OXFORD
G3 Genes | Genomes | Genetics

G3, 2023, jkad191
https://doi.org/10.1093/g3journal/jkad191
Advance Access Publication Date: 23 August 2023
Genome Report

Long-read genome assemblies for the study of chromosome expansion: *Drosophila kikkawai, Drosophila takahashii, Drosophila bipectinata*, and *Drosophila ananassae*

Wilson Leung,[1] Nicole Torosin,[2] Weihuan Cao,[2] Laura K. Reed,[3] Cindy Arrigo,[4] Sarah C.R. Elgin,[1] Christopher E. Ellison[2,*]

[1]Department of Biology, Washington University in St. Louis, St. Louis, MO 63130, USA
[2]Department of Genetics and Human Genetics Institute of New Jersey, Rutgers University, Piscataway, NJ 08854, USA
[3]Department of Biological Sciences, The University of Alabama, Tuscaloosa, AL 35487, USA
[4]Department of Biology, New Jersey City University, Jersey City, NJ 07305, USA

*Corresponding author: Department of Genetics and Human Genetics Institute of New Jersey, Rutgers University, 604 Allison Rd, Piscataway, NJ 08854, USA. Email: chris.ellison@rutgers.edu

Flow cytometry estimates of genome sizes among species of *Drosophila* show a 3-fold variation, ranging from ∼127 Mb in *Drosophila mercatorum* to ∼400 Mb in *Drosophila cyrtoloma*. However, the assembled portion of the Muller F element (orthologous to the fourth chromosome in *Drosophila melanogaster*) shows a nearly 14-fold variation in size, ranging from ∼1.3 Mb to >18 Mb. Here, we present chromosome-level long-read genome assemblies for 4 *Drosophila* species with expanded F elements ranging in size from 2.3 to 20.5 Mb. Each Muller element is present as a single scaffold in each assembly. These assemblies will enable new insights into the evolutionary causes and consequences of chromosome size expansion.

Keywords: *Drosophila kikkawai* strain 14028-0561.14; *Drosophila takahashii* strain IR98-3 E-12201; *Drosophila bipectinata* strain 14024-0381.07; *Drosophila ananassae* strain 14024-0371 13

Introduction

Genome size spans several orders of magnitude across eukaryotes: some fungi have genomes less than 10 Mb in size while some plant and protozoan species have genomes larger than 100 Gb (Gregory *et al.* 2007). This size variation is not associated with organismal complexity nor with the number of protein-coding genes, i.e. "the C-value paradox" (Elliott and Gregory 2015). While it is now clear that variation in the abundance of noncoding (usually repetitive) DNA is the major determinant of genome size variation in eukaryotes, many questions remain, including the evolutionary causes and consequences of the proliferation of noncoding DNA (Gregory 2005).

Drosophila genomes have 6 chromosome arms, known as Muller elements A–F, arranged in 4 to 6 chromosomes (Muller 1940). The Muller F element is the ancestral X chromosome across all Dipteran species; however, prior to the most recent common ancestor of *Drosophila*, the F element became an autosome and underwent a large reduction in size (Vicoso and Bachtrog 2013). The *Drosophila melanogaster* Muller F element exhibits characteristics distinct from the other Muller elements in several ways, including a low rate of recombination, late replication, enrichment of the histone modification H3K9me2/3, and high levels of Heterochromatin Protein 1a (HP1a) and Painting of fourth (Pof) (Larsson *et al.* 2004). The *D. melanogaster* F element is thus considered almost entirely heterochromatic, with an approximate size of 5.2 Mb (Locke and McDermid 1993). However, the distal 1.3 Mb contains approximately 80 protein-coding genes that show a range of expression levels similar to genes located in the euchromatic regions of the other autosomes (Riddle and Elgin 2018).

Across the *Drosophila* genus, which is paraphyletic (Finet *et al.* 2021; Suvorov *et al.* 2022), overall genome size is fairly constrained. Flow cytometry suggests the largest genome size (∼400 Mb, *Drosophila cyrtoloma*) is only ∼3-fold larger than the smallest (∼127 Mb, *Drosophila mercatorum*) (Bosco *et al.* 2007; Craddock *et al.* 2016; Gregory 2023). However, as inferred from current genome assemblies, the portion of the *Drosophila* Muller F element containing protein-coding genes shows a remarkable, nearly 14-fold variation in size, ranging from ∼1.3 Mb in the *D. melanogaster* Release 6 assembly to ∼17.8 Mb in the *Drosophila ananassae* dana_caf1 genome assembly (Drosophila 12 Genomes Consortium *et al.* 2007; Schaeffer *et al.* 2008).

Prior work in *D. ananassae* found that proliferation of retrotransposons played a major role in size expansion of the F element (Leung *et al.* 2017). Furthermore, while *D. ananassae* and *D. melanogaster* F element genes share distinct characteristics—including larger coding spans and introns, greater number of coding exons, and lower codon usage bias—these features are exaggerated in *D. ananassae*, potentially due to the F element

Received on 23 May 2023; accepted on 04 August 2023

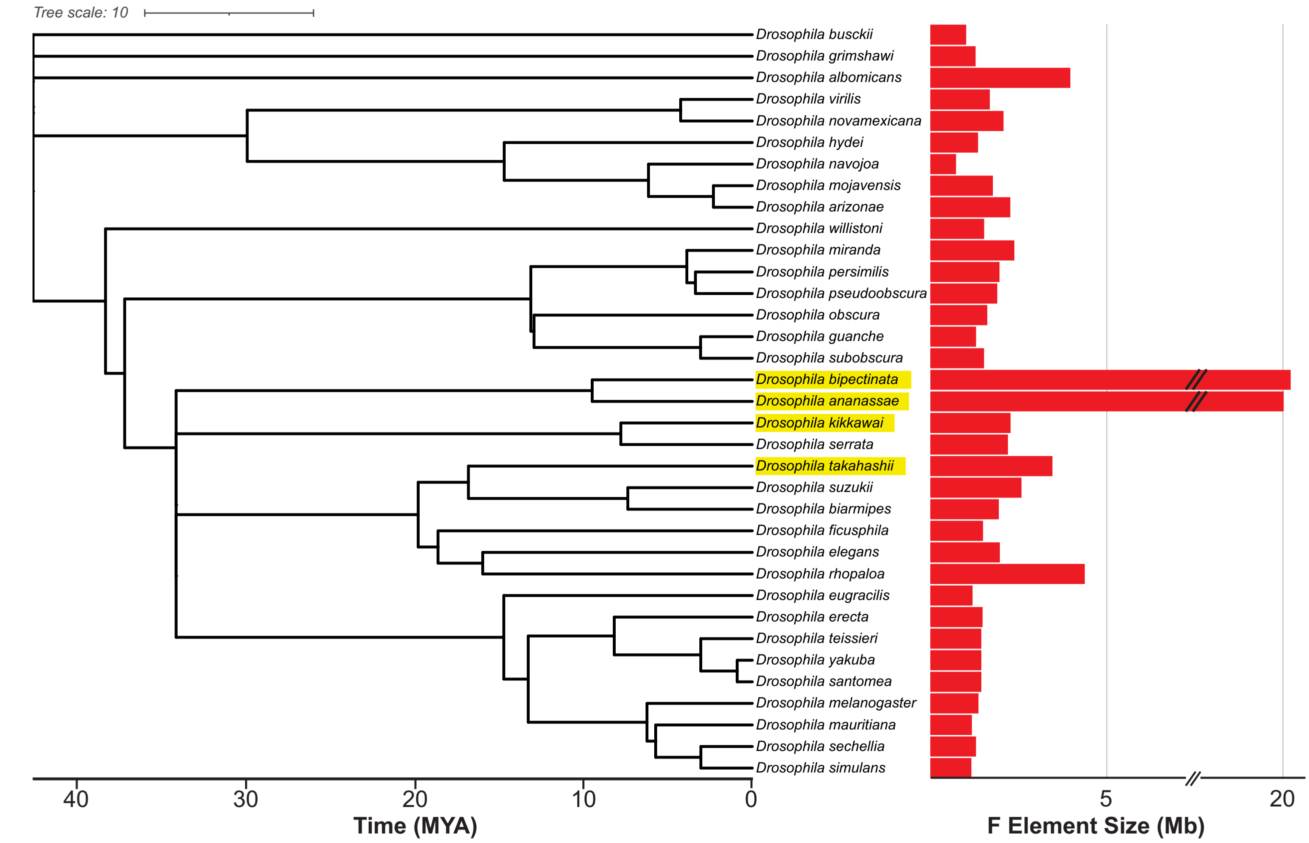

Fig. 1. Size variation of the Muller F element across the *Drosophila* genus. The size of the F element for 35 *Drosophila* species was estimated based on their NCBI RefSeq genome assemblies (see Materials and methods; Supplementary Table 2). The species considered here are highlighted in yellow Tree topology and timescale were obtained from the TimeTree of Life resource (Kumar *et al.* 2022). Note that the F element size x-axis is discontinuous due to space constraints and not all F elements were assembled as a single contig (see Supplementary File 2 for details). Also note that the *Drosophila busckii* F element is fused to the X chromosome and the F element of *Drosophila willistoni* is fused to an autosome (Muller E). Finally, the *Drosophila navojoa* genome assembly is highly fragmented, and thus, the F element size shown here is likely an underestimate.

size expansion (Leung *et al.* 2017). However, these observations were based on the analysis of a small portion of the *D. ananassae* F element (i.e. ~1.4 Mb from 2 scaffolds).

Here, we have used long-read sequencing and Hi-C scaffolding to construct chromosome-level genome assemblies for 3 *Drosophila* species that show different degrees of F element size expansion: *Drosophila takahashii*, *Drosophila kikkawai*, and *Drosophila bipectinata* (Fig. 1. We also perform Hi-C scaffolding for a recently generated long-read *D. ananassae* genome assembly (Tvedte *et al.* 2022). All Muller elements, including the F element, arepresent as single scaffolds in each of these assemblies and thus provide a valuable resource for future studies investigating the causes and consequences of size expansion in this chromosome.

Materials and methods

Software versions

See Supplementary Table 1 for the version information and the references for the bioinformatics tools used in this study.

Muller F size estimation

The F element sizes in each reference sequence (RefSeq) genome assembly were estimated based on the total size of the scaffolds that contain alignments to *D. melanogaster* F element proteins, transcripts, and coding exons. The RefSeq genome assemblies for 35 *Drosophila* species were obtained from NCBI (O'Leary *et al.* 2016). Note that the F element is not assembled into a single scaffold in all species and some of the scaffolds include gaps of unknown sizes (Supplementary Table 2); thus, the size estimates are conservative, and the true F element size may be larger in these species. The *D. melanogaster* proteins, transcripts, and coding exons sequences were obtained from FlyBase (FlyBase release FB2022_03; *D. melanogaster* annotation release 6.46) (Gramates *et al.* 2022). In addition, the *D. melanogaster* release 6 genome assembly was aligned against each RefSeq genome assembly to facilitate the identification of *D. melanogaster* F element genes that might have moved to another Muller element in other *Drosophila* species. The proteins, transcript, coding exons, and whole-genome alignments are available as evidence tracks (i.e. *D. melanogaster* proteins, *D. melanogaster* transcripts, CDS mapping, and *D. melanogaster* net) on the Genomics Education Partnership (GEP) mirror of the UCSC Genome Browser (https://gander.wustl.edu). The RefSeq accession numbers of the scaffolds assigned to the F element are listed in Supplementary File 2.

The *D. melanogaster* proteins were aligned against each RefSeq genome assembly with SPALN (Iwata and Gotoh 2012) with cross-species alignment parameters optimized for *D. melanogaster*: -yX1 -TInsectDm -LS -ya0 -yS100.

D. melanogaster transcripts were aligned against each RefSeq genome assembly using BLAT with the parameters -q=rnax -t=dnax -mask=lower. The transcript alignments were analyzed and filtered by utilities provided by the UCSC Genome Browser (Nassar *et al.* 2023). Transcript alignments were analyzed by pslReps to identify the genomic region with the best alignment having a minimum alignment ratio of 0.25 (-minAli=0.25). The alignments were then filtered by pslCDnaFilter using the following parameters: -minId=0.35 -minCover=0.15 -localNearBest=0.010 -minQSize=20 -ignoreIntrons -repsAsMatch -ignoreNs -bestOverlap.

D. melanogaster coding exons were aligned against each RefSeq genome assembly with the tblastn program provided by WU-BLAST with the following parameters: -e=1e-2 -topComboN=1 -links -hspsepSmax=10000 -hspsepQmax=1000 -matrix=PAM40 -Q=7 -R=2.

The *D. melanogaster* whole-genome assembly was obtained from FlyBase and aligned against each RefSeq genome assembly using LAST (Kiełbasa *et al.* 2011). The whole-genome alignments were then processed using the chain and net alignment protocol and utilities (e.g. axtSort, axtChain, chainAntiRepeat, chainFilter, chainPreNet, chainNet, netSyntenic, netClass, and netFilter) developed by the UCSC Bioinformatics Group (Kent *et al.* 2003).

Strain information

D. kikkawai *strain 14028-0561.14*

The original strain of *D. kikkawai* was obtained from the National *Drosophila* Species Stock Center (NDSSC) at Cornell University and then inbred (11 generations of full-sib crosses) by Professor Artyom Kopp at University of California, Davis (UC Davis). Details regarding the BioSample used for the Pacific Biosciences (PacBio whole-genome sequencing are available under the accession number SAMN33872896 , and the details for the BioSample used to generate the Hi-C data are available under the accession number SAMN34351228.

D. takahashii *strain IR98-3 E-12201*

The original strain of *D. takahashii* was obtained from EHIME-Fly and then inbred (10 generations of full-sib crosses) by Professor Artyom Kopp at UC Davis. Details regarding the BioSample used for the PacBio whole-genome sequencing are available under the accession number SAMN33872897, and the details for the BioSample used to generate the Hi-C data are available under the accession number SAMN34351229.

D. bipectinata *strain 14024-0381.07*

The strain of *D. bipectinata* was obtained from the NDSSC and then kept by the Elgin Lab and Kopp Lab but not inbred. Details regarding the BioSample used for the PacBio whole-genome sequencing are available under the accession number SAMN33872898, and the details for the BioSample used to generate the Hi-C data are available under the accession number SAMN34351230.

D. ananassae *strain 14024-0371.13*

This strain of *D. ananassae* was originally kept by the *Drosophila* Species Stock Center (DSSC) at the University of California, San Diego, and sequenced by the *Drosophila* 12 Genomes Consortium (Drosophila 12 Genomes Consortium *et al.* 2007). The strain of *D. ananassae* used to generate the whole-genome assembly produced by Professor Julie C. Dunning Hotopp at the University of Maryland (Tvedte *et al.* 2021) was treated with tetracycline and is cured of *Wolbachia* (see BioSample SAMN13901672 for details). The strain of *D. ananassae* used to generate the Hi-C data was maintained independently by the Elgin Lab and was not treated with tetracycline (see BioSample SAMN26507075 for details).

In addition to EHIME-Fly and the NDSSC, the strains of *D. kikkawai*, *D. takahashii*, *D. bipectinata*, and *D. ananassae* used in this study are currently available through the Ellison Lab (Rutgers University).

PacBio long-read sequencing

High molecular weight (HMW) DNA from *D. kikkawai* (adult females), *D. takahashii* (adult females), and *D. bipectinata* (adult males and females) flies was provided by Dr. Bernard Kim at Stanford University. An overview of the DNA extraction protocol has previously been described (Kim, Miller, *et al.* 2021; Kim, Wang, *et al.* 2021).

The library preparation and PacBio sequencing were performed by the McDonnell Genome Institute (MGI) at Washington University in St. Louis. The SMRTbell Express Template Prep Kit 2.0, Sequel Binding Kit 3.0, and Sequel Sequencing Kit 3.0 were used to prepare the samples for single-molecule real-time (SMRT) sequencing using the PacBio Sequel system. Each species was sequenced using one 1M SMRT cell in the continuous long-read (CLR) sequencing mode with the 6.0.0.45111 chemistry and a movie length of 600 min. The sequencing data were processed by version 7.0.1.66975 of PacBio SMRT Link.

To assess the quality of the PacBio sequencing data, the quality control (QC) tool in SequelTools (Hufnagel *et al.* 2020) and SEQUELstats were used to analyze the subreads and scraps BAM files for each species.

Assembly of PacBio reads

The PacBio subreads were analyzed by the icecreamfinder.sh script in BBMap to remove adapter sequences and filter potential chimeric reads. Trimmed PacBio subreads that passed the filter and have a read length of at least 5,000 nt were used as input to the Canu assembler (Koren *et al.* 2017) with the genomeSize parameter set to 205 m. The Canu assemblies then underwent 2 rounds of polishing by GCpp (with the Arrow algorithm) using the PacBio subreads. The assemblies from the second round of GCpp were then polished using Illumina genomic reads from each species by POLCA (part of the MaSuRCA assembler; Zimin and Salzberg 2020) and NextPolish (Hu *et al.* 2020). The Illumina genomic reads used for polishing were obtained from the NCBI Sequence Read Archive (SRA) under the accession numbers SRR345537 (*D. kikkawai*), SRR13070706 (*D. takahashii*), and SRR6425989 (*D. bipectinata*).

Assembly of Nanopore reads

Nanopore reads were obtained from the NCBI SRA under the accession numbers SRR13070622 (*D. kikkawai*), SRR13070623 (*D. takahashii*), and SRR13070724 (*D. bipectinata*). Adapter sequences and chimeric reads were identified and removed by Porechop. The trimmed Nanopore reads that passed the default Porechop filters were used as input to the Flye assembler (Kolmogorov *et al.* 2019) with the genome-size parameter set to 205 m. The Nanopore Flye assemblies then underwent 2 rounds of polishing with GCpp using the PacBio subreads from the corresponding species. The assemblies from the second round of GCpp were polished by POLCA and NextPolish using the same set of Illumina reads used to polish the PacBio assemblies.

Assembly merging

For each species, 2 rounds of quickmerge (Chakraborty *et al.* 2016) were used to combine the PacBio assembly produced by Canu with the Nanopore assembly produced by Flye. In the first round of quickmerge, the Nanopore assembly produced by Flye was used as the query and the PacBio assembly produced by Canu was used as the reference to produce the merged assembly. In the second round of quickmerge, the PacBio assembly produced by Canu was used as the query and the merged assembly produced by the first round of quickmerge was used as the reference. The assemblies produced by the second round of quickmerge were then polished by Hapo-G (Aury and Istace 2021) using the genomic Illumina reads that had been used for polishing the PacBio Canu and Nanopore Flye assemblies.

The polished quickmerge assemblies were analyzed by Purge Haplotigs (Roach et al. 2018) to identify haplotigs associated with the primary contigs. Different -low, -mid, and -high parameters were used with the cov command to identify haplotigs based on alignment coverage: *D. kikkawai* (-low 5 -mid 20 -high 95), *D. takahashii* (-low 5 -mid 100 -high 190), and *D. bipectinata* (-low 5 -mid 95 -high 190).

Hi-C scaffolding

All flies were maintained in population cages on molasses agar with yeast paste (https://bdsc.indiana.edu/information/recipes/hardagar.html). Embryos (8–16 h) for each species were collected and dechorionated in 50% commercial bleach for 2.5 min at room temperature. Nuclei were isolated from 250 to 500 mg of embryos and fixed in 1.8% formaldehyde for 15 min according to a previously published protocol (Sandmann et al. 2006). Hi-C libraries were constructed for each species using the in situ DNase Hi-C protocol in (Ramani et al. 2016), and 150-bp paired-end reads were sequenced on an Illumina HiSeq machine. Approximately 116–177 million read pairs were generated for each species.

Hi-C scaffolding was performed using the 3D-DNA pipeline (Dudchenko et al. 2017) with the following parameters: EDITOR_REPEAT_COVERAGE, 6; SPLITTER_COARSE_STRINGENCY, 70; SPLITTER_SATURATION_CENTILE, 7; and SPLITTER_COARSE_RESOLUTION, 50,000. A gap size of 500 bp (–gap_size 500) was added between adjacent contigs produced by quickmerge to construct the sequences in the Hi-C-scaffolded genome assemblies. Contact maps for each chromosome arm were generated using the *hicPlotMatrix* utility from HiCExplorer (Ramírez et al. 2018) with 40-kb bins.

The snail plots used to assess the quality of the Hi-C-scaffolded genome assemblies were generated by BlobTools2 (Challis et al. 2020).

Assembly-level classification

The NCBI Assembly Data Model defines a chromosome-level assembly as follows: "There is sequence for 1 or more chromosomes. This could be a completely sequenced chromosome without gaps or a chromosome containing scaffolds or contigs with gaps between them. There may also be unplaced or unlocalized scaffolds" (https://www.ncbi.nlm.nih.gov/assembly/help/#level). We therefore describe our assemblies as chromosome level despite the fact that they contain gaps as well as unplaced scaffolds.

Comparisons to RefSeq assemblies

For the 3 species with new primary assemblies, the new assembly presented here was compared to its current RefSeq assembly. The RefSeq accession numbers are as follows: *D. bipectinata* (GCF_018153845.1), *D. kikkawai* (GCF_018152535.1), and *D. takahashii* (GCF_018152695.1). Statistics and BUSCO summaries for these RefSeq assemblies are available on the Genome section of the NCBI Datasets webpage accessible via the assembly accession numbers.

Annotation

The diptera_odb10 (release date 2020-08-05) lineage dataset was used with BUSCO (Manni et al. 2021) in "genome" mode to assess the quality of the assembled genomes. The generate_plot.py script provided by BUSCO is used to produce the bar chart of the BUSCO summary results. The locations of the BUSCO matches in the full_table.tsv files are converted into bigBed format for display on the GEP UCSC Genome Browser [available under the "BUSCO (diptera_odb10)" evidence track].

The *D. melanogaster* proteins, transcripts, and coding exons sequences were obtained from FlyBase (FlyBase release FB2022_03; *D. melanogaster* annotation release 6.46) and aligned against the *D. kikkawai*, *D. takahashii*, *D. bipectinata*, and *D. ananassae* Hi-C genome assemblies using the methods described under the "Muller F size estimation" section above. The SPALN alignments of *D. melanogaster* proteins against the Hi-C genome assemblies include the locations where insertions or deletions (indels) will result in frameshifts. The locations of these potential frameshifts are available under the "Potential Frame Shifts" evidence track on the GEP UCSC Genome Browser.

RefSeq gene models from *D. kikkawai*, *D. takahashii*, *D. bipectinata*, and *D. ananassae* were aligned against the corresponding Hi-C genome assembly using BLAT with the following parameters q=rna -fine -minScore=20 -stepSize=5. The transcript alignments were analyzed by pslReps with the parameters -minCover=0.15 -minAli=0.98 -nearTop=0.001 and then filtered by pslCDnaFIlter with the parameters -minId=0.95 -minCover=0.15 -localNearBest=0.001-min QSize=20 -ignoreIntrons -repsAsMatch -ignoreNs -bestOverlap.

De novo repeat libraries were constructed for each species using Earl Grey (Baril et al. 2022). Repeat landscape plots were generated using the *createRepeatLandscape.pl* script from the RepeatMasker package (Smit et al. 2013).

Wolbachia BLAST searches

We used the *wAna* genome (accession number: GCF_008033215.1) and, for NCBI BLAST (Camacho et al. 2009), an *E*-value threshold of 1e−10 and a minimum alignment length of 1,000 bp for *D. ananassae*. For *D. bipectinata*, we used an *E*-value threshold of 1e −5 and a minimum alignment length of 500 bp.

Results and discussion

Long-read sequencing combined with Hi-C scaffolding results in chromosome-level scaffolds

We generated 12.9–13.7 Gb of PacBio CLR reads for each species, which amounts to approximately 65× sequencing coverage based on a genome size of 205 Mb for each species (inferred from flow cytometry; Gregory and Johnston 2008) (Fig. 2a). We compared the distributions of read lengths between species and calculated both the median and N50 values. For *D. bipectinata*, we obtained a median read length of 10.4 kb and an N50 of 20.6 kb. For *D. kikkawai*, the median read length was 13.1 kb and read N50 was 32.1 kb. For *D. takahashii*, the median read length was 10.3 kb and read N50 was 20.7 kb (Fig. 2b and c). Read quality was assessed using SequelTools and SEQUELstats (see Materials and methods; Supplementary Tables 3 and 4).

Previous work has shown assembly contiguity can be significantly improved by merging assemblies generated from different sequencing technologies and/or assembly algorithms (Alhakami et al. 2017). To implement this strategy, we generated 2 assemblies for each species using 2 different long-read sequencing platforms and assemblers: (1) contig assemblies for each species were

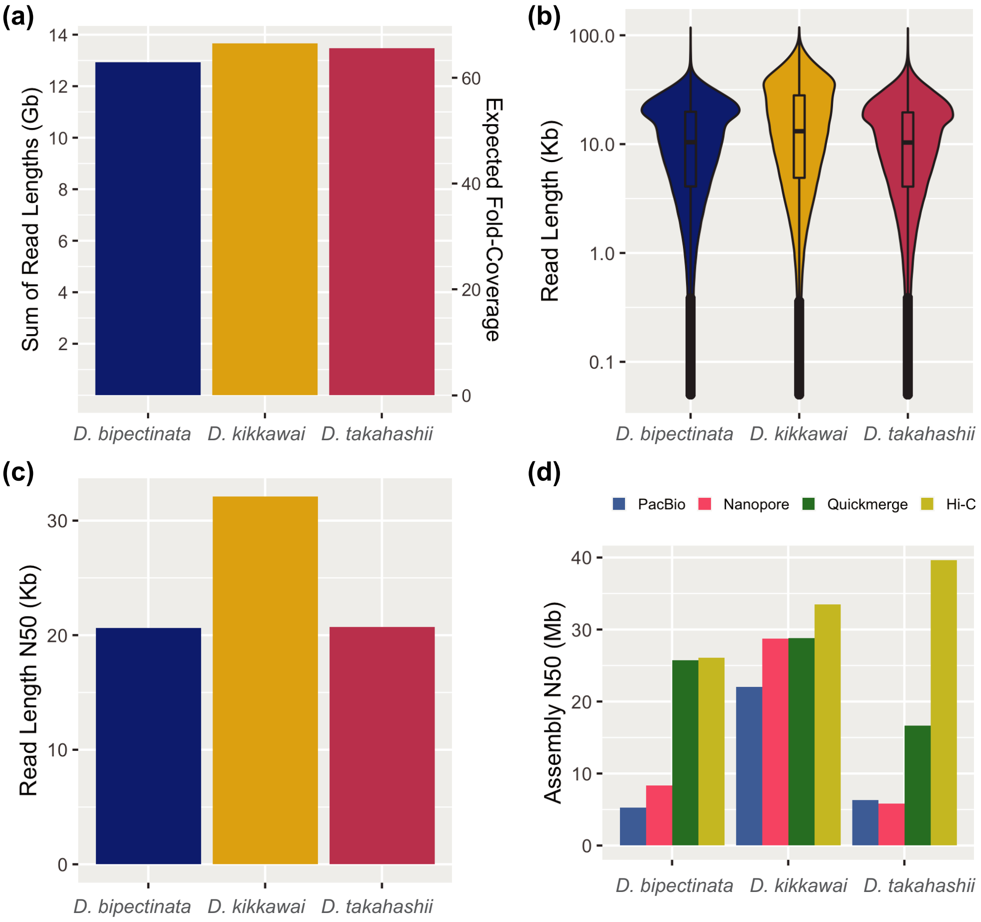

Fig. 2. Summary of sequencing read lengths and assembly contiguity. a) Total base pairs of PacBio long-read sequencing data generated for each species along with the expected sequencing coverage based on a genome size of 205 Mb for each species (estimated from flow cytometry). b) Distribution of PacBio read lengths for each species. c) The read length N50 for each species. Half of the total sequencing data is present in reads of length N50 or larger. d) Assembly N50 metrics for 4 different stages of the assembly pipeline: (1, blue) PacBio-only versions of each assembly were generated using Canu; (2, pink) Oxford Nanopore-only versions of each assembly were generated using Flye; (3, green) PacBio and Nanopore assemblies were merged with quickmerge; and (4, yellow) the merged assemblies were scaffolded using Hi-C data.

generated using the PacBio reads and the *Canu* assembler (Koren *et al.* 2017), which resulted in contig N50 values of 5.3, 22.0, and 6.3 Mb for *D. bipectinata*, *D. kikkawai*, and *D. takahashii*, respectively (Fig. 2d), and (2) contig assemblies were generated from Oxford Nanopore data (produced by Kim, Miller, *et al.* 2021; Kim, Wang, *et al.* 2021) using Flye, which resulted in contig N50 values of 8.4, 28.7, and 5.8 Mb for *D. bipectinata*, *D. kikkawai*, and *D. takahashii*, respectively. We then used quickmerge (Chakraborty *et al.* 2016) to merge our PacBio assemblies produced by Canu with our Nanopore assemblies produced by Flye. The merged assemblies showed improved contiguity with N50 values of 25.7, 28.8, and 16.6 Mb for *D. bipectinata*, *D. kikkawai*, and *D. takahashii*, respectively (Fig. 2d).

We next used Hi-C data to scaffold our contig assemblies with the 3D-DNA pipeline (Dudchenko *et al.* 2017), which resulted in chromosome-level scaffolds for each species (Fig. 3; Supplementary Table 5). We also generated Hi-C data for *D. ananassae*, which has an expanded F element similar in size to that of *D. bipectinata*. We used the *D. ananassae* Hi-C data to scaffold the contigs from a recently published *D. ananassae* long-read genome assembly (Tvedte *et al.* 2021). Note that our use of "chromosome-level" nomenclature is based on the NCBI Assembly Data Model designation (see Materials and methods).

To assess assembly completeness, we used BUSCO (Manni *et al.* 2021) to search for the presence of 3,285 dipteran single-copy orthologs. More than 99% of the single-copy orthologs were found in these assemblies, consistent with a high level of completeness (Fig. 4). The *D. bipectinata* DbipHiC1 assembly has the highest percentage (1.2%) of "complete (C) and duplicated (D)" genes among the 4 species. Examination of the duplicated BUSCO matches in the DbipHiC1 assembly shows that 27% (29/108) of the duplicated BUSCO matches correspond to Histone 2A (69,968 at 7,147) genes located at scaffold_165 and scaffold_175. Both scaffolds contain multiple copies of the histone genes His1, His2A, His2B, His3, and His4, which suggests that the duplicated genes in these scaffolds can be attributed to the histone gene cluster (Supplementary Fig. 1). Although these scaffolds were not detected by the *purge haplotigs* software package (see Materials and methods), it remains possible that they represent 2 different haplotypes from the same locus (i.e. haplotigs), rather than separate histone arrays.

In total, this work has resulted in chromosome-level genome assemblies for 4 *Drosophila* species with expanded F elements (Supplementary Fig. 2). These assemblies show significant improvements in contiguity compared to previous assembly versions (Supplementary Table 6) and each Muller element is now

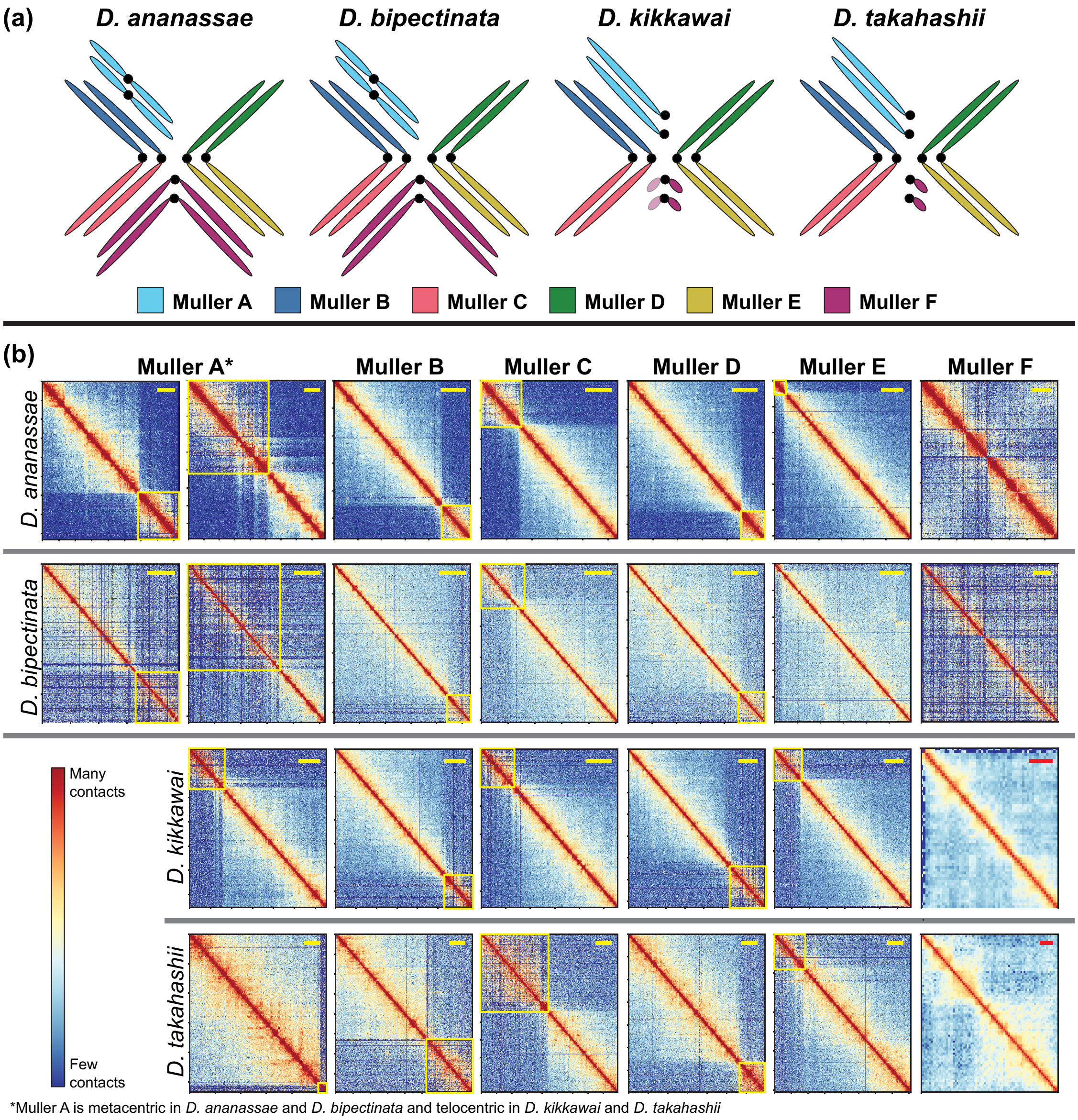

Fig. 3. Muller element Hi-C contact maps. a) The Muller elements A–F correspond to different chromosomes (or chromosome arms) in each species. Muller A is the X chromosome, which is telocentric in *D. melanogaster* but has become metacentric in *D. ananassae* and *D. bipectinata*. The B and C elements are orthologous to the left and right arms of *D. melanogaster* chromosome 2, respectively. The D and E elements are orthologous to the left and right arms of *D. melanogaster* chromosome 3, respectively. Published cytological data show that the larger F elements in *D. ananassae* and *D. bipectinata* are metacentric, while the smaller F element in *D. takahashii* is telocentric, similar to *D. melanogaster* (Deng *et al.* 2007). Previous cytological studies have reported metacentric and telocentric F elements in different populations of *D. kikkawai* (Baimai and Chumchong 1980). The *D. kikkawai* chromosome arm contains 93% (74/80) of the *D. melanogaster* F element genes and appears telocentric in our assembly. However, it is possible that the *D. kikkawai* F element is actually metacentric, but we were unable to assemble the other chromosome arm due to high repeat density. Note that chromosomes are not drawn to scale in this figure. b) Hi-C contact maps are shown for each Muller element (columns), for each species (rows). Yellow boxes show the pericentromeric heterochromatin of each chromosome arm, which is spatially segregated from the euchromatin in the nucleus. The horizontal bars in the upper right corner of each panel are shown for scale: yellow bars represent 5 Mb while red bars represent 400 kb. Note that 2 panels are shown for Muller A in *D. ananassae* and *D. bipectinata* because the chromosome is metacentric in these 2 species. Only 1 panel is shown for Muller A in *D. kikkawai* and *D. takahashii* because the chromosome is telocentric in these 2 species.

present as a single scaffold, which will inform future work related to chromosome structure and evolution.

Chromosome size variation among species

The chromosome-level scaffolds produced by this study allow us to compare the sizes of chromosome arms, including pericentromeric heterochromatin, among species (Fig. 5). The total assembly sizes for each species are 192.2 (*D. ananassae*, excluding a putative Y chromosome scaffold), 194.5 (*D. bipectinata*), 188.5 (*D. kikkawai*), and 198.1 Mb (*D. takahashii*), all close to, but slightly less than, the 205-Mb size estimate from flow cytometry (Gregory and Johnston 2008). The size of the F element scaffold in each species is 19.4 (*D. ananassae*), 20.5 (*D. bipectinata*), 2.3 (*D. kikkawai*), and 3.2 Mb (*D. takahashii*). F element expansion (compared to *D. melanogaster*) in these species therefore ranges from 1.8-fold (*D. kikkawai*) to 15.8-fold (*D. bipectinata*). Interestingly, despite the large increase in size of the F element in *D. ananassae* and *D. bipectinata*, the estimated total genome sizes based on the chromosome-level assemblies are similar across all 4 species (Fig. 5). In fact, the assembled portions of the Muller elements B, C, D, and E are all smaller in

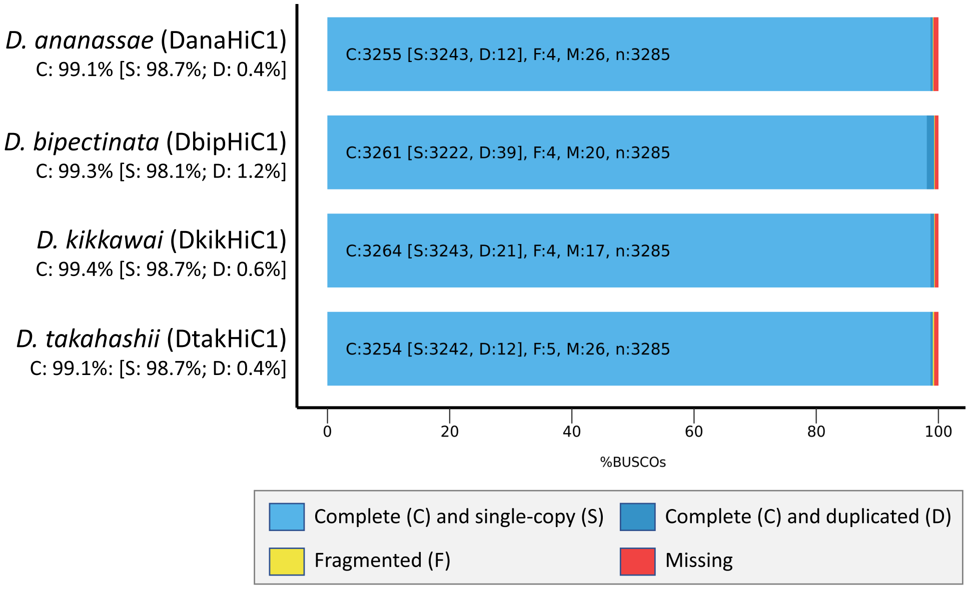

Fig. 4. BUSCO short summary results. BUSCO analysis shows that 99.1%–99.4% of the 3,285 single-copy orthologs in the diptera_odb10 lineage dataset are classified as "complete" in the 4 Hi-C-scaffolded genome assemblies. The percentages of fragmented (F) and missing (M) single-copy orthologs for each assembly are as follows: DanaHiC1 (F: 0.1%; M: 0.8%), DbipHiC1 (F: 0.1%; M: 0.6%), DkikHiC1 (F: 0.1%; M: 0.5%), and DtakHiC1 (F: 0.2%; M: 0.8%).

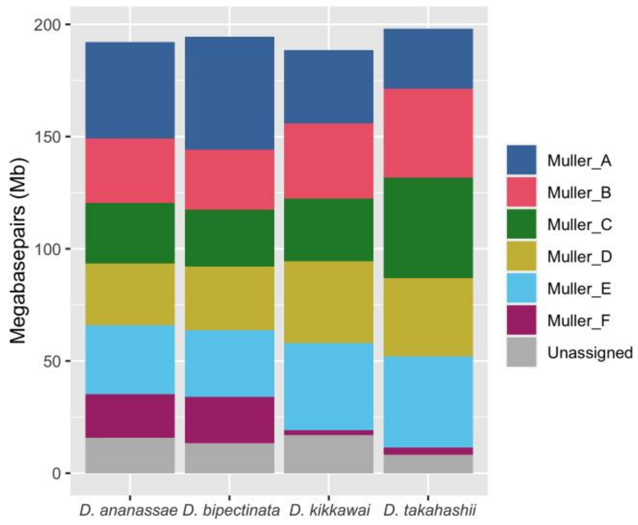

Fig. 5. Scaffold sizes for each species. The size in megabases is shown for each Muller element scaffold in the Hi-C assemblies for the 4 *Drosophila* species. Scaffolds that were not assigned to Muller elements are grouped together in the "unassigned" category. Note that a putative Y chromosome scaffold from *D. ananassae* is not included in the metrics shown here. For each species, the combined height of the colored boxes is equal to the assembly size.

both *D. ananassae* and *D. bipectinata* compared to *D. kikkawai* and *D. takahashii* (Fig. 5). Note, however, that these genome size estimates are derived from the sizes of the assembled scaffolds, and they could be confounded by differences in the number of sequences that cannot be assembled or scaffolded due to high repeat content in the four *Drosophila* species. Thus, in all cases, these are minimal estimates of Muller element size.

Table 1. SPALN alignments show that 86–93% of the protein-coding genes from *D. melanogaster* annotated release 6.46 (which consists of 30,799 proteins derived from 13,986 genes) can be placed in the Hi-C-scaffolded genome assemblies.

Species (assembly)	Number of genes with alignments	Number of isoform with alignments	Estimated number of frameshifts
D. kikkawai (DkikHiC1)	12,059 (86%)	27,997 (91%)	507
D. takahashii (DtakHiC1)	12,045 (86%)	27,970 (91%)	503
D. bipectinata (DbipHiC1)	12,119 (87%)	28,097 (91%)	512
D. ananassae (DanaHiC1)	13,037 (93%)	29,456 (96%)	552

Annotation of genes and repetitive elements

Students participating in the GEP will manually construct gene models for the F element and for a region near the base of the D element for the 4 *Drosophila* species discussed here using RNA-sequencing (RNA-Seq) data, computational gene predictions, and sequence similarity to *D. melanogaster* genes.

The high error rate of PacBio CLR sequencing data can lead to consensus errors in the resulting assembly. When found within gene coding sequences, these consensus errors can cause artifactual frameshift mutations, which decrease the accuracy of computational gene predictions. As part of the assessment of the quality of the Hi-C-scaffolded genome assemblies, 30,799 proteins from 13,986 genes in *D. melanogaster* annotation release 6.46 provided by FlyBase (Gramates *et al.* 2022) were aligned against each genome assembly. Between 91 and 96% of the *D. melanogaster* isoforms have at least 1 SPALN alignment in the Hi-C-scaffolded genome assemblies, accounting for 86 to 93% of the *D. melanogaster* genes (Table 1). The number of frameshifts in the SPALN protein alignments ranges from 507 to 552, representing ~4% of genes. The number of frameshifts is similar across all 4 assemblies, despite the fact that the *D. ananassae* contigs were generated from high-accuracy PacBio HiFi sequencing data (Tvedte *et al.* 2021). This comparison suggests that our *D. bipectinata*, *D. kikkawai*, and *D. takahashii* assemblies do not suffer from a high rate of consensus errors, despite being generated from lower accuracy PacBio CLR data.

Table 2. SPALN alignments shows that 91–96% of the F element protein-coding genes from *D. melanogaster* annotated release 6.46 (which consists of 298 proteins derived from 80 genes) can be placed in the Hi-C-scaffolded genome assemblies.

Species (assembly)	Number of genes with alignments	Number of isoforms with alignments	Estimated number of frameshifts
D. kikkawai (DkikHiC1)	74 (93%)	287 (96%)	5
D. takahashii (DtakHiC1)	77 (96%)	293 (98%)	1
D. bipectinata (DbipHiC1)	73 (91%)	281 (94%)	1
D. ananassae (DanaHiC1)	73 (91%)	278 (93%)	4

In the *D. melanogaster* annotation release 6.46, the F element has 298 isoforms derived from 80 genes. Between 93 and 98% of the *D. melanogaster* F element isoforms have at least 1 SPALN alignment in the Hi-C-scaffolded genome assemblies, accounting for the 91 to 96% of the F element genes (Table 2). The number of frameshifts in the protein alignments to F element genes ranges from 1 to 5.

We used Earl Grey (Baril *et al.* 2022) to create de novo repeat libraries for each species. The number of families identified and sum of consensus lengths for each species are as follows: in *D. ananassae*, 1,118 families sum to 4.5 Mb; in *D. bipectinata*, 1,014 families sum to 3.8 Mb; in *D. kikkawai*, 924 families sum to 3.5 Mb; and in *D. takahashii*, 1,066 families sum to 3.5 Mb. We then used RepeatMasker (Smit *et al.* 2013) to identify the locations of each repeat family within their respective genome assemblies. Using our custom repeat libraries, RepeatMasker masked a total of 88.8 Mb (41.5%) of the *D. ananassae* genome assembly, 69.3 Mb (35.6%) of the *D. bipectinata* genome assembly, 58.6 Mb (31.1%) of the *D. kikkawai* genome assembly, and 60.3 Mb (30.4%) of the *D. takahashii* genome assembly (Supplementary Fig. 3). We then used the sequence divergence among individual insertions from the same repeat family, along with the percentage of the genome occupied by each family, to visualize the repeat landscape for each species (Fig. 6).

Contribution of *Wolbachia* to F element expansion

Previous work has suggested that lateral transfer of DNA from the *Wolbachia* endosymbiont into the nuclear genome of *D. ananassae* has contributed to the size expansion of the F element in this species (Klasson *et al.* 2014). A recent study constructed a long-read genome assembly (accession number: GCF_017639315.1) from a strain of *D. ananassae* that was treated with tetracycline and cured of *Wolbachia* infections. The study showed that approximately 4.9 Mb of Wolbachia sequences have integrated into the *D. ananassae* genome (Tvedte *et al.* 2022), likely in the F element (Tvedte *et al.* 2021).

Our Hi-C data allowed us to place many contigs within chromosome-level scaffolds. We therefore sought to determine whether the Hi-C scaffolding process allowed the *Wolbachia* sequences previously identified in the *D. ananassae* assembly to be assigned to 1 or more Muller element scaffolds. We also investigated whether *Wolbachia* sequence was present within our *D. bipectinata* assembly, which shows a level of F element expansion similar to that of *D. ananassae*. We used BLAST (Camacho *et al.* 2009) to search the complete *wAna Wolbachia* genome assembly (accession number: GCF_008033215.1) against the Hi-C-scaffolded *D. ananassae* assembly reported here. This BLAST search identified a similar amount of *Wolbachia* sequence as the previous study (4.97 Mb). However, all of the significant BLAST hits to *wAna* were on scaffolds that could not be assigned to *D. ananassae* Muller elements using the Hi-C data; in particular, there were no significant BLAST hits to the *D. ananassae* F element scaffold.

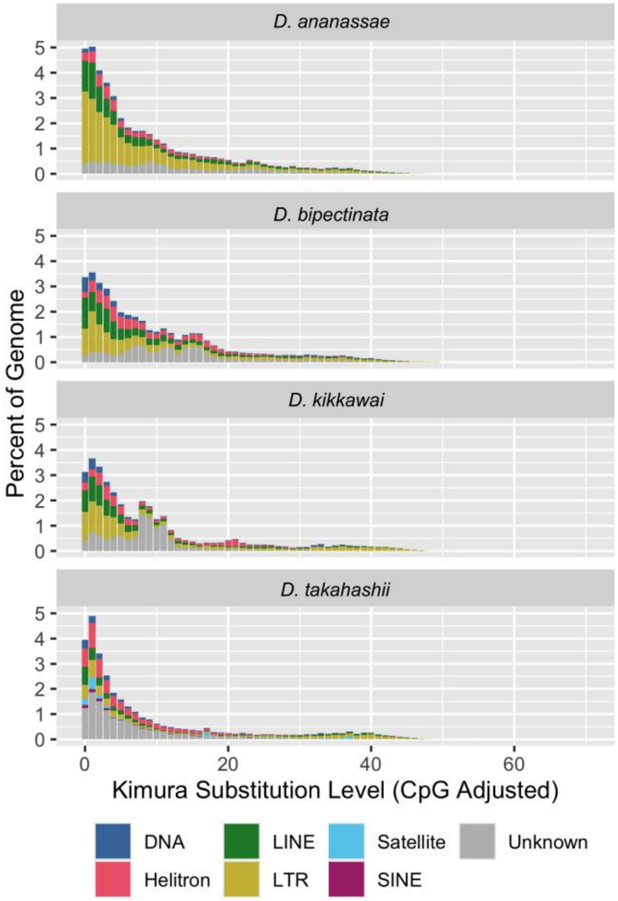

Fig. 6. Repeat landscape plots. Repeat landscapes were generated from de novo repeat libraries created for each species using the assemblies reported here. The x-axis shows the sequence divergence among individual copies from the same repeat family, corrected using the Kimura 2 parameter model. Each bar summarizes the percent of the genome occupied by each repeat superfamily/subclass for a given divergence level.

We also performed a BLAST search of the *wAna* genome assembly against the Hi-C-scaffolded *D. bipectinata* assembly, using less stringent parameters to account for the possibility that the *D. bipectinata* nuclear genome contains DNA from a different *Wolbachia* subtype. In contrast to the *D. ananassae* results, we only identify ∼19 kb of sequence in the *D. bipectinata* assembly that matches the *wAna* genome. All of the matches are located in scaffolds that could not be assigned to the *D. bipectinata* Muller elements. Our *D. bipectinata* strain was not treated with antibiotics before DNA extraction; thus, it remains unclear whether the 19 kb of *Wolbachia* sequence is integrated into the *D. bipectinata* genome or derived from the endosymbiont itself.

Collectively, our analysis of the Hi-C-scaffolded assemblies cannot rule out potential contributions of horizontal transfer of *Wolbachia* DNA to the expansion of the *D. ananassae* F element. However, our results strongly suggest that horizontal transfer of *Wolbachia* DNA is not a major contributor to the expansion of the *D. bipectinata* F element. Furthermore, incorporation of

Wolbachia DNA would explain, at most, ~20% of the expansion of the Muller F element in *D. ananassae*, which means that ~80% of the size increase is due to other factors, such as accumulation of mobile DNA and other repeats.

Data availability

See Supplementary File 1 for the list of supplementary tables, figures, and files associated with this study. The PacBio sequencing data are available through the NCBI BioProject database under the accession number PRJNA948012, and the Hi-C data are available through accession numbers PRJNA961071 and PRJNA967347. The Hi-C-scaffolded genome assemblies for *D. bipectinata*, *D. takahashii*, *D. kikkawai*, and *D. ananassae* have been deposited at GenBank under the accession numbers JARPSB000000000, JARPSC000000000, JARPSD000000000, and JASIRA000000000, respectively. The versions of the *D. bipectinata*, *D. takahashii*, *D. kikkawai*, and *D. ananassae* genome assemblies described in this paper are JARPSB010000000, JARPSC010000000, JARPSD010000000, and JASIRA010000000, respectively. Gene annotations for each assembly in GFF3 format are available via figshare under the DOI: 10.6084/m9.figshare.23737671. The genome assemblies and evidence tracks described in this manuscript are displayed on the GEP UCSC Genome Browser, available through the links under the "UCSC Genome Browser" column of the "Hi-C Genome Assemblies for F Element Expansion Project" landing page. These Genome Browsers include additional gene predictions, repeat analysis, and RNA-Seq evidence tracks that will be used in the subsequent comparative analyses of the expansion of the F elements.

Supplemental material available at G3 online

Acknowledgments

The authors would like to thank Professor Artyom Kopp (UC Davis) for providing the flies used for the PacBio sequencing and for clarifying the history of the strains of *D. kikkawai*, *D. takahashii*, and *D. bipectinata* used by the different sequencing projects. We would also like to thank Dr. Bernard Kim (Stanford University) for providing the high molecular weight DNA used for the PacBio sequencing and the McDonnell Genome Institute (MGI) for performing the PacBio sequencing. The authors acknowledge the Office of Advanced Research Computing (OARC) at Rutgers, The State University of New Jersey for providing access to the Amarel cluster and associated research computing resources that have contributed to the results reported here.

Funding

This material is based upon work supported by the National Science Foundation (NSF) under grant no. 2114661 to CJA at New Jersey City University and the National Institutes of Health (NIH) under grant number R01GM130698 to CE and Fellowship F32GM140669 to NT. Funding for the PacBio sequencing was supported by NSF under grant no. 1915544 and the National Institute of General Medical Sciences of the National Institutes of Health under award number R25GM130517 to LKR. at The University of Alabama for the Genomics Education Partnership (GEP).

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of interest

The authors declare no conflict of interest.

Literature cited

Alhakami H, Mirebrahim H, Lonardi S. A comparative evaluation of genome assembly reconciliation tools. Genome Biol. 2017;18(1):93. doi:10.1186/s13059-017-1213-3.

Aury J-M, Istace B. Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads. NAR Genom Bioinform. 2021;3(2):lqab034. doi:10.1093/nargab/lqab034.

Baimai V, Chumchong C. Karyotype variation and geographic distribution of the three sibling species of the *Drosophila kikkawai* complex. Genetica. 1980;54(2):113–120. doi:10.1007/BF00055979.

Baril T, Imrie RM, Hayward A. Earl Grey: a fully automated user-friendly transposable element annotation and analysis pipeline: In Review preprint. 2022.

Bosco G, Campbell P, Leiva-Neto JT, Markow TA. Analysis of *Drosophila* species genome size and satellite DNA content reveals significant differences among strains as well as between species. Genetics. 2007;177(3):1277–1290. doi:10.1534/genetics.107.075069.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10(1):421. doi:10.1186/1471-2105-10-421.

Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. 2016;44(19):e147. doi:10.1093/nar/gkw654.

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolkit—interactive quality assessment of genome assemblies. G3 (Bethesda). 2020;10(4):1361–1374. doi:10.1534/g3.119.400908.

Drosophila 12 Genomes Consortium; Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, *et al*. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature. 2007;450(7167):203–218. doi:10.1038/nature06341.

Craddock EM, Gall JG, Jonas M. Hawaiian *Drosophila* genomes: size variation and evolutionary expansions. Genetica. 2016;144(1):107–124. doi:10.1007/s10709-016-9882-5.

Deng Q, Zeng Q, Qian Y, Li C, Yang Y. Research on the karyotype and evolution of *Drosophila melanogaster* species group. J Genet Genomics. 2007;34(3):196–213. doi:10.1016/S1673-8527(07)60021-6.

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, *et al*. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356(6333):92–95. doi:10.1126/science.aal3327.

Elliott TA, Gregory TR. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philos Trans R Soc Lond B Biol Sci. 2015;370(1678):20140331. doi:10.1098/rstb.2014.0331.

Finet C, Kassner VA, Carvalho AB, Chung H, Day JP, Day S, Delaney EK, De Ré FC, Dufour HD, Dupim E, *et al*. Drosophyla: resources for *Drosophilid* phylogeny and systematics. Genome Biol Evol. 2021;13(8):evab179. doi:10.1093/gbe/evab179.

Gramates LS, Agapite J, Attrill H, Calvi BR, Crosby MA, dos Santos G, Goodman JL, Goutte-Gattat D, Jenkins VK, Kaufman T, *et al*. Flybase: a guided tour of highlighted features. Genetics. 2022;220(4):iyac035. doi:10.1093/genetics/iyac035.

Gregory TR. The C-value enigma in plants and animals: a review of parallels and an appeal for partnership. Ann Bot. 2005;95(1):133–146. doi:10.1093/aob/mci009.

Gregory TR. Animal Genome Size Database. [Accessed 2023 May 9]. 2023. http://www.genomesize.com/.

Gregory TR, Johnston JS. Genome size diversity in the family Drosophilidae. Heredity (Edinb). 2008;101(3):228–238. doi:10.1038/hdy.2008.49.

Gregory TR, Nicol JA, Tamm H, Kullman B, Kullman K, Leitch IJ, Murray BG, Kapraun DF, Greilhuber J, Bennett MD. Eukaryotic genome size databases. Nucleic Acids Res. 2007;35(Database):D332–D338. doi:10.1093/nar/gkl828.

Hu J, Fan J, Sun Z, Liu S. Nextpolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics. 2020;36(7):2253–2255. doi:10.1093/bioinformatics/btz891.

Hufnagel DE, Hufford MB, Seetharam AS. SequelTools: a suite of tools for working with PacBio sequel raw sequence data. BMC Bioinformatics. 2020;21(1):429. doi:10.1186/s12859-020-03751-8.

Iwata H, Gotoh O. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. Nucleic Acids Res. 2012;40(20):e161. doi:10.1093/nar/gks708.

Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A. 2003;100(20):11484–11489. doi:10.1073/pnas.1932072100.

Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011;21(3):487–493. doi:10.1101/gr.113985.110.

Kim BY, Miller DE, Wang JR, 2021 DNA extraction and Nanopore library prep from 15–30 whole flies V.1. protocols.io. Published July 15, 2021. [Accessed 2023 May 3]. https://dx.doi.org/10.17504/protocols.io.bdfqi3mw.

Kim BY, Wang JR, Miller DE, Barmina O, Delaney E, Thompson A, Comeault AA, Peede D, D'Agostino ERR, Pelaez J, *et al.* Highly contiguous assemblies of 101 drosophilid genomes. eLife. 2021;10:e66405. doi:10.7554/eLife.66405.

Klasson L, Kumar N, Bromley R, Sieber K, Flowers M, Ott SH, Tallon LJ, Andersson SGE, Dunning Hotopp JC. Extensive duplication of the *Wolbachia* DNA in chromosome four of *Drosophila ananassae*. BMC Genomics. 2014;15(1):1097. doi:10.1186/1471-2164-15-1097.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37(5):540–546. doi:10.1038/s41587-019-0072-8.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27(5):722–736. doi:10.1101/gr.215087.116.

Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, Stecher G, Hedges SB. Timetree 5: an expanded resource for species divergence times. Mol Biol Evol. 2022;39(8):msac174. doi:10.1093/molbev/msac174.

Larsson J, Svensson MJ, Stenberg P, Mäkitalo M. Painting of fourth in genus *Drosophila* suggests autosome-specific gene regulation. Proc Natl Acad Sci U S A. 2004;101(26):9728–9733. doi:10.1073/pnas.0400978101.

Leung W, Shaffer CD, Chen EJ, Quisenberry TJ, Ko K, Braverman JM, Giarla TC, Mortimer NT, Reed LK, Smith ST, *et al.* Retrotransposons are the major contributors to the expansion of the *Drosophila ananassae* Muller F element. G3 (Bethesda). 2017;7(8):2439–2460. doi:10.1534/g3.117.040907.

Locke J, McDermid HE. Analysis of *Drosophila* chromosome 4 using pulsed field gel electrophoresis. Chromosoma. 1993;102(10):718–723. doi:10.1007/BF00650898.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021;38(10):4647–4654. doi:10.1093/molbev/msab199.

Muller HJ. Bearings of the "*Drosophila*" work on systematics. In: Huxley J, editor. The New Systematics. Oxford: Clarendon Press; 1940. p. 185–268.

Nassar LR, Barber GP, Benet-Pagès A, Casper J, Clawson H, Diekhans M, Fischer C, Gonzalez JN, Hinrichs AS, Lee BT, *et al.* The UCSC genome browser database: 2023 update. Nucleic Acids Res. 2023;51(D1):D1188–D1195. doi:10.1093/nar/gkac1072.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–D745. doi:10.1093/nar/gkv1189.

Ramani V, Cusanovich DA, Hause RJ, Ma W, Qiu R, Deng X, Blau CA, Disteche CM, Noble WS, Shendure J, *et al.* Mapping 3D genome architecture through in situ DNase Hi-C. Nat Protoc. 2016;11(11):2104–2121. doi:10.1038/nprot.2016.126.

Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, Habermann B, Akhtar A, Manke T. High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat Commun. 2018;9(1):189. doi:10.1038/s41467-017-02525-w.

Riddle NC, Elgin SCR. The *Drosophila* dot chromosome: where genes flourish amidst repeats. Genetics. 2018;210(3):757–772. doi:10.1534/genetics.118.301146.

Roach MJ, Schmidt SA, Borneman AR. Purge haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19(1):460. doi:10.1186/s12859-018-2485-7.

Sandmann T, Jakobsen JS, Furlong EEM. ChIP-on-chip protocol for genome-wide analysis of transcription factor binding in *Drosophila melanogaster* embryos. Nat Protoc. 2006;1(6):2839–2855. doi:10.1038/nprot.2006.383.

Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O'Grady PM, Rohde C, Valente VLS, Aguadé M, Anderson WW, *et al.* Polytene chromosomal maps of 11 *Drosophila* species: the order of genomic scaffolds inferred from genetic and physical maps. Genetics. 2008;179(3):1601–1655. doi:10.1534/genetics.107.086074.

Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. 2013.

Suvorov A, Kim BY, Wang J, Armstrong EE, Peede D, D'Agostino ERR, Price DK, Waddell PJ, Lang M, Courtier-Orgogozo V, *et al.* Widespread introgression across a phylogeny of 155 *Drosophila* genomes. Curr Biol. 2022;32(1):111–123.e5. doi:10.1016/j.cub.2021.10.052.

Tvedte ES, Gasser M, Sparklin BC, Michalski J, Hjelmen CE, Johnston JS, Zhao X, Bromley R, Tallon LJ, Sadzewicz L, *et al.* Comparison of long-read sequencing technologies in interrogating bacteria and fly genomes. G3 (Bethesda). 2021;11(6):jkab083. doi:10.1093/g3journal/jkab083.

Tvedte ES, Gasser M, Zhao X, Tallon LJ, Sadzewicz L, Bromley RE, Chung M, Mattick J, Sparklin BC, Dunning Hotopp JC. Accumulation of endosymbiont genomes in an insect autosome followed by endosymbiont replacement. Curr Biol. 2022;32(12):2786–2795.e5. doi:10.1016/j.cub.2022.05.024.

Vicoso B, Bachtrog D. Reversal of an ancient sex chromosome to an autosome in *Drosophila*. Nature. 2013;499(7458):332–335. doi:10.1038/nature12235.

Zimin AV, Salzberg SL. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. PLoS Comput Biol. 2020;16(6):e1007981. doi:10.1371/journal.pcbi.1007981.

Editor: K. Vogel