Title

Chromosome-Level Assembly of Drosophila bifasciata Reveals Important Karyotypic Transition of the X Chromosome

Journal

G3: Genes, Genomes, Genetics, 10(3)

ISSN

2160-1836

Authors

Bracewell, Ryan
Tran, Anita
Chatla, Kamalakar
et al.

Publication Date

2020-03-01

DOI

10.1534/g3.119.400922

Peer reviewed

G3 Genes | Genomes | Genetics

GENOME REPORT

Chromosome-Level Assembly of *Drosophila bifasciata* Reveals Important Karyotypic Transition of the X Chromosome

Ryan Bracewell,[1] Anita Tran, Kamalakar Chatla, and Doris Bachtrog
Department of Integrative Biology, University of California Berkeley, Berkeley, CA 94720
ORCID IDs: 0000-0003-2678-4758 (R.B.); 0000-0001-9724-9467 (D.B.)

ABSTRACT The *Drosophila obscura* species group is one of the most studied clades of Drosophila and harbors multiple distinct karyotypes. Here we present a *de novo* genome assembly and annotation of *D. bifasciata*, a species which represents an important subgroup for which no high-quality chromosome-level genome assembly currently exists. We combined long-read sequencing (Nanopore) and Hi-C scaffolding to achieve a highly contiguous genome assembly approximately 193 Mb in size, with repetitive elements constituting 30.1% of the total length. *Drosophila bifasciata* harbors four large metacentric chromosomes and the small dot, and our assembly contains each chromosome in a single scaffold, including the highly repetitive pericentromeres, which were largely composed of Jockey and Gypsy transposable elements. We annotated a total of 12,821 protein-coding genes and comparisons of synteny with *D. athabasca* orthologs show that the large metacentric pericentromeric regions of multiple chromosomes are conserved between these species. Importantly, Muller A (X chromosome) was found to be metacentric in *D. bifasciata* and the pericentromeric region appears homologous to the pericentromeric region of the fused Muller A-AD (XL and XR) of *pseudoobscura/affinis* subgroup species. Our finding suggests a metacentric ancestral X fused to a telocentric Muller D and created the large neo-X (Muller A-AD) chromosome ~15 MYA. We also confirm the fusion of Muller C and D in *D. bifasciata* and show that it likely involved a centromere-centromere fusion.

KEYWORDS
chromosome
Muller element
centromere
Nanopore

Recent advances in DNA sequencing technology have dramatically improved the quality and quantity of genome assemblies in both model and non-model species. Long-read sequencing technologies (*e.g.*, PacBio and Nanopore) combined with long-range scaffolding information generated through chromatin conformation capture methods such as Hi-C (Lieberman-Aiden *et al.* 2009) or Chicago (Putnam *et al.* 2016) can produce assemblies of unprecedented length and accuracy. However, there are still relatively few assemblies that traverse through megabase-long stretches of highly repetitive sequence, thereby limiting our understanding of the evolution of pericentromere/heterochromatic regions of the genome and the genes, satellites, and transposable elements that inhabit them (Chang *et al.* 2019, Miga 2019).

Drosophila has been at the forefront of genetics and genomics research for over a century and new chromosome-level assemblies are now becoming available for several non-model species (Mahajan *et al.* 2018, Miller *et al.* 2018, Bracewell *et al.* 2019, Karageorgiou *et al.* 2019, Mai *et al.* 2020). Recent comparative genomic analysis in the *Drosophila obscura* group has revealed extensive karyotype evolution and turnover of centromeric satellites that alters chromosome morphology (Bracewell *et al.* 2019) (Figure 1). Unfortunately, our understanding of karyotype and genome evolution is currently limited because no high-quality assembly of a species from the *obscura* subgroup is available (Figure 1). Given the phylogenetic placement of *D. bifasciata* (Figure 1) and its putative chromosomal configuration (Buzzati-Traverso and Scossiroli 1955, Moriwaki and Kitagawa 1955), it is an important species for reconstructing karyotype evolution in the *obscura* group for several reasons. First, a high-quality *D. bifasciata* genome assembly allows us to better understand the

doi: https://doi.org/10.1534/g3.119.400922
Manuscript received November 18, 2019; accepted for publication January 20, 2020; published Early Online January 22, 2020.

Supplemental material available at figshare: https://doi.org/10.25387/g3.11561892.
[1]Corresponding author: Department of Integrative Biology, University of California, Berkeley, 3040 Valley Life Sciences Building #3140, Berkeley, CA 94720-3140. E-mail: ryan.bracewell@berkeley.edu

emergence of metacentric chromosomes and determine if metacentric pericentromeres are conserved over evolutionary time (Figure 1). Second, the configuration of the Muller A chromosome (the ancestral X chromosome in Drosophila) is particularly interesting, since it became fused to Muller D in some members of the *obscura* group ~15 million years ago (Figure 1) thereby creating a large neo-sex chromosome (Carvalho and Clark 2005). The location of the centromere (metacentric or telocentric) prior to the fusion is not known, and the A-to-D fusion has been a matter of some debate (Schaeffer 2018). If Muller A was metacentric prior to the fusion, that could explain the presence of ancestral Muller A genes on the long arm of the fused A-D chromosome (denoted XR in *D. pseudoobscura*) (Mahajan *et al.* 2018, Bracewell *et al.* 2019) (hereafter referred to as Muller A-AD). Third, the putative Muller C-D fusion is only present in some species of the *obscura* subgroup, suggesting it occurred recently. How the chromosomes fused is unknown (centromere-centromere, centromere-telomere, telomere-telomere) and the relative size and gene content of this new pericentromeric region is unknown. Here, we report on our genome assembly and annotation of *D. bifasciata* and we characterize chromosome structure, the distribution of transposable elements (TE), and explore the putative Muller C-D fusion.

METHODS AND MATERIALS

Genome sequencing and assembly

We sequenced the *D. bifasciata* isofemale line 14012-0181.02, which was originally collected in Hokkaido, Japan and obtained from the National Drosophila Species Stock Center at Cornell University. High molecular weight DNA for sequencing was extracted from ~60 female flies using a Qiagen Blood & Cell Culture DNA Midi Kit and the resulting DNA was size selected for fragments >15 kb using BluePippin (Sage Science). For size selection, we used 6 μg of total DNA (100 ng/μl) run in two wells. The elute was bead purified, resulting in a total of 2.7 μg total DNA in a 50 μl solution. We generated long-reads using Nanopore and the SQK-LSK109 sequencing kit on one 9.4.1RevD flow cell and with the minKNOW software version 3.1.13. Raw output files from our sequencing run were base called using Albacore Sequencing Pipeline software version 2.3.3 (Oxford Nanopore Technologies) with default parameters for quality score filtering.

We used Canu version 1.8 (Koren *et al.* 2017) to first error-correct the raw sequencing reads using slightly modified parameters (correctedErrorRate = 0.065 corMinCoverage = 8 batOptions="-dg 3 -db 3 -dr 1 -ca 500 -cp 50" trimReadsCoverage = 4 trimReadsOverlap = 500 genomeSize = 200m). The resulting error-corrected reads were then assembled into contigs using the WTDBG2 assembler (Ruan and Li 2019) with default settings. We then BLAST searched all contigs <1 MB to the nt database and returned the top two hits to identify any contigs from non-target species (typically *Acetobacter* and *Saccharomyces*).

After removing contaminant contigs we polished the genome assembly using three rounds of Racon (Vaser *et al.* 2017) followed by three rounds of Pilon (Walker *et al.* 2014). This method of combining multiple rounds of Racon and Pilon has been shown to increase genome assembly quality in other Drosophila species (Miller *et al.* 2018). To polish with Racon we mapped our raw nanopore reads each round using minimap2 and specified -x ava-ont (Li 2018). For genome polishing with Pilon we used reads derived from our high coverage Hi-C Illumina data (below). Because of the inherent properties of Hi-C data (paired-end reads with atypical orientations, highly variable insert sizes, chimeric reads) that could lead to spurious genome polishing, we first mapped our Illumina reads to the genome using BWA mem (Li and Durbin 2009) and extracted only those reads with correct pairing using samtools (view -bf 0x2) (Li *et al.* 2009). We then used those reads as single-end reads for genome polishing. A fraction of the single-end reads will be chimeric but read mapping with BWA mem soft-clips reads and these soft-clipped reads should be randomly distributed across the genome (Figure S1) and not contribute significantly to genome polishing. At each step of assembly and polishing we assessed genome completeness using BUSCO *v3* (Simão *et al.* 2015) and the odb9 eukaryota database.

Hi-C scaffolding

Prior to scaffolding we compared our polished contigs with other chromosome-length genome assemblies from *obscura* group species (Mahajan *et al.* 2018, Bracewell *et al.* 2019) using whole genome alignments with D-Genies (Cabanettes and Klopp 2018). We then identified the largest contigs belonging to Muller elements to help guide any potential manual manipulations during Hi-C scaffolding. To scaffold the assembly, we used chromatin conformation capture to generate

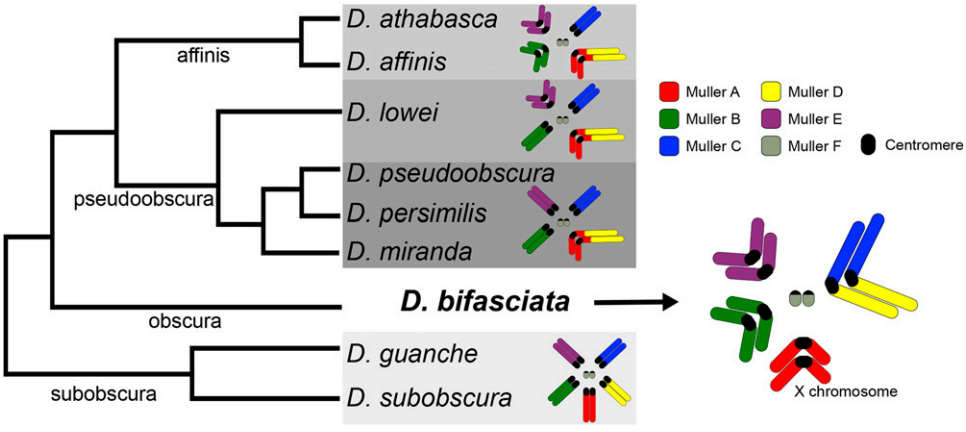

Figure 1 Evolutionary relationships and karyotype transitions of *obscura* group flies. The ancestral karyotype of the *obscura* group (shown here as *Drosophila subobscura*) consists of five large and one small pair of telocentric chromosomes, referred to as Muller elements A-F (reviewed in Schaeffer 2018), and shown color coded. Significant karyotypic changes have occurred across the *obscura* group (highlighted with gray boxes) with chromosomal fusions and centromere movement altering chromosome structure (Bracewell *et al.* 2019). *Drosophila bifasciata* represents an important karyotype to understand evolutionary transitions since Muller A (the X chromosome), B and E are thought to be metacentric and Muller A is unfused (Moriwaki and Kitagawa 1955). In *D. bifasciata*, it is thought that Muller C and D fused, although C-D fusions are only present in some *obscura* subgroup species (Buzzati-Traverso and Scossiroli 1955). Shown phylogenetic relationships adapted from (Gao *et al.* 2007) with subgroup designations shown along the branches.

Hi-C data (Lieberman-Aiden *et al.* 2009). We generated Hi-C libraries as outlined in Bracewell *et al.* (2019) using a DNase digestion method (Ramani *et al.* 2016). The resulting DNA library was prepped using an Illumina TruSeq Nano library prep kit and was sequenced on a HiSeq 4000 with 100 bp PE reads. We used Juicer (Durand *et al.* 2016b) to map raw Hi-C reads and generate contact maps based on 3D interactions to scaffold the genome assembly. We then used the 3D-DNA pipeline (Dudchenko *et al.* 2017) to orient and place contigs. 3D-DNA output files were visualized and checked for accuracy using Juicebox (Durand *et al.* 2016a) with verification and modifications to scaffolding done using built-in tools. The final assembly was scaffolded together with 300 Ns between each contig.

Estimating residual isofemale line genetic variation and genome assembly accuracy

Residual genetic variation can complicate genome assembly and lead to varying assembly quality across chromosomes. To characterize genomic patterns of variation in the sequenced isofemale line, we used the mapped Illumina polishing reads (above) and GATK's UnifiedGenotyper (DePristo *et al.* 2011) to call single nucleotide polymorphisms (SNPs). We filtered SNPs using VCFtools (Danecek *et al.* 2011) (–minGQ 20 –minDP 3 –min-alleles 2) and estimated nucleotide diversity (π) in 50 kb non-overlapping windows, with the expectation that diversity should be nearly zero for genomic regions that are isogenic. To estimate base level accuracy (QV) of the genome assembly, we followed methods outlined in Koren *et al.* (2018) and used FreeBayes (Garrison and Marth 2012) to identify variants in our mapped Illumina polishing reads (same as above) with the command -C 2 -0 -O -q 20 -z 0.10 -E 0 -X -u –p 2 –F 0.75 -b input.bam -v output.vcf -f reference.fasta. Heterozygous calls (0/1) with a reference allele were filtered out and QV was calculated as:

$$-10 \log_{10} \frac{B}{T}$$

where B is the sum of all bases changed from indels and homozygous alternate SNPs with sequencing depth ≥3×, and T is total bases with sequencing depth ≥3×.

Repetitive element identification and genome masking

We first used REPdenovo (Chu *et al.* 2016) to identify novel repeats from our single-end Hi-C Illumina sequencing data (above) using parameters described in detail in Bracewell *et al.* (2019). We then concatenated the REPdenovo repeats with the Repbase *Drosophila* repeat library (downloaded March 22, 2016, from www.girinst.org) and used this combined file to mask the genome with RepeatMasker version 4.0.7 using the -no_is and -nolow flags. To characterized the genomic distribution of specific transposable element (TE) families we used a TE library developed from *obscura* group flies (Hill and Betancourt 2018) and again used RepeatMasker and then bedtools coverage (Quinlan and Hall 2010) to determine the proportion of masked bases per TE family.

Genome annotation and characterization of assembly

To annotate our *D. bifasciata* genome assembly we used the REPdenovo/Repbase repeat-masked genome (above) and the MAKER annotation pipeline (Campbell *et al.* 2014) to identify gene models. The *ab initio* gene predictors SNAP (Korf 2004) and Augustus (Stanke and Waack 2003) were used to guide the annotation and we used protein sets from *D. pseudoobscura* and *D. melanogaster* (FlyBase) to aid in gene prediction. We used karyoploteR (Gel and Serra 2017) to plot features of the *D. bifasciata* genome assembly.

Gene orthologs, genome synteny, and Muller element fusion orientation

To compare our genome assembly with *D. athabasca* which has metacentric Muller A-AD, B and E chromosomes (Figure 1), and *D. subobscura*, which harbors the ancestral karyotype and is composed entirely of telocentric chromosomes (Figure 1), we performed BLASTP reciprocal best hit searches between proteins from our annotations of each species (Bracewell *et al.* 2019). We used the blast_rbh.py script (Cock *et al.* 2015) and genomic coordinates of reciprocal best hits were plotted using the *genoPlotR* package (Guy *et al.* 2010). To determine if the Muller C-D fusion in *D. bifasciata* was the result of a centromere-centromere, centromere-telomere, or telomere-telomere fusion, we identified the 50 most proximal pericentromeric genes from telocentric Muller C and D in *D. subobscura* and plotted the location of orthologs in *D. bifasciata*.

Data availability

Raw Nanopore and Hi-C (Illumina) reads have been deposited in the NCBI SRA and are under the BioProject (PRJNA565796). The genome assembly and annotation have been deposited with NCBI (accession WIOZ00000000.1). Figure S1 shows an Integrative Genomics Viewer image of a region on Muller A of the genome assembly with mapped Hi-C reads filtered for Pilon polishing. Figure S2 contains a plot of Illumina and Nanopore sequencing coverage over the draft genome assembly. Figure S3 shows the Hi-C heatmap and scaffolding as in Figure 1 with nucleotide diversity (π) estimated from the isofemale line in 50 kb non-overlapping windows across the assembly. Figure S4 displays the genomic distribution of the ten most frequently encountered transposable elements (TEs) for each Muller element in the genome assembly. Figure S5 shows the genomic location of *D. subobscura* pericentromeric orthologs from Muller C and D on the fused Muller CD of *D. bifasciata*. Supplemental material available at figshare: https://doi.org/10.25387/g3.11561892.

RESULTS AND DISCUSSION

Using one Nanopore flow cell, we generated 538,757 reads that passed Albacore's standard quality filtering. Our Nanopore reads had an N50 read length of 23,957 bases and provided ∼45× coverage over the genome given an estimated genome size of ∼200 Mb for *D. bifasciata*. Our initial hybrid Canu/WTDBG2 assembly resulted in a genome assembly that consisted of 796 contigs with an N50 of 2,325,530. BLAST results flagged multiple putative bacterial contigs (primarily *Acetobacter*) and 49 contigs (5.5 Mb of total sequence) were removed. As expected, rounds of Racon polishing (3×) and subsequent Pilon polishing (3×) led to an appreciable increase in our BUSCO scores (Table 1) although the most significant increases in genome completeness were detected after the initial round of Racon or Pilon. Pilon polishing did not lead to as dramatic an increase in genome completeness as seen in other studies (Bracewell *et al.* 2019) and this was likely due to limitations of our Illumina polishing data that was single-end and was of modest coverage (mean 18×) over the genome (Figure S2). However, we did see a significant increase in genome completeness suggesting that polishing the genome with Hi-C reads can be a viable strategy for increasing genome assembly quality. Our polished genome assembly consisted of 747 contigs with an N50 of 2,386,451. The longest contig was 18,852,285 bp with a total genome assembly length of 192,589,718 bp.

Table 1 BUSCO results from the genome assembly and polishing process

	Canu/WTDBG2 only	Canu/WTDBG2 + Racon 3x	Canu/WTDBG2 + Racon 3x + Pilon 3x	Final Hi-C scaffolded Dbif_1.0
Complete BUSCOs	958	961	1020	1020
Single-copy BUSCOs	947	955	1009	1009
Duplicated BUSCOs	11	6	11	11
Fragmented BUSCOs	47	43	5	5
Missing BUSCOs	61	62	41	41
% BUSCOs complete	89.9%	90.2%	95.7%	95.7%

Our Hi-C library generated a total of 13,018,415 sequenced read pairs of which 73.8% were alignable to the draft genome. The Juicer pipeline identified 6,734,204 Hi-C contacts which were used to scaffold the genome. The final scaffolded genome assembly placed 126 contigs to the fused Muller CD, 54 to Muller A, 238 to Muller B, 119 to Muller E, 1 to Muller F, and 209 were left unplaced (Figure 2A). Hi-C scaffolding revealed clear associations between euchromatic arms of the same chromosome thereby increasing our confidence in the assembly of metacentric chromosomes (Figure 2A). For example, Muller CD is thought to be the result of a fusion of telocentric Muller C and D elements (Moriwaki and Kitagawa 1955) and our assembly showed clear associations between the C and D arms (Figure 2A). Importantly, there were also clear associations between Muller C and D euchromatic arms with adjacent pericentromeric contigs (Figure 2A), thus providing evidence for the placement of the repeat-rich pericentromeric sequence as well (Figure 2B). Muller A also showed clear associations that extend into highly repetitive pericentromeric regions highlighting this chromosome is indeed metacentric (Figure 2A). The combination of inter-arm and arm-pericentromere Hi-C associations allowed us to determine the correct orientation for all arms of the *D. bifasciata* chromosomes.

BUSCO results suggest our final scaffolded genome assembly is of high quality and 95.7% of BUSCOs were found complete (Table 1). We found the BUSCO statistics to be slightly lower than our other high-quality *obscura* group assemblies which average 98.7% complete (Bracewell *et al.* 2019). To investigate this reduction, we looked for missing BUSCOs in a species with a similar karyotype and higher score (*D. athabasca*) and found that 49% of missing BUSCOs (20 of 41 total) were in pericentromeric regions. Therefore, residual genome assembly and polishing issues of highly repetitive pericentromeric regions are likely the main contributor to the slightly lower scores of *D. bifasciata*.

Base level accuracy (QV) was found to be rather high at 35.9. We identified 25,889 indels and 19,151 homozygous SNPs that differed from the genome assembly over the 174,670,905 bases with coverage ≥3×. The assembly QV for *D. bifasciata* is slightly lower than assemblies of the reference genome strain of *Drosophila melanogaster* (Koren *et al.* 2017, Solares *et al.* 2018) but slightly higher than that of the domestic goat (Bickhart *et al.* 2017). A likely contributor to the lower accuracy than *D. melanogaster* (Solares *et al.* 2018) was the levels of genetic variation we identified in the isofemale line used for sequencing (Figure S3). We found moderately high nucleotide diversity over most of Muller CD, and the long arms of Muller B and E. In contrast, the

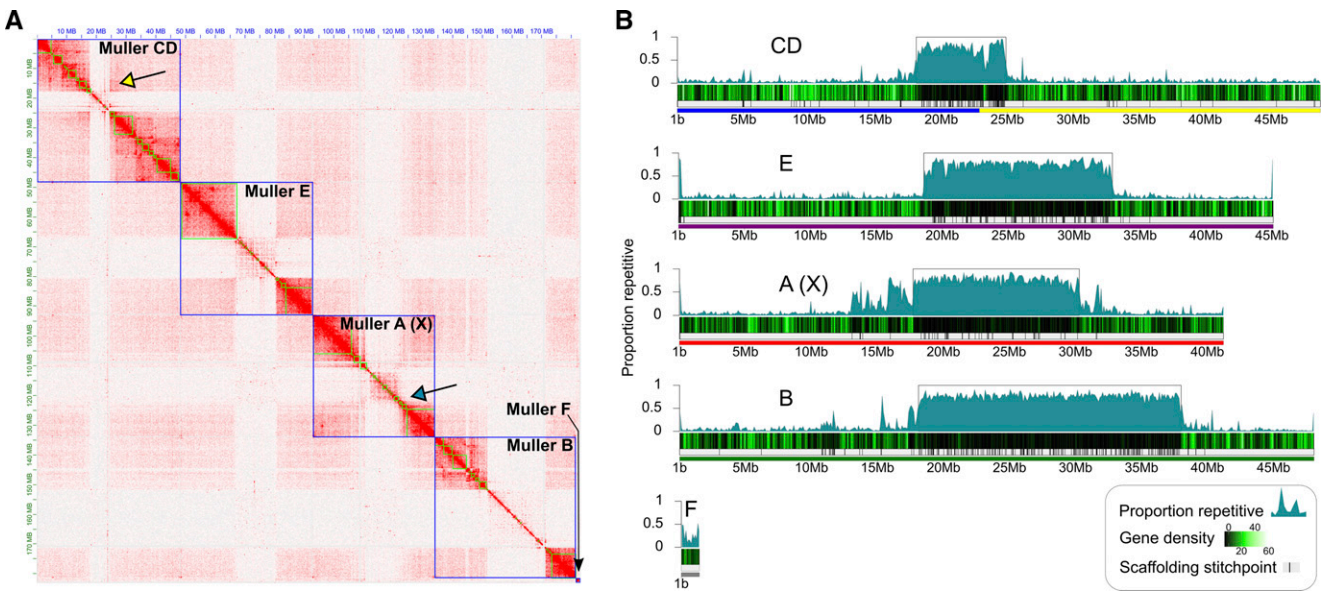

Figure 2 Chromosome-level genome assembly of *Drosophila bifasciata* using Hi-C. A) Hi-C heatmap showing long-range contacts and scaffolding of the genome assembly. Green and blue squares denote contigs and chromosomes, respectively. Euchromatic chromosome arms and heterochromatic pericentromeres for each chromosome show distinct and primarily isolated associations that resemble a 'checkerboard' pattern. Note that chromosome arms on opposite sides of a pericentromere often show associations on the diagonal confirming their placement (yellow arrow) while pericentromeres show finer-scale associations with their chromosome arms (blue arrow). B) Shown is the *D. bifasciata* genome assembled into Muller elements (color coded as in Figure 1), scaffolding stitch points, gene density (genes per 100 kb) and repeat content (proportion of bases repeat-masked in 100 kb non-overlapping windows). Boxes around highly repetitive regions indicate putative pericentromere boundaries (defined as ≥40% repeat-masked sequence in sliding windows away from the center).

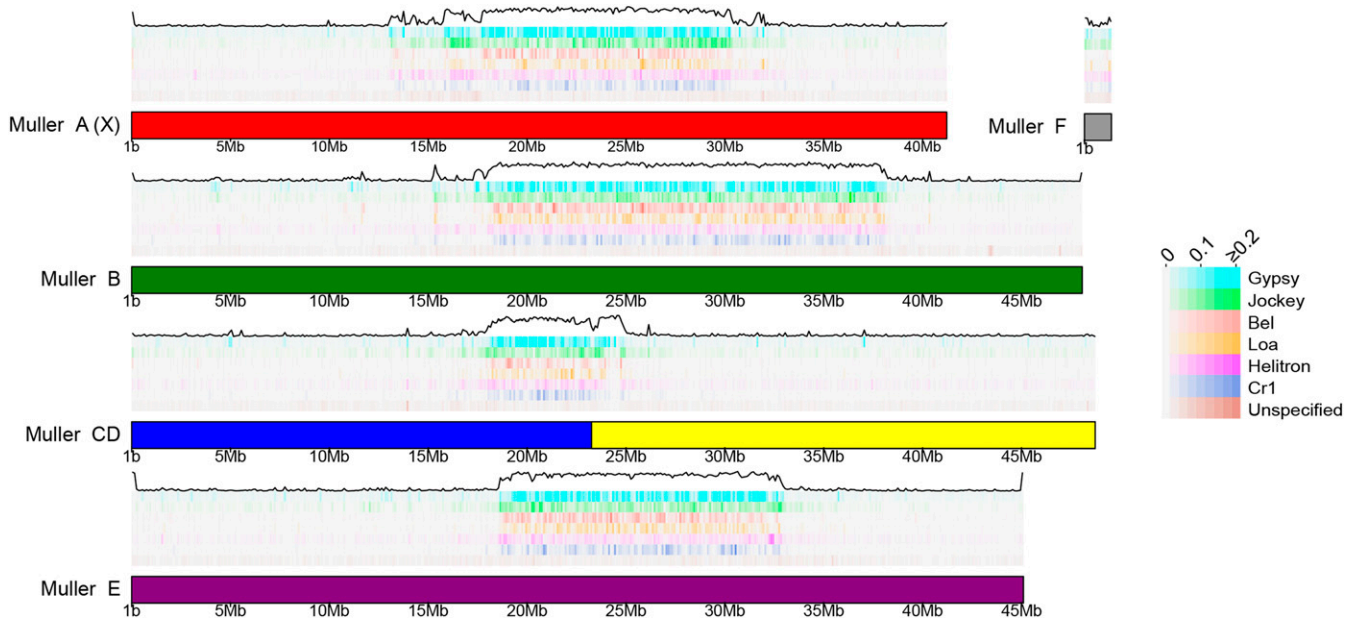

Figure 3 Transposable elements enriched in pericentromeres. Genomic distribution of common transposable element (TE) families in the *D. bifasciata* genome assembly. For each Muller element, TEs are arranged in horizontal tracks of decreasing abundance from top to bottom with the total TE abundance (black line) plotted on top. Shown is the proportion of bases repeat-masked per TE family in 100 kb non-overlapping windows.

short arms of Muller B and E, all of Muller A, and Muller F showed very low levels of variation, consistent with being nearly isogenic (Figure S3). These patterns of elevated diversity are likely driven by chromosomal inversions still present within the sequenced isofemale line.

A total of 57,947,182 bp of the genome assembly was identified as being repetitive (30.1% of the total length of the assembly) and large fractions of all Muller elements were repeat-masked (Figure 2B). The exceptionally high level of repeat-masking located in the middle of chromosome-length scaffolds is indicative of pericentromeric regions on metacentric chromosomes that harbor large numbers of TEs (Kaminker *et al.* 2002, Bracewell *et al.* 2019). Indeed, we find that TEs from a few specific families are highly abundant in the pericentromeric region of all Muller elements (Figure 3). Gypsy and Jockey elements are frequently encountered in the pericentromeres of *D. bifasciata* (Figure 3). Nearly 10 Mb, and over 5 Mb, of assembled sequence (28.4% and 15.8% of all bases masked for TEs) was classified as either Gypsy or Jockey elements, respectively. One specific element, *Daff_Jockey_18*, is at high frequency in all pericentromeres of *D. bifasciata* (Figure S3) and was also the most frequently encountered TE in *D. athabasca*, which also has large metacentric chromosomes (Bracewell *et al.* 2019).

Our MAKER annotation identified a total of 12,821 protein coding genes models in our *D. bifasciata* genome assembly. This number is very similar to other *obscura* groups species, which have been found to harbor anywhere from 12,714 - 14,547 genes (Mahajan *et al.* 2018, Puerma *et al.* 2018, Bracewell *et al.* 2019, Karageorgiou *et al.* 2019). We find a total of 2,279 protein-coding genes on Muller A, 2,499 on Muller B, 4,599 on the fused Muller CD, 3,276 on Muller E and 90 on Muller F (Table 2). Comparisons of orthologs between *D. bifasciata* and *D. athabasca* (*affinis* subgroup), which also has a metacentric Muller A-AD, Muller B, and Muller E indicates that the large pericentromeric region in these species is homologous (Figure 4). Surprisingly, the pericentromeric regions in *D. bifasciata* are remarkably similar in size to those of *Drosophila athabasca* suggesting some level of pericentromere stability over long periods of evolutionary time. Conservation of the pericentromere for Muller A between *D. bifasciata* and *D. athabasca* strongly suggests the fusion between Muller A and D involved a telomere-centromere or telomere-telomere fusion between the metacentric Muller A and the telocentric Muller D. This type of fusion would have resulted in the large neo-X (Muller A-AD) we see in species from the *pseudoobscura*/*affinis* subgroup and would account for the excess of Muller A genes on XR of the fused chromosome. For Muller B and Muller E, we find clear evidence of multiple paracentric inversions that differentiate the *D. bifasciata* and *D. athabasca* chromosomes (Figure 4). However, we find no signatures of pericentric inversions, and each arm of Muller B and E appears to be conserved

Table 2 Genome assembly and annotation results

	Contigs	Length (bp)[a]	Repetitive (%)	Pericentromere (Mb)	Gene models
Muller A (X)	54	41,219,968	32.4	12.6	2,279
Muller B	238	48,071,810	36.6	19.9	2,499
Muller CD	126	48,727,904	15.8	6.8	4,599
Muller E	119	45,099,364	28.2	14.3	3,276
Muller F	1	1,364,133	26.9	NA	90
unplaced	209	8,267,939	76.9	NA	78

[a] Includes Ns introduced from scaffolding Muller elements.

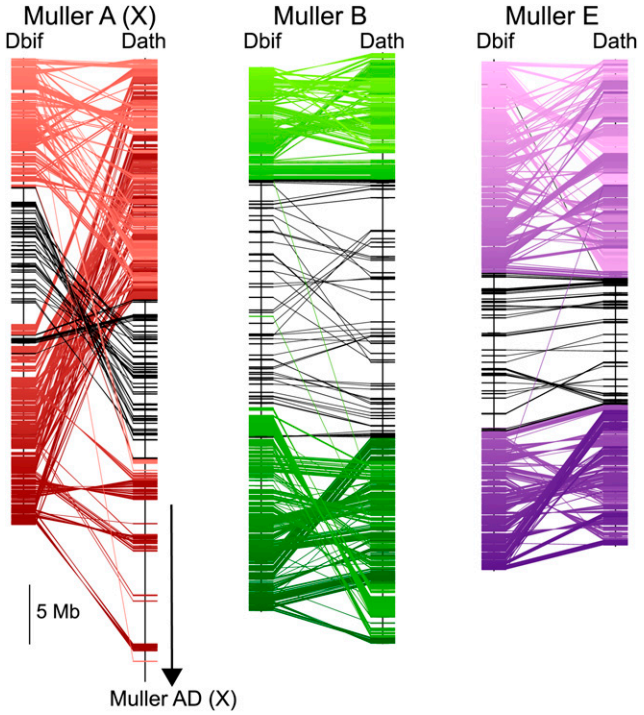

Figure 4 Muller element evolution and synteny. Comparisons of synteny between *D. bifasciata* and *D. athabasca* Muller elements A (X) (red), B (green), and E (purple) with each line representing a protein-coding gene. Genes previously identified as pericentromeric in *D. athabasca* (Bracewell *et al.* 2019) are shown in black. Only Muller A (X) genes shown for *D. athabasca*.

(Figure 4). This pattern contrasts with Muller A where we find evidence of both paracentric and pericentric inversions that differentiate these species (Figure 4).

We also sought to determine the orientation of the fusion between Muller C and D in *D. bifasciata* and we find that the current configuration most likely occurred via a fusion of the two chromosomes at their centromeres. Orthologs of pericentromeric C and D genes in *D. subobscura* are adjacent to one another in our scaffolded assembly (Figure S4) and Hi-C results strongly support this relationship (Figure 2A). Interestingly, the pericentromeric region of the fused C-D chromosome appears smaller than all other pericentromeres in our assembly (Figure 2A). Although speculative, this may be due to the young age of this pericentromere which may be just beginning to expand through the proliferation of repetitive sequences. For example, the 50 pericentromeric C genes in *D. subobscura* are in a 1.0 Mb region while orthologs in *D. bifasciata* are spread out across 4.6 Mb (Figure S4).

In conclusion, our chromosome-level assembly of *D. bifasciata* provides a valuable resource for future work in this species and will allow for more comprehensive comparative genomic analyses of Drosophila. Our genome assembly method highlights how long-read Nanopore sequencing combined with Hi-C scaffolding can assemble long stretches of highly repetitive pericentromeric sequence, resulting in the assembly of entire metacentric chromosomes. These chromosome-level assemblies allow for evolutionary comparisons of pericentromeric regions that until recently have not been possible. As more chromosome-level genome assemblies become available, we will begin to better understand large-scale changes in chromosome morphology and their impact on genome architecture, gene evolution and speciation.

ACKNOWLEDGMENTS

We thank K Wei and D Mai for discussions of repetitive DNA and genome assembly. This work was supported by NIH grants R01GM076007, R01GM101255 and R01GM093182 to D. Bachtrog and 5F32GM123764-02 to R. Bracewell.

LITERATURE CITED

Bickhart, D. M., B. D. Rosen, S. Koren, B. L. Sayre, A. R. Hastie *et al.*, 2017 Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat. Genet. 49: 643–650. https://doi.org/10.1038/ng.3802

Bracewell, R., K. Chatla, M. J. Nalley, and D. Bachtrog, 2019 Dynamic turnover of centromeres drives karyotype evolution in Drosophila. eLife 8. https://doi.org/10.7554/eLife.49002

Buzzati-Traverso, A. A., and R. E. Scossiroli, 1955 The "Obscura Group" of the Genus Drosophila, pp. 47–92 in *Advances in Genetics*, edited by M. Demerec, Academic Press, Cambridge, MA.

Cabanettes, F., and C. Klopp, 2018 D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ 6: e4958. https://doi.org/10.7717/peerj.4958

Campbell, M. S., C. Holt, B. Moore, and M. Yandell, 2014 Genome annotation and curation using MAKER and MAKER-P. Current Protocols in Bioinformatics **48**: 4.11.11–14.11.39. https://doi.org/10.1002/0471250953.bi0411s48

Carvalho, A. B., and A. G. Clark, 2005 Y chromosome of *D. pseudoobscura* is not homologous to the ancestral Drosophila Y. Science 307: 108–110. https://doi.org/10.1126/science.1101675

Chang, C.-H., A. Chavan, J. Palladino, X. Wei, N. M. C. Martins *et al.*, 2019 Islands of retroelements are major components of Drosophila centromeres. PLoS Biol. 17: e3000241. https://doi.org/10.1371/journal.pbio.3000241

Chu, C., R. Nielsen, and Y. Wu, 2016 REPdenovo: Inferring de novo repeat motifs from short sequence reads. PLoS One 11: e0150719. https://doi.org/10.1371/journal.pone.0150719

Cock, P. J. A., J. M. Chilton, B. Grüning, J. E. Johnson, and N. Soranzo, 2015 NCBI BLAST+ integrated into Galaxy. Gigascience 4: 39. https://doi.org/10.1186/s13742-015-0080-7

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The variant call format and VCFtools. Bioinformatics 27: 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

DePristo, M. A., E. Banks, R. Poplin, K. Garimella, J. R. Maguire *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43: 491–498. https://doi.org/10.1038/ng.806

Dudchenko, O., S. S. Batra, A. D. Omer, S. K. Nyquist, M. Hoeger *et al.*, 2017 De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science 356: 92–95. https://doi.org/10.1126/science.aal3327

Durand, N. C., J. T. Robinson, M. S. Shamim, I. Machol, J. P. Mesirov *et al.*, 2016a Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 3: 99–101. https://doi.org/10.1016/j.cels.2015.07.012

Durand, N. C., M. S. Shamim, I. Machol, S. S. P. Rao, M. H. Huntley *et al.*, 2016b Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 3: 95–98. https://doi.org/10.1016/j.cels.2016.07.002

Garrison, E., and G. Marth, 2012 Haplotype-based variant detection from short-read sequencing. arXiv: 1207.3907. https://arxiv.org/abs/1207.3907

Gao, J., H. Watabe, T. Aotsuka, J. Pang, and Y. Zhang, 2007 Molecular phylogeny of the *Drosophila obscura* species group, with emphasis on the Old World species. BMC Evol. Biol. 7: 87. https://doi.org/10.1186/1471-2148-7-87

Gel, B., and E. Serra, 2017 karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics 33: 3088–3090. https://doi.org/10.1093/bioinformatics/btx346

Guy, L., J. Roat Kultima, and S. G. E. Andersson, 2010 genoPlotR: comparative gene and genome visualization in R. Bioinformatics 26: 2334–2335. https://doi.org/10.1093/bioinformatics/btq413

Hill, T., and A. J. Betancourt, 2018 Extensive exchange of transposable elements in the *Drosophila pseudoobscura* group. Mob. DNA 9: 20. https://doi.org/10.1186/s13100-018-0123-6

Kaminker, J. S., C. M. Bergman, B. Kronmiller, J. Carlson, R. Svirskas *et al.*, 2002 The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. Genome Biology **3**: RESEARCH0084.

Karageorgiou, C., V. Gámez-Visairas, R. Tarrío, and F. Rodríguez-Trelles, 2019 Long-read based assembly and synteny analysis of a reference *Drosophila subobscura* genome reveals signatures of structural evolution driven by inversions recombination-suppression effects. BMC Genomics 20: 223. https://doi.org/10.1186/s12864-019-5590-8

Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman *et al.*, 2017 Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27: 722–736. https://doi.org/10.1101/gr.215087.116

Koren, S., A. Rhie, B. P. Walenz, A. T. Dilthey, D. M. Bickhart *et al.*, 2018 De novo assembly of haplotype-resolved genomes with trio binning. Nat. Biotechnol. 36: 1174–1182. https://doi.org/10.1038/nbt.4277

Korf, I., 2004 Gene finding in novel genomes. BMC Bioinformatics 5: 59. https://doi.org/10.1186/1471-2105-5-59

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, H., 2018 Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy *et al.*, 2009 Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326: 289–293. https://doi.org/10.1126/science.1181369

Mahajan, S., K. H. C. Wei, M. J. Nalley, L. Gibilisco, and D. Bachtrog, 2018 De novo assembly of a young Drosophila Y chromosome using single-molecule sequencing and chromatin conformation capture. PLoS Biol. 16: e2006348. https://doi.org/10.1371/journal.pbio.2006348

Mai, D., M. Nalley, and D. Bachtrog, 2020 Patterns of genomic differentiation in the *Drosophila nasuta* species complex. Mol. Biol. Evol. 37: 208–220. https://doi.org/10.1093/molbev/msz215

Miga, K. H., 2019 Centromeric satellite DNAs: hidden sequence variation in the human population. Genes (Basel) 10: 352. https://doi.org/10.3390/genes10050352

Miller, D. E., C. Staber, J. Zeitlinger, and R. S. Hawley, 2018 Highly contiguous genome assemblies of 15 Drosophila species generated using nanopore sequencing. G3: Genes|Genomes|Genetics **8**: 3131–3141.

Moriwaki, D., and O. Kitagawa, 1955 Salivary gland chromosomes of *Drosophila bifasciata*. Cytologia (Tokyo) 20: 247–257. https://doi.org/10.1508/cytologia.20.247

Puerma, E., D. J. Orengo, F. Cruz, J. Gómez-Garrido, P. Librado *et al.*, 2018 The high-quality genome sequence of the oceanic island endemic species *Drosophila guanche* reveals signals of adaptive evolution in genes related to flight and genome stability. Genome Biol. Evol. 10: 1956–1969. https://doi.org/10.1093/gbe/evy135

Putnam, N. H., B. L. O'Connell, J. C. Stites, B. J. Rice, M. Blanchette *et al.*, 2016 Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 26: 342–350. https://doi.org/10.1101/gr.193474.115

Quinlan, A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ramani, V., D. A. Cusanovich, R. J. Hause, W. Ma, R. Qiu *et al.*, 2016 Mapping 3D genome architecture through in situ DNase Hi-C. Nat. Protoc. 11: 2104–2121. https://doi.org/10.1038/nprot.2016.126

Ruan, J., and H. Li, 2019 Fast and accurate long-read assembly with wtdbg2. bioRxiv https://doi.org/doi: 10.1101/530972

Schaeffer, S. W., 2018 Muller "Elements" in Drosophila: how the search for the genetic basis for speciation led to the birth of comparative genomics. Genetics 210: 3–13. https://doi.org/10.1534/genetics.118.301084

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. https://doi.org/10.1093/bioinformatics/btv351

Solares, E. A., M. Chakraborty, D. E. Miller, S. Kalsow, K. Hall *et al.*, 2018 Rapid low-cost assembly of the *Drosophila melanogaster* reference genome using low-coverage, long-read sequencing. G3 (Bethesda) 8: 3143–3154. https://doi.org/10.1534/g3.118.200162

Stanke, M., and S. Waack, 2003 Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19: ii215–ii225. https://doi.org/10.1093/bioinformatics/btg1080

Vaser, R., I. Sovic, N. Nagarajan, and M. Sikic, 2017 Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27: 737–746. https://doi.org/10.1101/gr.214270.116

Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel *et al.*, 2014 Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9: e112963. https://doi.org/10.1371/journal.pone.0112963

Communicating editor: K. Thornton