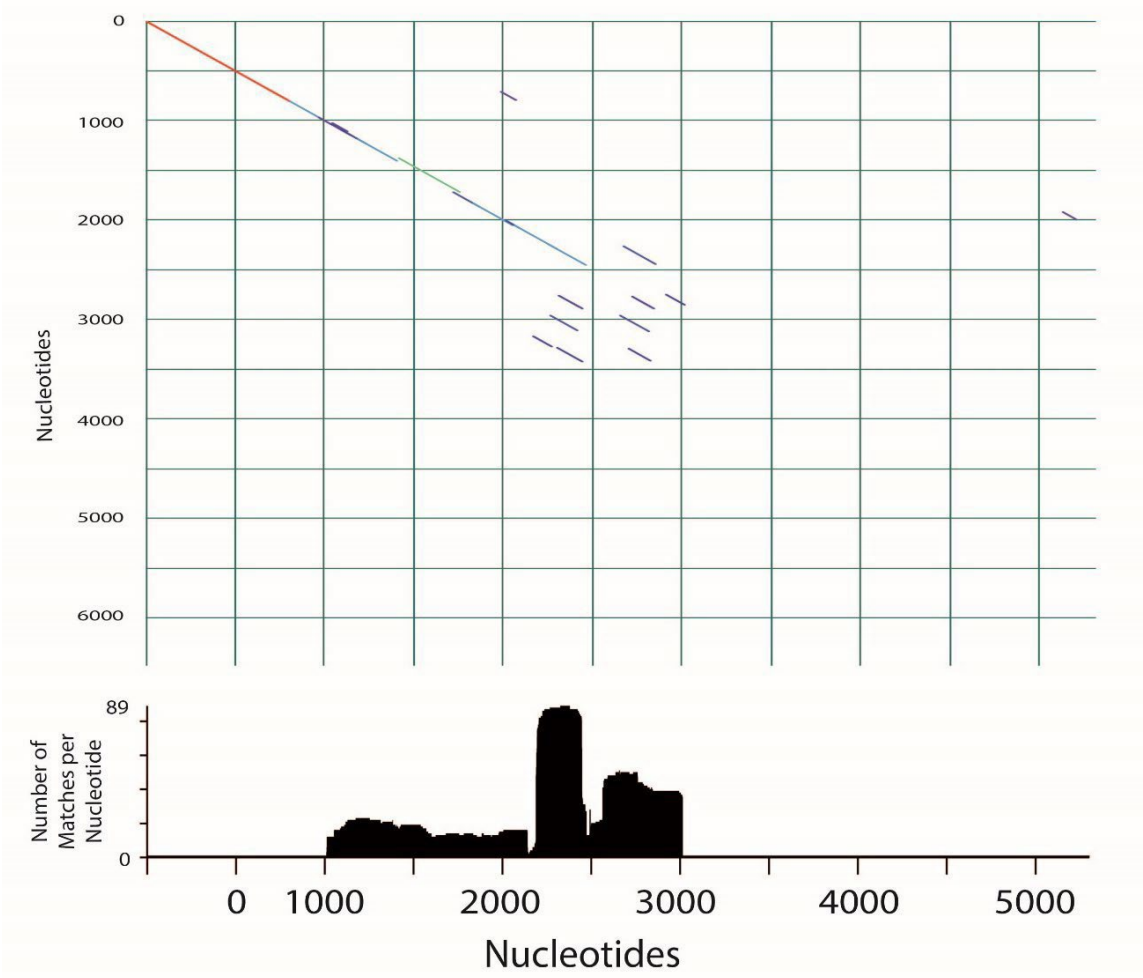


Figure 7. The top figure shows a dot plot comparison of the breakpoint sequence for the dSTAR from the *D. pseudoobscura* AR reference genome MV25 (horizontal axis) and the pSTKL from *D. persimilis* KL SN104-2 genome (vertical axis). Diagonal lines indicate similar sequences between the two sequences with blue being less similar and red being more similar. The top 88 of 1744 (5.0%) matches were included in the plot. The bottom figure is a histogram that integrates the set of BLAST High-Scoring Segment Pair from the comparison of the *D. pseudoobscura* AR reference genome MV25 sequence to the *D. persimilis* KL SN104-2 genome. The elevated match numbers in the bottom graph represent regions within the breakpoint sequence that are repeated in the *D. persimilis* genome.

Between the dSTKL from Arrowhead *D. pseudoobscura* genome and the pSTKL from the Klamath *D. persimilis* genome, there was similarity between the 5' ends of the Klamath and Arrowhead arrangement. There was no similarity between the 3' ends of both genomes as a breakpoint populated with repeat sequences, as large as 39 nucleotides, interrupted the Klamath sequence.

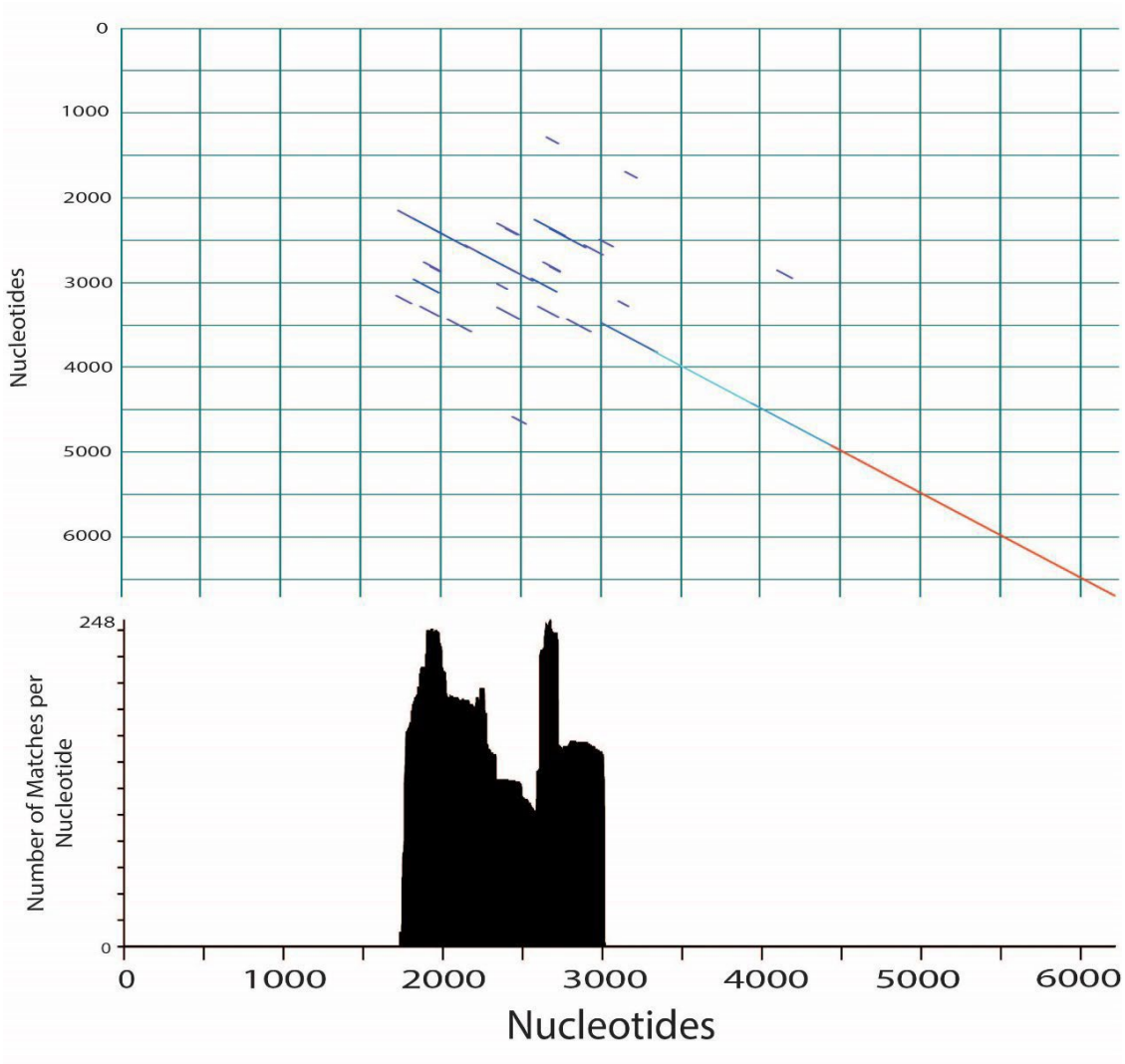


Figure 8. The top figure shows a dot plot comparison of the breakpoint sequence for the pSTAR from the *D. pseudoobscura* AR reference genome MV25 (horizontal axis) and the pSTKL from *D. persimilis* KL SN104-2 genome (vertical axis). Diagonal lines indicate similar sequences between the two sequences with blue being less similar and red being more similar. The top 46 of 1745 (2.6%) matches were included in the plot. The bottom figure is a histogram that integrates the set of BLAST High-Scoring Segment Pair from the comparison of the *D. pseudoobscura* AR reference genome MV25 sequence to the *D. persimilis* KL SN104-2 genome. The elevated match numbers on the bottom graph represent regions within the breakpoint sequence that are repeated in the *D. persimilis* genome.

Between the pSTKL from arrowhead *D. pseudoobscura* genome and the pSTKL from the Klamath *D. persimilis* genome, there was similarity between the 3' ends of the Klamath and Arrowhead arrangement. There was no similarity between the 5' ends of both genomes as a breakpoint populated with repeat sequences, as large as 248 nucleotides, interrupted the Klamath sequence.

Chapter 4

Discussion:

Overall, the *D. persimilis* KL assembly size produced is similar to that of the Arrowhead of the *D. pseudoobscura* reference genome (Dpse_UCI_MV25_SWS) given that it has a total read length of 173,005,162 and a N50 of 21,153,950. Additionally, the assembly chromosomes 2-5 are similar in size to the reference chromosomes, as the assembly chromosomes are 90-96% the length of the reference chromosomes, with specifically chromosome 3 being 94.4% of the reference length. This similarity in lengths between the assembly chromosomes and reference chromosomes is a sign of satisfactory assembly quality. As for chromosome X, while the assembly was only 80.5% of the reference chromosome's length, this assembly only accounted for the euchromatin region, not the heterochromatin regions, as is common with other assemblies of chromosome X[27]. Furthermore, while the size of the overall assembly was large, the reported BUSCO score does indicate some shortcomings in the assembly, as the assembly had a BUSCO score of 88%, which means that only 88% of the Dipteran reference set of protein coding loci were detected as complete. This is considerably lower than the 97% BUSCO score reported for the 15 Drosophila species reported by Miller *et al*. (2021)[32]. The lower score from this study is likely due to indel mutations within coding genes that can lead to frameshift mutations that prevent detection of core genes. The use of short less error-prone reads for additional polishing of the sequence could reduce the number indels in the final assembly, which can improve the quality of the assembly.

By locating the breakpoint regions, inferences about the mechanisms of breakpoints were able to be made. In the proximal and distal breakpoints of the Klamath arrangement from *D. persimilis* and the Arrowhead arrangement from *D. pseudoobscura*, discontinuity between the chromosomal sequences were present within regions of repeat sequences. After this repeat mediated region, there was no significant similarity between the Klamath and Arrowhead arrangements in the proximal or distal orientation. This indicates that the breakpoints in both chromosomal arrangements occur in a region populated with repeat sequences supporting the hypothesis that the Standard to Klamath inversion was mediated by ectopic exchange within the repeat regions.

As in Richards *et al*. (2005), the breakpoint regions of the *D. pseudoobscura* genome appear to be mediated by repeat elements[2]. The same appears true for the *D. persimilis* genome. Similar to the findings in Richards *et al*. (2005), the breakpoint was staggered and at opposite ends of the repeat elements[2]. When comparing the Arrowhead arrangement and the Klamath arrangement, it appears the repeats are between 1250-3000 bp in length, which is greater than what was observed in Richards *et al*. (2005), where the repeat lengths were between 128-315 bp[2].

Overall, the presence of repeat elements located in the breakpoint region of the Klamath arrangement from *D. persimilis* reinforce the emerging evidence that there is a close relationship between increased repeat elements and chromosomal breakpoints. To further explore this concept in the future, the repeat structures located at the Klamath region should be analyzed on a more fine scale nucleotide level.

BIBLIOGRAPHY

1. Griffiths, A. J. F., S. R. Wessler, R. C. Lewontin, W. M. Gelbart, D. T. Suzuki *et al.*, 2005 *An Introduction to Genetic Analysis*. W. H. Freeman & Co., New York, NY

2. Richards, S., Y. Liu, B. R. Bettencourt, P. Hradecky, S. Letovsky *et al.*, 2005 Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene and *cis*-element evolution. Genome Research 15: 1-18.

3. Ranz, J. M., D. Maurin, Y. S. Chan, M. v. Grotthuss, L. W. Hillier *et al.*, 2007 Principles of genome evolution in the *Drosophila melanogaster* species group Public Library of Science Biology 5: 1366-1381.

4. Hsu, L. Y. F., P. A. Benn, H. L. Tannenbaum, T. E. Perlis, A. D. Carlson et al., 1987 Chromosomal polymorphisms of 1, 9, 16, and Y in 4 major ethnic groups: A large prenatal study. American Journal of Medical Genetics 26: 95-101.
5. Pettenati, M. J., P. N. Rao, M. C. Phelan, F. Grass, K. W. Rao *et al.*, 1995 Paracentric inversions in humans: a review of 446 paracentric inversions with presentation of 120 new cases. American Journal of Medical Genetics 55: 171-187.

6. Liehr, Thomas, et al. "Recombinant Chromosomes Resulting from Parental Pericentric Inversions—Two New Cases and a Review of the Literature." Frontiers in Genetics, vol. 10, 2019, https://doi.org/10.3389/fgene.2019.01165.

7. Puig, Marta et al. "Human inversions and their functional consequences." Briefings in functional genomics vol. 14, 5 2015: 369-79. doi:10.1093/bfgp/elv020

8. Lande, R., 1984 The expected fixation rate of chromosomal inversions. Evolution 38**:** 743-752.
9. Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favorable gene. Genetical Research 23: 23-35.

10. Nei, M., K. I. Kojima and H. E. Schaffer, 1967 Frequency changes of new inversions in populations under mutation-selection equilibria. Genetics 57: 741-750.

11. Ohta, T., and K. I. Kojima, 1968 Survival probabilities of new inversions in large populations. Biometrics 24: 501-516.
12. Kirkpatrick, M., and N. Barton, 2006 Chromosome inversions, local adaptation and speciation. Genetics 173: 419-434.

13. Dobzhansky, T., 1950 The genetics of natural populations. XIX. Origin of heterosis through natural selection in populations of *Drosophila pseudoobscura*. Genetics 35**:** 288-302.

14. Painter, T. S., 1933 A new method for the study of chromosome rearrangements and the plotting of chromosome maps. Science 78: 585-586.

15. Dobzhansky, T., and A. H. Sturtevant, 1938 Inversions in the chromosomes of *Drosophila pseudoobscura*. Genetics 23: 28-64.

16. Dobzhansky, T., et al. Contributions to the Genetics, Taxonomy, and Ecology of Drosophila Pseudoobscura and Its Relatives, 1944, Carnegie Institution

17. Dobzhansky, T., and A. H. Sturtevant, 1938 Inversions in the chromosomes of *Drosophila pseudoobscura*. Genetics 23: 28-64.

18. Anderson, W. W., F. J. Ayala and R. E. Michod, 1977 Chromosomal and allozymic diagnosis of three species of *Drosophila*, *Drosophila pseudoobscura*, *Drosophila persimilis,* and *Drosophila miranda*. Journal of Heredity 68: 71-74.

19. Schaeffer, S. W., A. Bhutkar, B. F. McAllister, M. Matsuda, L. M. Matzkin *et al.*, 2008 Polytene chromosomal maps of 11 Drosophila species: The order of genomic scaffolds inferred from genetic and physical maps. Genetics 179**:** 1601-1655.

20. Noor, Mohamed A. F. et al. "Divergence between the Drosophila pseudoobscura and D. persimilis genome sequences in relation to chromosomal inversions." Genetics vol. 177,3 (2007): 1417-28. doi:10.1534/genetics.107.070672

21. Fuller, Z. L., S. A. Koury, N. Phadnis and S. W. Schaeffer, 2019 How chromosomal rearrangements shape adaptation and speciation: Case studies in *Drosophila pseudoobscura* and its sibling species *Drosophila persimilis*. Molecular Ecology 28: 1283-1301.

22. Fuller, Zachary L, et al. "Genomics of Natural Populations: How Differentially Expressed Genes Shape the Evolution of Chromosomal Inversions in Drosophila pseudoobscura." Genetics, vol. 204, no. 1, 2016, pp. 287–301.

23. Kolmogorov, M., J. Yuan, Y. Lin and P. A. Pevzner, 2019 Assembly of long, error-prone reads using repeat graphs. Nature Biotechnology 37: 540-546.

24. Lin, Y., J. Yuan, M. Kolmogorov, M. W. Shen, M. Chaisson *et al.*, 2016 Assembly of long error-prone reads using de Bruijn graphs. Proceedings of the National Academy of Sciences 113**:** E8396-E8405.

25. Anderson, W. W., F. J. Ayala and R. E. Michod, 1977 Chromosomal and allozymic diagnosis of three species of *Drosophila*, *Drosophila pseudoobscura*, *Drosophila persimilis,* and *Drosophila miranda*. Journal of Heredity 68: 71-74.

26. Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva and E. M. Zdobnov, 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210-3212.

27. Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne *et al.*, 2000 The genome sequence of *Drosophila melanogaster*. Science 287: 2185-2195.

28. Bartolome, C., and B. Charlesworth, 2006 Rates and patterns of chromosomal evolution in Drosophila pseudoobscura and D. miranda. Genetics 173: 779-791.

29. Wallace, A. G., D. Detweiler and S. W. Schaeffer, 2011 Evolutionary history of the third chromosome gene arrangements of Drosophila pseudoobscura inferred from inversion breakpoints. Molecular Biology and Evolution 28: 2219-2229.

30. Fuller, Z. L., G. D. Haynes, S. Richards and S. W. Schaeffer, 2017 Genomics of natural populations: Evolutionary forces that establish and maintain gene arrangements in Drosophila pseudoobscura. Molecular Ecology 26: 6539-6562.

31. Bingham, P. M., R. Levis and G. M. Rubin, 1981 Cloning of DNA sequences from the white locus of D. melanogaster by a novel and general method. Cell 25: 693-704.

32. Miller, D. E., C. Staber, J. Zeitlinger and R. S. Hawley, 2018 GENOME REPORT: Highly Contiguous Genome Assemblies of 15 Drosophila Species Generated Using Nanopore Sequencing. G3: Genes|Genomes|Genetics 8: 3131-3141.

33. Moore, B. C., and C. E. Taylor, 1986 Drosophila of southern California. III Gene arrangements of Drosophila persimilis. Journal of Heredity 77: 313-323

ACADEMIC VITA

Academic Vita of Amira Ellison
aue234@psu.edu

The Pennsylvania State University
Bachelors of Science in Biology

Thesis Title: Genomic Analysis of the Klamath Gene Rearrangement of Drosophila Persimilis

Thesis Supervisor: Dr. Stephen Schaeffer

Work Experience
Schaeffer Lab
10/2019-04/2021
Research Assistant
Pennsylvania State University, State College PA
Dr. Stephen Schaeffer

Hanchard Lab
06/2020-08/2020
Research Assistant
Baylor College of Medicine Human Genome Sequencing Center, Houston TX
Dr. Neil Hanchard

Shendure Lab
06/2021-08/2021
Research Assistant
University of Washington Genome Sciences Department, Seattle WA
Dr. Lea Starita

Awards: Dean's List

Community Service Involvement: Envision Event