Topologically associating domains and their role in the evolution of genome structure and function in *Drosophila*

Yi Liao[1,*], Xinwen Zhang[1], Mahul Chakraborty[1] and J.J. Emerson[1,2,*]

[1]Department of Ecology and Evolutionary Biology, University of California, Irvine, CA, USA.

[2]Center for Complex Biological Systems, University of California, Irvine, CA, USA.

Key Words: *Drosophila*, Topologically associating domains, Evolution, Genome synteny, Structural variants, Gene regulation.

Corresponding Authors:

Yi Liao

University of California Irvine

Department of ecology and evolutionary biology

5427 McGaugh Hall, Irvine, CA 92697 Email: liaoy12@uci.edu

J. J. Emerson

University of California Irvine

Department of ecology and evolutionary biology

Center for complex biological systems

321 steinhaus Hall, Irvine, CA 92697 Telephone: (949) 824-9527

Email: jje@uci.edu

Running Title: TAD evolution in *Drosophila*

ABSTRACT

Topologically associating domains (TADs) are regarded as functional and structural units of higher-order spatial genome organization of many eukaryotic genomes. However, our knowledge of how evolution affects TADs remains limited. To decipher the evolutionary significance of TADs, we *de novo* assembled the genome of *D. pseudoobscura* and created a high-resolution (~800 bp) Hi-C contact map to annotate TADs. Remarkably, more than 40% of TADs between *D. pseudoobscura* and *D. melanogaster* are conserved, despite extensive chromosomal rearrangements in the ~49 million years since they shared a common ancestor. Comparison of 17 diverse *Drosophila* species genomes revealed enrichment of genome rearrangement breakpoints at the TAD boundaries but depletion of such breaks inside the TADs themselves. We show that conservation of TADs is associated with gene expression stability across tissues. Surprisingly, despite being larger mutational targets, a substantial proportion of long (>50kb) genes in *D. melanogaster* (42%) and *D. pseudoobscura* (26%) are individually spanned by complete TADs, implying the formation and maintenance of TADs via 3D cis-regulatory interactions commonly found within long genes. Using high-confidence genome-wide structural variant datasets from 14 *D. melanogaster* strains, its 3 closest sibling species from the *D. simulans* species complex, and two obscura clade species, we show evidence of natural selection operating on structural variants at the TAD boundaries, but with the nature of selection differing between the SV types. Deletions are significantly depleted at TAD boundaries for both divergent and polymorphic SVs, suggesting that deletions at TAD boundaries are under purifying selection, whereas divergent duplications are enriched at the TAD boundaries, possibly pointing to positive selection. Our results offer novel insights into the evolutionary role and maintenance of TADs and their significance in genome structure evolution and gene regulation.

INTRODUCTION

Eukaryotic genomes are organized in a hierarchical fashion, ranging from DNA loops to chromatin domains to compartments (Finn and Misteli 2019; Rowley and Corces 2018). This spatial organization plays crucial roles in genome function and cellular processes such as DNA replication (Pope et al. 2014; Marchal et al. 2019), transcription (Schoenfelder and Fraser 2019), DNA-damage repair (Schmitt et al. 2016), development and cell differentiation (Zheng and Xie 2019). Topologically associating domain (TADs), one of these organizational features, was originally discovered in Hi-C contact maps as domains within which DNA sequences physically contact each other more densely than they are outside (Sexton et al. 2012; Dixon et al. 2012). TADs or similar domains have been widely observed across the kingdom of life, from yeast (Mizuguchi et al. 2014) and bacteria (Le et al. 2013) to plants (Liu et al. 2017; Xie et al. 2019) and animals (Fishman et al. 2019).

One important role of TADs in genome function is their role in regulation of gene expression by promoting and constraining long-range enhancer-promoter interactions (Schoenfelder and Fraser 2019). Many intriguing examples came from the facts that reorganization of spatial genome are associated with gene expression in development (Lupiáñez et al. 2015; Bonev et al. 2017). Additional insights from interspecies comparisons demonstrated that evolutionary conserved TADs are associated with stability of gene expression (Krefting et al. 2018) and 3D genome reorganization may contribute to gene regulatory evolution (Eres et al. 2019). However, the role of TADs in modulating gene expression has recently been challenged by studies that reported that gene regulation is not strongly coupled to genome topology (Ghavi-Helm et al. 2019; Despang et al. 2019). On the other hand, a reversal of the relationships above has been proposed. Namely, transcription may affect genome topology, at least at a fine scale (van Steensel and Furlong 2019). Thus, the relationship between gene transcription and spatial genome structure is in need of further exploration.

TADs are regarded as basic structural units of chromosome organization (Szabo et al. 2018). The disruption of TADs by changes to genome structure (i.e. deletions, duplications, inversions and translocations) might contribute to changes in gene regulation, including the dysregulation of disease-specific genes (Kim et al. 2019). Indeed, such changes have been observed in human cancer genomes (Akdemir et al. 2020). How SVs affect the spatial genome and the potential regulatory effect are comprehensively reviewed elsewhere (Shanta et al. 2020; Spielmann et al. 2018). Given the potential deleterious consequences of disrupting genome organization and downstream effects on genome function (e.g. gene regulation), SV mutations are likely constrained by genome spatial organization. Comparative genomics studies revealed chromosomal arrangement breaks are enriched at the TAD boundaries but depleted inside TADs, possibly due to selection against chromosomal rearrangements that disrupt the TAD integrity (Krefting et al. 2018; Lazar et al. 2018). Other studies revealed deletions are strongly depleted at the TAD boundaries due to negative selection (Fudenberg and Pollard 2019) as the boundaries are important for insulating neighborhood TADs. However, it is still unclear whether nature selection effect is universal to all classes of structural variants.

In *Drosophila*, TADs have been extensively analyzed using Hi-C in embryos of early development stages (Dixon et al. 2012; Hug et al. 2017) and cell lines of different origins (Li et al. 2015; Cubeñas-Potts et al. 2017; Chathoth and Zabet 2019; Wang et al. 2018). These studies revealed prominent TAD structures across the genome with numbers varying from ~1300 to ~4000 based on the resolutions of Hi-C data can generate. Additionally, many intriguing and unique properties of *Drosophila* TAD boundaries were revealed, such as the fact that they coincide with conserved noncoding sequences (Harmston et al. 2017), housekeeping genes and certain transposons elements (Gong et al. 2018; Dixon et al. 2012) and motif sequences (Ramírez et al. 2018). However, most of these studies are conducted in *D. melanogaster*. Hi-C data available for other *Drosophila* species are more commonly used for genome scaffolding and are too low in coverage for spatial organization analysis (Bracewell et al. 2019; Mahajan et al. 2018). Thus, in-depth comparison between species is still lacking for *Drosophila*, while such studies are needed

to reveal the evolutionary role of spatial genome organization on *Drosophila* genome biology.

To better understand the evolutionary role of the spatial genome organization on genome biology in *Drosophila*, we resequenced the genome of *D. pseudoobscura* to reference-quality and created the first high resolution chromatin contact map (~800 bp) for this species using full body Hi-C. We assessed the conservation of spatial genome organization by comparing the TAD structures on the syntenic genomic regions between *D. melanogaster* and *D. pseudoobscura*, which diverged from each other approximately 49 million years ago (Thomas and Hahn 2017). Leveraging a genus-wide highly-contiguous genome assemblies of 17 *Drosophila* species (Miller et al. 2018) spanning ~72 million years of evolution, we characterized evolutionary rearrangements in the context of TAD structures. We also compared the gene expression pattern across 7 tissus in both sexes between *D. melanogaster* and *D. pseudoobscura* to see whether evolutionary TAD arrangements were associated with gene expression divergence. Finally, we used high-confident structural variants identified from reference-quality genome assemblies in both intraspecies (14 *D. melanogaster* strains) (Chakraborty et al. 2019, 2018) and interspecies (*D. melanogaster* versus three simulans clade species; *D. pseudoobscura* versus *D. miranda*)(Chakraborty et al. 2020; Mahajan et al. 2018) comparisons to test the impact of spatial genome organization on SVs evolution. The genome resource and Hi-C contact map generated in the current study add valuable genetic information to *Drosophila* genome biology research. Our analysis provides new insights into the evolutionary significance of spatial genome organization on *Drosophila* genome function, structure and evolution.

RESULTS

A nearly complete genome assembly of *D. pseudoobscura*

We resequenced females of *D. pseudoobscura* with deep coverage (~280X; assuming genome size G=164Mb) long reads and scaffolded it using high resolution Hi-C (~726X Hi-C read coverage) chromosome contact maps(Supplemental Table S1). The assembly is highly contiguous, with 90% of the 164Mb assembly being represented by 6 contigs that are 9.7 Mb or larger (i.e. N90 = 9.7 Mb and N50= 30.7 Mb; Supplemental Table S2). This new genome assembly of *D. pseudoobscura* is also highly accurate at the nucleotide level (concordance with Illumina reads yields a QV = 52) and harbors 99.6% of the 1066 complete universal Arthropoda single-copy orthologs (BUSCO)(Supplemental Table S3). The three telocentric autosomes (Chr2, Chr3, and Chr4), the dot chromosome and the mitochondrial DNA (mtDNA) are each assembled into single contigs. The X chromosome has two sequence gaps including one at the centromeric region. Additionally, there are 64 unplaced contigs totaling ~6.6 Mb, 98% of which is annotated as repeats. Among these, 37 contigs are likely centromeric sequences as they are composed of centromere specific repeats (Supplementary Table S4). High-throughput chromatin conformation capture (Hi-C) data is used to verify the global continuity of the assembly and identified a large (~9.7 Mb) pericentromeric inversion (Supplemental Fig. S1) on the X chromosome between our assembly and previous assemblies of this species (Bracewell et al. 2019; Richards et al. 2005). We annotated 13,413 gene models using supporting evidence from RNA-seq data and full length mRNA sequencing (Iso-seq) from females (22,237 isoforms) and males (15,372 isoforms) (Supplemental Table S5-6). Approximately 26.9% of the *D. pseudoobscura* genome is annotated as repetitive sequences, including 13.95% in retro-elements retrotransposons (Supplemental Table S5). The distribution of genomic features is generally conserved with its closely related species *D. miranda (Mahajan et al. 2018)*, despite at least 10 large inversions reshuffling their genome synteny (Supplemental

Fig. S2). The high quality genome assembly generated here is a valuable genetic source for genome topology and evolution study in the *Drosophila* genus.

TAD annotation using high resolution Hi-C data from the adult full body of *D. pseudoobscura*

The *Drosophila* TADs have been extensively studied using Hi-C data collected from cell lines (AlHaj Abed et al. 2019; Chathoth and Zabet 2019; Wang et al. 2018) and embryos (Pal et al. 2019; Hug et al. 2017; Ghavi-Helm et al. 2019; Sexton et al. 2012), chromosome contact maps in adults have largely been ignored. We used an optimized Arima-HiC protocol (Arima Genomics Inc, San Diego) to generate the Hi-C data for *D. pseudoobscura* from adult full body. This protocol uses multiple cutting restriction enzymes for chromatin digestion which can obtain a theoretical mean restriction fragment resolution about ~160 bp in *D. pseudoobscura* genome. A total of 397 million raw paired end reads (2x150 bp) were generated, of which half are retained after filtering and are finally valid for constructing the Hi-C contact map (Supplemental Table S7). The map resolution is about ~800bp calculated by the method described in Rao et al (Rao et al. 2014).

Since TAD annotation may vary moderately among computational methods (Forcato et al. 2017; Zufferey et al. 2018; Dali and Blanchette 2017), we used three different tools, HiCExplorer (Ramírez et al. 2018), Armatus (Filippova et al. 2014) and Arrowhead include in the Juicer package (Durand et al. 2016) to identify TADs. TAD calling is optimized with various combinations of parameters for each tool by comparing the identified TADs to the Hi-C contact heatmap (Supplemental Table S8; Supplemental Fig. S3). When the bin size of the contact map was arbitrarily chosen to be 5kb for all tools (Fig. 1A), Armatus annotated 858 TADs (>30Kb) with an average size of 123 kb; Arrowhead reported TADs in a nested format and discrete TADs are allowed (the neighbour TADs don't have to share the same border). It annotated a total of 795 TADs with a mean size of 148 kb; HiCExplorer predicted 996 continuous (i.e. the neighbour TADs share the same border) TADs (>30Kb) with an average size of 146 kb (Fig. 1B,C). Although these callers reported TADs that do not

completely overlap, there are 589 domains reported in at least two callers, covering 58% of the genome (Fig. 1B) .

To ask whether these bioinformatically inferred TADs are also supported by biological evidence, we investigated the chromatin landscape around TAD boundaries. In agreement with previous results in *D.* melanogaster (Wang et al. 2018; Chathoth and Zabet 2019), we found that TAD boundaries are enriched in the two insulator proteins, BEAF-32 (Yang et al. 2012) and CTCF (Ni et al. 2012), both with publicly available ChiP-seq data in *D. pseudoobscura* (Fig. 2D, Supplemental Fig. S4), although the data are not collected from the same tissue. We also found that TAD boundaries are enriched in open chromatin as measured by ATAC-seq data (Fig. 2E, Supplemental Fig. S4)(Jacobs et al. 2018). Furthermore, We found that TAD boundaries are highly enriched for the active chromatin mark, H3K4me3, but depleted for the repressive chromatin mark, H3K27me3 (Fig. 2F, Supplemental Fig. S4)(Schuettengruber et al. 2014). All these results are consistent with previous results in *D. melanogaster* (Hug et al. 2017). Collectly, these results demonstrate that Hi-C data from the full adult bodies can identify a considerable proportion of biology meaningful TADs in *Drosophila*, implying this spatial feature conserved across cells and tissues for a large number of TADs.

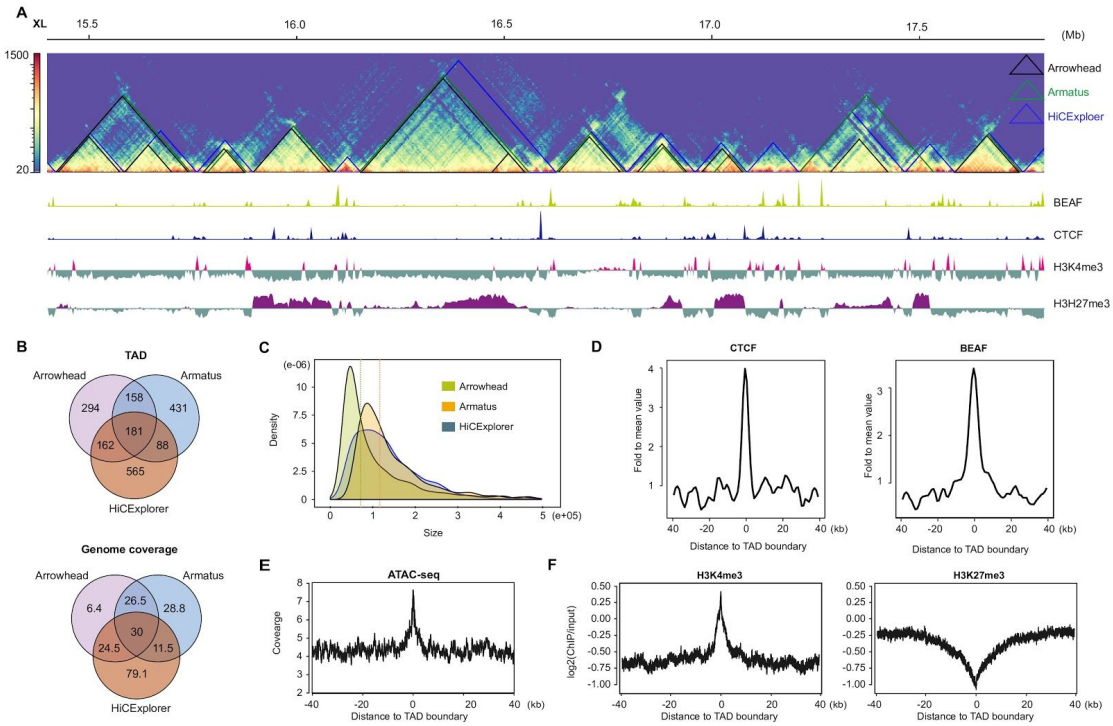


Figure 1: In situ Hi-C map and topologically associating domains (TADs) from the *D. pseudoobscura* adult full body. (**A**) Top: Hi-C contact map from 14.8 Mb to 17.8 Mb on X chromosome with 5kb bin size and TADs identified by three methods, Arrowhead, Armatus and HiCExplore ; Bottom: ChIP-seq for two insulator proteins, BEAF-32 and CTCF, and two epigenetic marks, H3K4me3 and H3H27me3. (**B**) TAD annotation consistency among three methods in terms of number (top) and genome coverage (below). (**C**) TAD size distribution for three methods. Vertical dashed lines are the mean value for each method. Armatus and HiCExplorer have the similar mean value. (**D**) Enrichment of CTCF and BEAF-32 at TAD boundaries. (**E**) Open chromatin measured by ATAC-seq signal around the TAD boundaries. (**F**) Enrichment (H3K4me3) and depletion (H3K27me3) of epigenetic marks at the TAD boundaries.

Remarkable evolutionary conservation of TAD structure between *D. pseudoobscura* and *D. melanogaster*

TADs are highly conserved across cell types, tissues and species (Dixon et al. 2012; Vietri Rudan et al. 2015; Fishman et al. 2019). To investigate conservation of this spatial

structure during *Drosophila* evolution, we compared TADs between *D. pseudoobscura* and *D. melanogaster*, which diverged about ~49 million years ago (Thomas and Hahn 2017). To do so, we first constructed the genome synteny map between these two species. The resulting map consists of 985 orthologous blocks larger than 10kb (Fig. 2A), with an average length of 101kb in *D.* melanogaster and 109kb in *D. pseudoobscura*, spanning 72% (100/140Mb) of the *D. melanogaster* genome and 69% (110/164 Mb) of the *D. pseudoobscura*, respectively (Supplemental Table S9). We found that most orthologous blocks are placed in the same Muller elements, suggesting that even on a small scale, translocations rarely occur between chromosomes during *Drosophila* evolution (Fig. 2A), consistent with a recent study (Renschler et al. 2019).

We identify TADs using published Hi-C data from three cell lines (Kc167, BG3 and S2) (Chathoth and Zabet 2019; Wang et al. 2018) for *D. melanogaster*. TAD calling is carried out as described above for our *D. pseudoobscura* data to make the TADs we identify as comparable as possible between these two species (Supplemental Table S10). In agreement with previous studies (Ulianov et al. 2016; Hou et al. 2012), we found that TAD structures are highly conserved across cell lines in *D. melanogaster* by both analyzing the entire TADs or their boundaries in all tools (Fig. 2B; see Supplemental Fig. S5 for another 7 syntenic regions). For example, among HiCExplorer-TADs, ~68% of the TADs and 76% of their boundaries are shared at least in two cell lines and only ~32% of the TADs and 24% of the boundaries occur specifically in one cell line (Fig. 2C). This observation is also shown for TADs annotated with two other tools (Supplemental Fig. S6).

Remarkably, we found that TAD structures are also highly conserved between *D. melanogaster* and *D. pseudoobscura* in any pairwise comparison of the entire TADs or boundaries from *D. melanogaster* cells of different origin and those from *D. pseudoobscura* full body (Fig. 2B; Supplemental Fig. S5). Among HiCExplorer-TADs, in the pairwise comparison of *D. melanogaster* S2 cell and *D. pseudoobscura*, 44% (339/776) of the S2 TAD boundaries (Fig. 2D) are conserved with *D. pseudoobscura* TADs (full body), which is 2.4 fold enrichment relative to random expectation (~18.5 %; Fisher's exact test p-value < 2.2e-16). Correspondingly, 46% (314/678) of *D. pseudoobscura* TAD boundaries (Fig. 2D)

are also detected in the S2 cell line, compared to random expectation of 21% (Fisher's exact test p-value < 2.2e-16). Similar trends are also shown in the analysis with Kc167 and BG3 cell lines and with TADs identified in different tools (Supplemental Fig. S6). Our results show that at least 40% of the TADs spanning 30% of the genome (Fig. 2E; Supplemental Fig. S7) in each species of *D. melanogaster* and *D. pseudoobscura* still share TAD structure in the other species, although they diverged over ~49 million years ago and with extensive genome reshuffling.

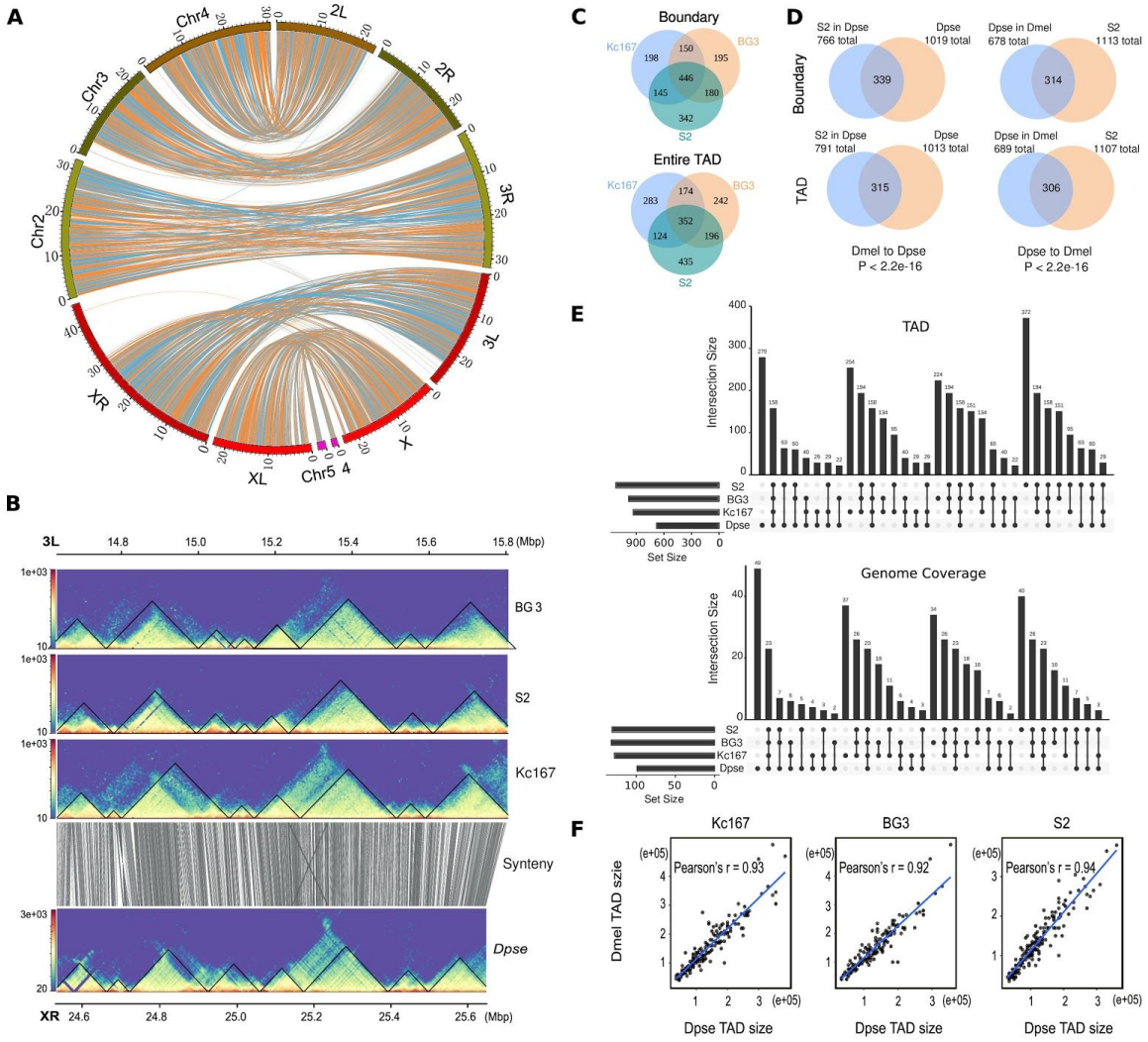


Figure 2: Evolutionary conservation of TAD structure between *D. melanogaster* and *D. pseudoobscura*. (**A**). Genome-wide synteny map between *D. melanogaster* and *D. pseudoobscura*

constructed using 985 syntenic blocks which are larger than 10kb. (**B**) Hi-C contact maps and TAD structures from a 1.2 Mb region between *D. melanogaster* and *D. pseudoobscura*. (**C**) Conservation of TAD structure among three cell types in *D. melanogaster* in terms of boundaries and number. (**D**) Conservation of TAD boundaries between *D. melanogaster* S2 cell line and the full body in *D. pseudoobscura*. (**E**). Upset plot showing the overlap of conserved TAD structures among three cell lines in *D. melanogaster* and *D. pseudoobscura*. Top: TAD number; Below: Genome coverage. (**F**) Correlation of TAD size between *D. melanogaster* and *D. pseudoobscura* conserved TADs.

Interestingly, we found that the sizes of syntenic orthologous TADs between *D. melanogaster* and *D. pseudoobscura* are correlated with their genome sizes or local genomic region sizes (Fig. 2F). This observation indicates that TADs can proportionately expand or contract with the genome or local genomic regions, suggesting that TADs may be the robust entities that can tolerate TE insertions, deletions, and other events that can cause genome size change.

Assessment of the evolutionary conservation of TAD boundaries implies relatively important factors in maintenance of TAD

We next investigate which properties of TAD boundaries are more likely to be conserved between species. Firstly, we considered TAD boundaries that overlapped with binding sites of different architectural proteins. For *D. melanogaster*, we obtained published ChIP-seq data for six architectural proteins (BEAF-32, CTCF, CP190, Chromator, Su(Hw), Trl) for three cell types (Kc167, BG3, and S2) (Supplemental Table S11). We found that boundaries that overlap with BEAF-32, CP190 and Chromator are substantially more conserved than those that do not, while boundaries overlap with CTCF, Su(Hw) and Trl don't (Fig. 3A; Supplemental Fig. S8). These results suggest that some architectural proteins may act as important factors in maintenance of TAD boundaries during evolution.

We next considered boundaries from TADs classified according to chromatin modifications. In a study conducted by (Ramírez et al. 2018), TADs are classified into four

groups: active (enriched for either H3K36me3, H3K4me3, and H4K16ac), polycomb group silenced (PcG; enriched for H3K27me3), HP1 (enriched for H3K9me3) and inactive (enriched for no mark). We found that TAD boundaries adjacent to at least one active TAD are generally more conserved (Fig. 3B), which were also classified as strong TAD boundaries (Ramírez et al. 2018). While, TAD boundaries between inactive-inactive TADs, inactive-PcG TADs or PcG-PcG TADs are less conserved.

We also classify TAD boundaries according to whether they are shared across cell lines or are specific to certain cell lines. Since our data for *D. pseudoobscura* comes from a single full body sample, we performed this classification in *D. melanogaster* based on the TAD boundaries from three cell types (Kc167, BG3 and S2). For HiCExplorer TADs, we identified 921 boundaries that are shared at least in two cell types, while 198 appeared in only Kc167, 195 in BG3 and 342 in S2 (Supplemental Table S12). We found that 58% (391/672) of boundaries shared by two or more cell types are also TAD boundaries in *D. pseudoobscura* (23% as expected), which is significantly more than those cell-specific boundaries (Fisher exact test, P-value < 0.001) which only shares only 34% (165/483)of the boundaries with *D. pseudoobscura* (Fig. 3C). Similar results were obtained when TAD boundaries were annotated with Arrowhead and Armatus (Fig. 3C). These results indicate that TAD boundaries across cell types are more likely to be conserved than cell type specific boundaries. But we can't rule out the possibility that this pattern is caused by the fact that cell type specific TAD boundaries are underrepresented in the *D. pseudoobscura* full body TAD set.

Finally, we classify TAD boundaries into strong or weak groups (Chathoth and Zabet 2019) based on what stringent thresholds are used in the HiCExplorer. For *D. pseudoobscura*, we found that strong boundaries are significantly conserved than weak boundaries in any comparisons between *D. pseudoobscura* TAD set and those from three cell types in *D. melanogaster* (Fig. 3D). For example, 64% (235/367) of strong TADs boundaries are conserved between Kc167 and *D. pseudoobscura* (29% as expected), while only 40% (130/326) of weak boundaries are found to be conserved (Fig. 3D). Correspondingly, this pattern is also observed in the reciprocal comparison focusing on *D.*

melanogaster TAD boundaries (Fig. 3D). This result suggests that stronger TAD boundaries should be more conserved in the course of evolution.

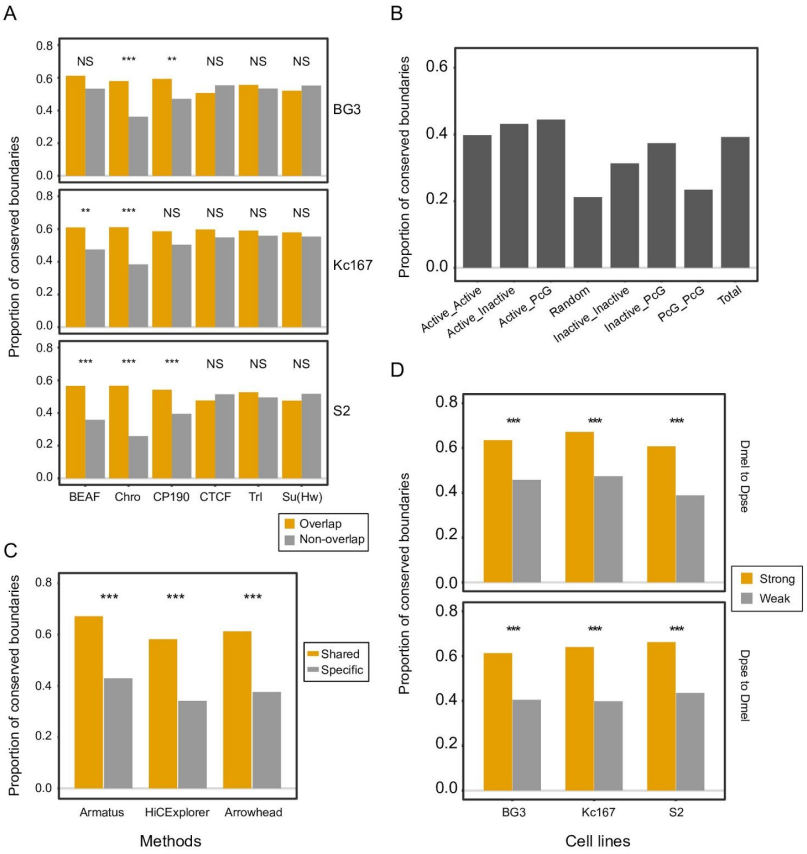


Figure 3: Conservation of TAD boundaries between *D. melanogaster* and *D. pseudoobscura*. The exact number of each category of TAD boundaries is provided in Supplemental Table S12. (A) Conservation of TAD boundaries overlapped with architectural proteins. (B) Conservations of boundaries from TAD with different chromatin modifications. (C) Comparison between boundaries that share at least in two cell types and those only specific to one cell type. (D) Comparison between strong and weak boundaries. Significance is determined by Chi-square test (***P < 0.001; **<0.01; NS: no significance).11

Topologically associating domains and gene regulation: insights from long genes coincide with entire TADs

TADs have been reported to be associated with evolutionary stability of gene expression, suggesting that disruption of TADs may result in perturbation of gene expression (Krefting et al. 2018). To test this in *Drosophila*, we analyzed the conservation of gene expression of a total of 10,921 one-to-one orthologs between *D. melanogaste*r and *D. pseudoobscura* across 7 tissues from both males and females (Yang et al. 2018). Krefting et al. (Krefting et al. 2018) suggested that orthologs exhibiting strong correlation of gene expression between species across the matching tissues are likely to have similar functions modulated by similar regulatory programs(Ludwig et al. 2000). We first classified the orthologs into two sets: 1) genes inside TADs (i.e. genes fall in the genomic regions where TAD bodies are annotated using Juicer) and 2) genes outside of TADs. We found that the genes inside TADs have a higher gene expression correlation than genes outside TADs (mean $r = 0.785$ versus $r = 0.678$, $p < 1.4e\text{-}12$, Fig. 4A). Next, we focused on genes that fall inside TAD bodies and classified them into: 1) genes fall in the conserved TADs between *D. melanogaster* and *D. pseudoobscur*a and 2) genes fall in the non-conserved TADs. The expression correlation is higher for genes in conserved TADs than genes in non-conserved TADs (mean $r = 0.821$ versus $r = 0.765$, $p = 7.45e\text{-}5$, Fig. 4B). These results suggest that TAD arrangements are associated with variance in conservation of gene expression during *Drosophila* genome evolution.

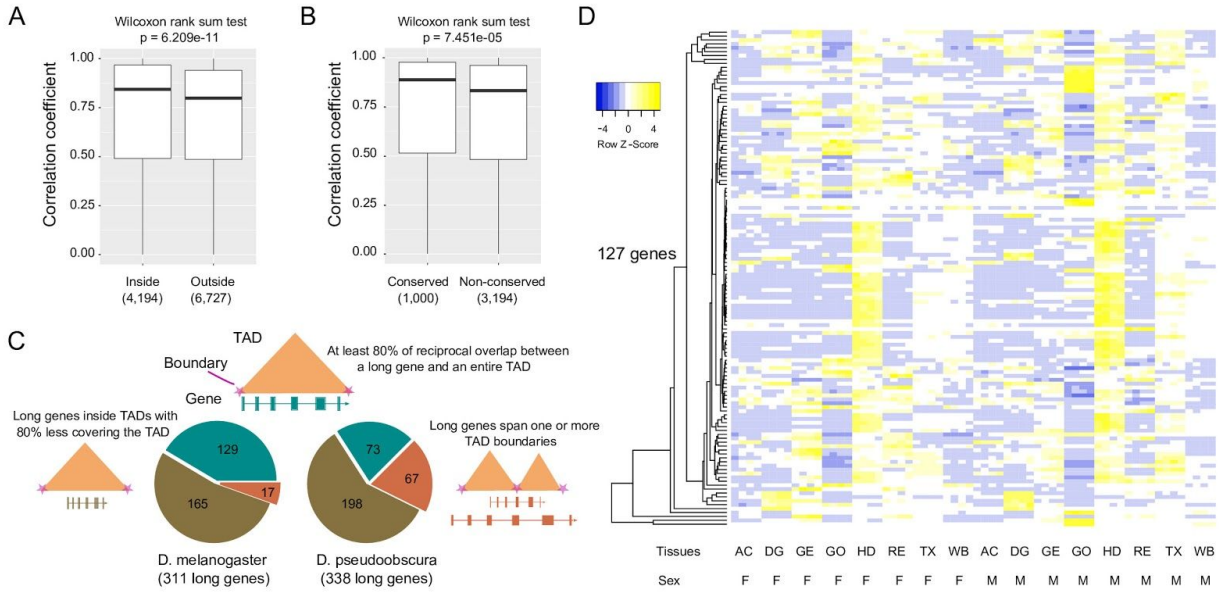


Figure 4: **The role of TADs in gene expression evolution and regulation of long genes.** (A) Expression correlation of one-to-one orthologs across 7 tissues and both sexes in *D. melanogaster* and *D. pseudoobscura* for *D. melanogaster* genes within or outside the TADs. (B) Expression correlation of one-to-one orthologs across 7 tissues and both sexes in *D. melanogaster* and *D. pseudoobscura* for genes in conserved or non-conserved TADs. (C) Distribution of physical overlap between long genes and annotated TADs spanning them in *D. melanogaster* and *D. pseudoobscura*. (D) Expression profile of 127 long genes that are fully occupied by full TADs across 8 tissues in both female and male of *D. melanogaster*. AC, abdomen without digestive or reproductive system; DG, digestive plus excretory system; GE, genitalia; GO, gonad; HD, head; RE, reproductive system without gonad; TX, thorax without digestive system; WB, whole body. F, female; M, male.

To characterize the possible relationship between gene regulation and TAD structure, we selected the 311 longest coding genes (those > 50kb), totaling 34 Mb of the genome in *D. melanogaster*, and examined their overlap with annotated TADs. Among these long genes, we found that TAD boundaries are significantly depleted inside gene bodies compared to genome background (342 versus 765; Fisher's exact test, P< 10e-23), consistent with the observation that TAD boundaries tend to be enriched at gene

promoters (Ramírez et al. 2018). For only 17 long genes were TAD boundaries found to be present inside their spans for all three cell types (Fig. 4C). More interestingly, we identified 129 long genes (Supplementary Table S13) in *D. melanogaster*, comprising 15 Mb in total, that each individually occupied a full TAD predicted in at least one of the three cell lines or tools (Supplementary Fig. S9). This significant association (Permutation test, $P < 0.0001$) might reflect that TADs play a functional role in long distance gene regulation. Similarly, we observed the analogous trend by analyzing the 338 longest genes (those > 50Kb) in the genome of *D. pseudoobscura*. However, we found relatively fewer long genes (73) that span full TADs in *D. pseudoobscura* data (Supplemental Table S14), which may perhaps be due to the fact that TADs are annotated only in the whole body, whereas TADs in *D. melanogaster* come from separate experiments on three separate cell types. This observation suggests a possibility that some TAD structures emerged specifically in certain cell types or developmental stages to regulate their matched genes. This prediction can be further bolstered by the fact that most of the long genes are involved with development processes and have a relatively narrow expression profile (Fig. 4D, Supplemental Fig. S10). Taken together, our results indicate that gene structure might be an important factor determining the TAD formation and maintenance or perhaps that TADs are important in the regulation of these genes.

Breaks in synteny enriched at TAD boundaries

Conservation of TAD structures between the two distantly related *Drosophila* species further prompted us to ask if disruption of TAD integrity (i.e. TAD shuffling or fusing as a result of large-scale chromosomal arrangements) in *Drosophila* is also constrained during evolution as previously shown in other animals such as mammals (Krefting et al. 2018; Lazar et al. 2018) and birds (Fishman et al. 2019). We test this by analyzing the distribution of chromosomal rearrangement breakpoints (using synteny breaks as the proxy) along the TAD body. In the comparison between *D. melanogaster* and *D.*

pseudoobscura, we found that synteny breaks largely coincide with TAD boundaries (Fig. 5A; Supplemental Fig. S11). About 33% (280/859) the synteny breaks are found to be overlapped with TAD boundaries corresponding to 2.5 fold enrichment relative to random expectation (Fisher exact test, *P-value* < 0.001), suggesting chromosomal arrangement breakpoints are not randomly distributed across the genome and might be influenced by genome spatial organization. This pattern was also observed in a recent work (Renschler et al. 2019) in which the authors compared *D. melanogaster* to two distantly related species.

Further, we extended the analysis to a total of 17 *Drosophila* species, spanning a period of 72 million years evolution (Fig. 5B; (Thomas and Hahn 2017)). All these species have highly contiguous genome assemblies with N50 at least 4Mb (Miller et al. 2018; Mahajan et al. 2018), enabling accurate and reliable identification of chromosomal arrangement breaks. Using the whole genome alignment pipeline(see Methods), we identified a total of 108 to 1180 synteny breaks and 10 to 314 inversion breaks in the comparisons between the 16 query species and *D. melanogaster* (Supplemental Table S15), respectively. The number of synteny breaks detected in the 16 species is positively correlated with their divergence times with *D. melanogaster* (Supplemental Table S15), with some discordant cases may result from the quality of genome assembly. Using this data, we found that the evolutionary synteny and inversion breaks were strongly enriched with the TAD boundaries, while inside TADs the frequency of breaks was slightly depleted (Fig. 5C; Supplemental Fig. S12), except in the comparisons between *D. melanogaster* and the three species (*D. sechellia*, *D. mauritiana*, and *D. simulans*) in the *D. simulans* clade. This exception is caused by lacking enough informative data for interpretation.

We also repeated the analysis using *D. pseudoobscura* as reference. Correspondingly, we identified a total of 259 to 1,242 synteny breaks and 60 to 359 inversion breaks (Supplemental Table S15) in comparisons between 16 query species and *D. pseudoobscura*. With these dataset, we observed the analogous trend as above (Fig. 5D; Supplemental Fig 12).

The enrichment of chromosomal arrangement breaks at the TAD boundaries suggest that breaks occur at the TAD boundaries ensuring the stability of the TADs during genome

shuffling. As shown in Fig. 5E, three TADs are still well preserved on an ~ 450kb inverted genomic region between *D. melanogaster* and *D. paseudoobscura*. Thus, our results from a genus-wide genome dataset demonstrate that at least a considerable fraction of TAD structure, if not most, in the *Drosophila* genome is preserved and chromosomal arrangements resulting in their disruption is constrained.

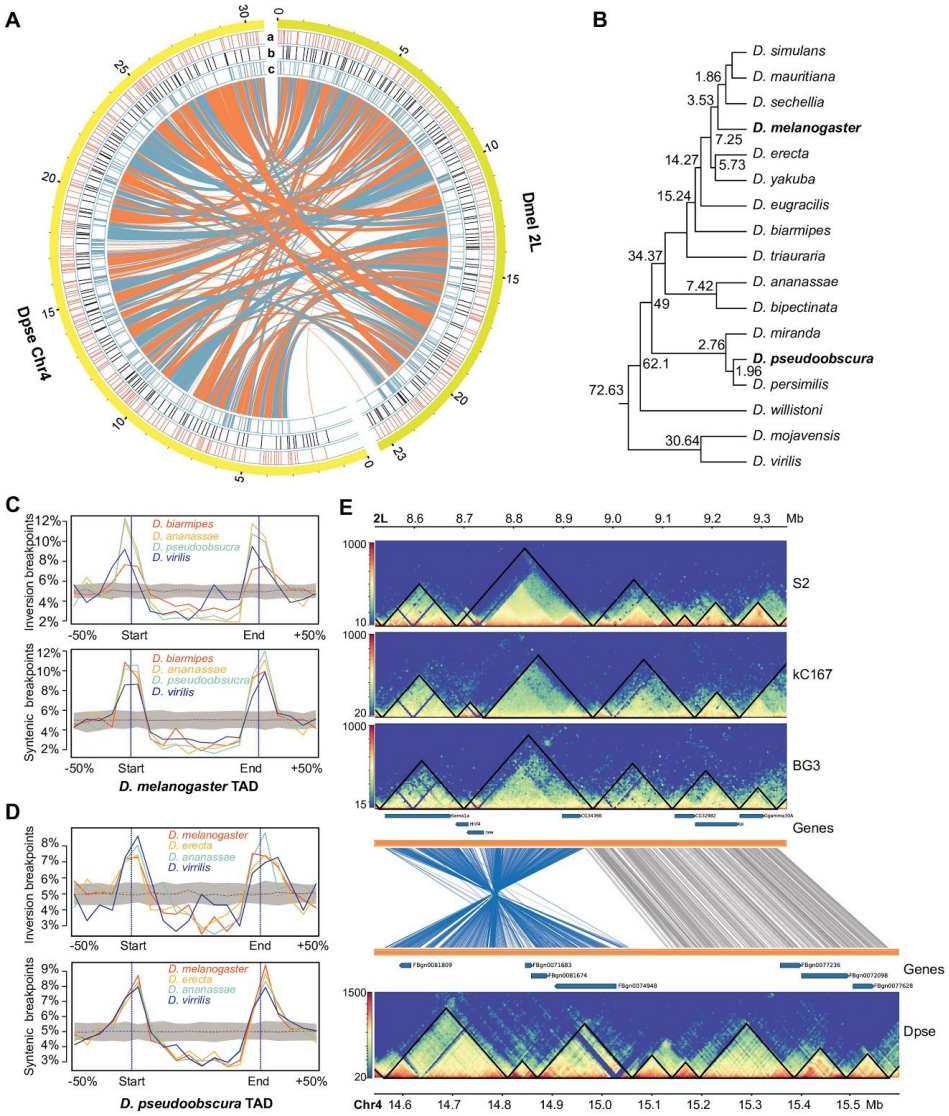


Figure 5: Evolutionary synteny breaks are enriched with TAD boundaries in D*rosophila* genomes. (A) Synteny map between *D. melanogaster* 2L and *D. pseudoobscura* Chr4. Other pairs of homologous chromosomes are shown in Supplementary Fig. S11. Tracks a: TAD boundaries annotated by HiCExperor at restriction fragment resolution; b: 10 kb resolution; c: synteny

breakpoints. (B) Phylogenetic trees of the 17 *Drosophila* species. Estimated divergence times are obtained from (Thomas and Hahn 2017). (C) Distribution of evolutionary synteny/inversion breakpoints between *D. melanogaster* and 4 query species around *D. melanogaster* TADs. There are another 12 species shown in Supplemental Fig. S12. (D) Distribution of evolutionary synteny/inversion breakpoints between *D. pseudoobscura* and 4 query species around *D. pseudoobscura* TADs. There are another 12 species shown in Supplemental Fig. S12. (E) Conservation of TAD structure in an inverted genomic region (*D.mel* 2L: 8.55 - 8.95 Mb) between *D. melanogaster* and *D. pseudoobscura*.

Natural selection constrains structural variants at *Drosophila* TAD boundaries

Beyond large-scale chromosomal arrangements, small or medium structural variations, such as deletions, insertions, and duplications can also affect TAD stability by disrupting their boundaries. Variants that disrupt TAD boundaries have been shown to cause alteration on chromatin topology and gene expression (Arzate-Mejía et al. 2020; Ghavi-Helm et al. 2019; Despang et al. 2019; Sadowski et al. 2019). Thus, structural variants should also be constrained by chromatin topology, an observation recently reported in humans. For example, deletions within human populations (Sadowski et al. 2019), and between their close relatives (Fudenberg and Pollard 2019) and cancer genomes (Akdemir et al. 2020) have all shown to be depleted at the TAD boundaries. To explore the distribution of SVs in the context of spatial genome organization in *Drosophila* and whether they are constrained at TAD boundaries by natural selection, we obtained two high-confidence SVs datasets based on reference-quality genome assemblies (Fig. 6A): 1) polymorphic SVs that detected within 14 *D. melanogaster* strains (Chakraborty et al. 2019, 2018) and 2) between species and/or divergent (~3 million years) SVs that detected in the three species (*D. simulans, D. sechellia,* and *D. mauritiana*) of the *Drosophila* simulans clade (Chakraborty et al. 2020). Both are relative to the reference genome assembly of *D. melanogaster* ISO1 strain.

The SV types include polarized deletions and non-TE insertions when using *D. erecta* and *D. yakuba* as outgroup (Fig. 6A), TE insertions and duplications for 14 *D. melanogaster*

strains (Fig. 6B) and the three species in simulans clade (Fig. 6C). For polymorphic SVs in *D. melanogaster*, most are strain-specific, with TE insertions being most common and deletions the least (Fig. 6D). For interspecific SVs, although most are species-specific, a large proportion is found to be present in at least two simulans complex species, suggesting that they are more likely to be conserved than intraspecific mutations (Fig. 6E).

To detect signatures of selection on SVs at *Drosophila* TAD boundaries, we exploited the method previously described in (Fudenberg and Pollard 2019)). It simply compares the observed SV event counts and affected sequence coverage in the TAD boundary regions with a uniform genome-wide expectation, which imposes the simplifying assumption that the SV mutation rate is fairly similar across the genome (Methods). We used 2,185 TAD boundaries that were shared in two independent works by Ramírez et al (2018) and Wang et al (2018), respectively, in the euchromatin region (Supplemental Table S16). A total of 8.74 Mb genomic regions (4kb for each of 2,185) was annotated as the genomic context of TAD boundary. We then assessed the relative abundance (in terms of breakpoints and coverage) of deletions, Non-TE insertions, TE insertions and duplications in the 8.74 genomic regions compared to expectation. We separate deletions and insertions into large (>10bp) and small (<11bp) groups to see whether selection strength is different for SV size. This analysis shows that both polymorphic deletions in *D. melanogaster* and deletions from *D. simulans* clade species are strongly depleted at the TAD boundary regions, which represents a signature of purifying selection (Fig. 6F, G; Supplemental Fig. S13). Insertions show a relatively complex pattern. Although the overall insertions are found to be depleted at the TAD boundaries (Fig. 6H), some special types of insertions are not. For example, among TE insertions, LTR elements are the most strongly depleted at TAD boundaries, followed by LINE, but DNA-type TEs and simple repeats seem not to be depleted at the TAD boundaries (Supplemental Fig S14-15). It is worth noting that deletions and insertions from the *D. melanogaster* population are more depleted than those from *D. simulans* clade species (Fig. 6H-G).

An interesting contrast is observed for duplications. Duplications from the *D. simulans* species complex are found to be enriched at the TAD boundaries but duplications in the

sample of *D. melanogaster* strains are not, suggesting that this excess divergence may indicate the action of positive selection on duplications at the *Drosophila* TAD boundaries. To further test this, we identified duplications in the D. *miranda* genome in relation to *D. pseudoobscura*. We found duplications in the *D. miranda* genome are also enriched in the TAD boundaries identified from *D. pseudoobscura* (Fig. 6F), suggesting TAD boundaries may act as relatively frequent targets of duplications as also shown in human TAD boundaries(Sadowski et al. 2019).

Finally, we investigated the breakpoints of both deletions and insertions around the TAD boundaries. The results show that deletions and insertions are both more depleted at the TAD boundaries than the surrounding regions (Fig. 6G,H). Larger deletions and insertions in the *D. melanogaster* population dataset are more depleted than shorter ones (Fig. 6G). But this trend was not observed in the *D. simulans* clade dataset (Fig. 6H). Taken together, our results revealed empirical evidence of selection on structural variants at *Drosophila* TAD boundaries at a wide variety of timescales.

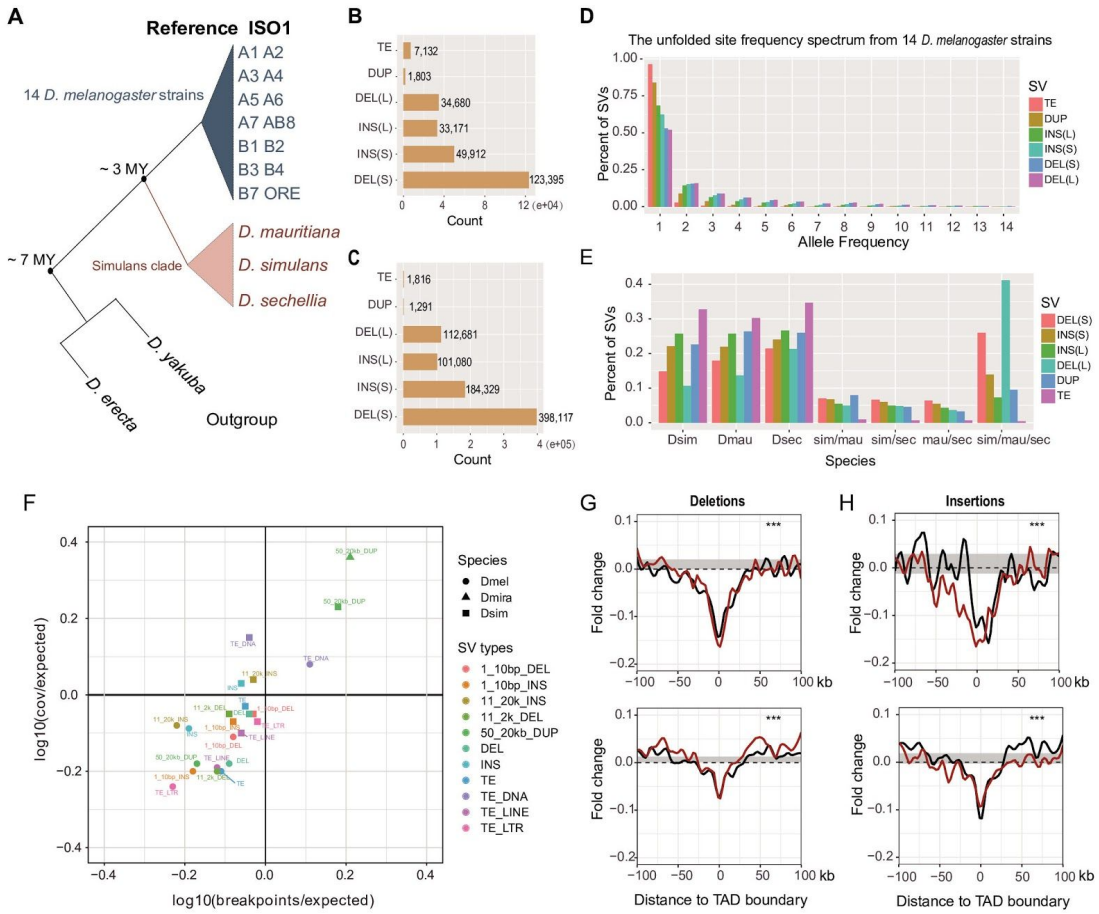


Figure 6: Structural variants are under nature selection at *Drosophila* TAD boundaries. (**A**) Highly continuous genome assemblies from 14 *D. melanogaster* strains and three *D. simulans* clade species (*D. mauritiana, D simulans* and *D. sechellia*), together with two outgroup species, *D. erecta* and *D. yakuba*, were used to identify and polarize structural variations. **(B)** Non redundant structural variations, including TE insertions, duplications (DUP), insertions (INS(S): 1-10bp; INS(L): 11bp ~ 20Kb), and deletions (DEL(S):1-10bp, DEL(L): 11~2Kb) identified in the 14 *D. melanogaster* strains. **(C)** Non redundant SVs identified in three *D. simulans* clade species. **(D)** The unfold site frequency spectrum of structural variants from 14 *D. melanogaster* strains. **(E)** Evolutionary placement of structural variants among three *D. simulans* species. **(F)** Test of natural selection signature for structural variants at the TAD boundaries**. (G)** Deletions from both 14 *D. melanogaster* and three *D. simulans* clade species are depleted at the TAD boundaries. Red line represents deletions larger than 10bp and black line represents deletions smaller than 11bp. (**H**) Insertions from both 14 *D. melanogaster* and three *D. simulans* clade species are depleted at the TAD

boundaries. Red line represents insertions larger than 10bp and black line represents insertions smaller than 11bp.

DISCUSSION

Our knowledge of the evolution genome topology and structure remains in its infancy (Yang et al. 2019). Chromosome conformation capture techniques such as Hi-C have enabled the high-resolution characterization of chromosomal organization. Interspecies comparison of Hi-C contact maps and the topological domains or TADs reveal robust conservation of 3D genome organization between species in mammals (Vietri Rudan et al. 2015), plants (Xie et al. 2019) and recently in *Drosophila* (Renschler et al. 2019). The reference-quality genome assembly and high-resolution Hi-C contact map (~800bp) of *D. pseudoobscura* generated here enable us to decipher the evolutionary significance of spatial genome organization on genome function and structure such as the impacts on chromosomal arrangements, structure variants and gene expression in *Drosophila*.

Conservation and evolution of spatial genome organization

We generated Hi-C data from the adult full body for *D. pseudoobscura* and detected pronounced TAD structures on the Hi-C contact map based on the high depth of Hi-C data. The identified TADs vary slightly for different methods, but all show biologically meaningful signals insofar as they tend to coincide with epigenetic domains (e.g. H3K4me3 and H3K27me3) and their boundaries are significantly enriched for insulator proteins (e.g. CTCF and BEAF-32). This suggests that our TAD set is valid for investigating genome topology and evolution in the *Drosophila* genus.

TAD features have been found to be conserved across species. Evolutionary conservation of TAD structures are associated with stabilization of gene expression and regulation of gene expression. About ~43% of the TADs were found to be shared between humans and chimpanzees (Eres et al. 2019). About 10% of the TADs are shared among all

three distantly related *Drosophila* species-*D. melanogaster*, *D. busckii* and *D. virilis* -even though with the low degree synteny of their genomes (Renschler et al. 2019). Our comparative Hi-C analysis of *D. melanogaster* and *D. pseudoobscura* revealed that at least 40% of the TADs, spanning nearly half of syntenic genomic regions, are still shared. The relatively higher conservation degree observed here than Renschler's study may be due to the differences in methods of estimation, depth of Hi-C data used in each study, and the evolutionary distance between the species pairs. Despite the differences, both results show that a substantial proportion of TADs are maintained during *Drosophila* evolution. It is worth noting that the estimated degree of conservation is likely underestimated in our study. Our estimates are derived from pairwise comparisons between Hi-C data from three cell lines (Kc167, S2 and BG3) in *D. melanogaster* and Hi-C data from adult full body in *D. pseudoobscura*. Given the extensive cell and allele-specific variability of TAD structures observed in Hi-C analysis of individual cell types (Nagano et al. 2013) and super-resolution fluorescence in situ hybridization (FISH) imaging approaches (Bintu et al. 2018), TADs detected in the full body in *D. pseudoobscura* will represent represent topolog averaged multiple tissues and millions of cells. Thus many cell-, tissue-, or developmental stage-specific TADs can be underrepresented in such a population-averaged data-set set. This is also probably one of the reasons why TADs shared across cell lines are more conserved than cell lines specific TADs (Fig. 3C). Future experiments that match samples (i.e. cells and tissues derived from the same condition such as developmental stages) between species are needed to further characterize this aspect of genome topology.

Role of gene transcription in the formation and maintenance of TAD structures

In addition to being modulated by genome topology, it has also been suggested that gene regulation and transcription in turn affects genome topology (van Steensel and Furlong 2019). Our results together with those of others (Krefting et al. 2018) show that evolutionary conservation of TAD structures correlates with stability of gene expression across tissues, while TAD rearrangements are generally associated with more variation of gene expression across tissues. However, two recent studies report that disruption of TAD

structure either by chromosomal rearrangements (Ghavi-Helm et al. 2019) or removal of the crucial TAD boundary insulator protein sites (Despang et al. 2019) does not significantly alter gene expression. This discrepancy suggests a possibility that TAD structure has a stronger effect on balancing gene expression across tissues or developmental stages than only effect in certain cells, tissues or developmental stages. An alternative explanation is that functional perturbations and their concomitant effects on fitness are subtle and only lead to observable signatures on evolutionary timescales(Crow et al. 1970) .

One intriguing finding in our study is that a substantial proportion of genes with long gene bodies (>50 kb) were found to each coincidentally overlap with one complete TAD. This feature is similar to the self-loop structure of genes (i.e. contacts between 5′ and 3′ ends of genes, forming local chromatin loops) found in *Arabidopsis thaliana* (Liu et al. 2016) and gene crumples in *S. cerevisiae* (Hsieh et al. 2015). Thus, our findings provide direct evidence that gene-based chromatin topology also exists in the animal system. Furthermore, these kinds of gene-based domains are more likely to be cell-, tissue-, or developmental stage-specific since we detected 129 (46%) genes in three cell lines in *D. melanogaster* and only 73 (26%) genes in the whole body in *D. pseudoobscura.* Additionally, our findings also show that TAD size may be influenced by gene body span as TAD size correlates with the expansion or contraction of local genomic regions. Thus, gene structure or transcription might serve as a major factor in the formation and maintenance of TAD in cell differentiation.

TADs as physical units during the course of genome structure evolution

Our analysis of chromosomal rearrangement breakpoints across 17 diverse *Drosophila* species strongly suggests that TADs are persistent structural features that may, among other things, govern where in the genome landscape breaks are likely to occur during the course of the evolution of genome structure. Our findings that chromosomal breaks are enriched at the TAD boundaries but depleted inside TAD bodies are consistent with the previous studies in mammals (Krefting et al. 2018; Lazar et al. 2018) and birds (Fishman et

al. 2019), suggesting that mutations that result in disruption of TAD integrity are negative selected. The non-random distribution of chromosomal breaks has two non-mutually exclusive explanations (Berthelot et al. 2015): 1) the breaks of chromosomal preferentially occur at "fragile regions" (e.g. regions with high GC content, high gene density, replication origins, repeat sequences and DNA hypomethylation); and/or 2) nature selection mediates the locations of breaks by favoring some types of locations and disfavoring others. We observed that the degree of enrichment of chromosomal breaks at the TAD boundaries increases as species diverge, suggesting that natural selection plays a role. Further experiments designed to characterize recurrent breaks across the genome or in the context of spatial genome organization are needed to further explain why chromosomal breaks are enriched at the TAD boundaries. Additionally, TAD structures are also found to affect other types of large scale chromosomal arrangements such as translocations (Zhang et al. 2012; Tao et al. 2017). Thus, TADs may play roles not only as functional units but also structural units in the evolution of genome structure.

Structural variation and spatial genome organization

Understanding the factors that affect the distribution of variation in the genome is a major goal of comparative genomics and population genetics. Structural variants can alter chromatin architecture (Shanta et al. 2020), while chromatin architecture can conversely affect the distribution of SVs through imposing selectional constraints on certain SVs. Our results and the results of others (Fudenberg and Pollard 2019) reveal that deletions, even over broad timescales, are found to be strongly depleted at the TAD boundaries, suggestive of purifying selection. Thus, it is likely that deletions are purged by selection in the TAD boundaries due to their potentially detrimental influence on chromatin architecture and/or their perturbation of functional sequences found at the boundaries (perhaps either coding or noncoding regulatory elements). TAD boundaries are found to coincide with gene-dense genomic regions and are enriched for noncoding regulatory sequences (Harmston et al. 2017). For example, ~77% of the TAD boundaries in *D. melanogaster* coincide with

promoters (Ramírez et al. 2018). Thus, it remains unclear which functional aspects are major factors governing constraint of deletions at the TAD boundaries.

Insertions have a relatively complex pattern in our analysis. Generally, the total insertions are found to be depleted at the TAD boundaries. However, if we separate insertions into separate classes such as TE insertions and Non-insertions, or even different types of TEs, we observe different patterns. Non-TE insertions are generally depleted at the TAD boundaries. LTRs are the most depleted insertion type at the TAD boundaries, then followed by LINE, most likely because they are able to directly alter the chromatin modification of the flanking regions where they inserted. While DNA TEs show a weak pattern of depletion at the TAD boundaries.

Intriguingly, we found that TAD boundaries are enriched for duplications from divergent species, suggesting that duplications can be an important evolutionary mechanism of spatial genome organization (Sadowski et al. 2019). Duplications may alter the copy number of regulatory elements or they may modify the 3D genome topology by disrupting the topological chromatin organization (Spielmann et al. 2018). Additionally, strong TAD boundaries are found to frequently co-duplicated with super-enhancers (Gong et al. 2018). The lack of enrichment of duplications at the TAD boundaries in the polymorphism data suggests that this pattern is unlikely to be mutationally driven, suggesting that duplication may be experiencing positive selection. Together, our results offer novel insight into the evolutionary significance of spatial genome organization on genome function and structure in *Drosophila* species.

MATERIALS AND METHODS

Genome sequencing, assembly and annotation

DNA was extracted from *D. pseudoobscura* adult females following a previously published protocol (Chakraborty et al. 2016). DNA was sheared using 21 gauge needles and size selected using the 30-80 Kb cutoff on Blue Pippin (Sage Science). Size selected DNA was sequenced on 10 SMRT cells using the Pacific Biosciences Sequel Platform. Illumina paired end (150bp * 2) reads were generated on Hiseq 4000 using the same DNA that was used for PacBio sequencing. All sequencing was performed at UC Irvine GHTF.

We obtained ~280 X (G =164Mb) PacBio long reads and ~283 X 150 bp PE short reads. PacBio long reads were assembled with Canu v1.7(Koren et al. 2017). After removal of redundant contigs and gap filling using raw reads with finisherSC (Lam et al. 2015), the assembly was polished twice with Arrow (Smrtanalysis 5.1.0) and three times with Pilon (Walker et al. 2014).

Tandem repeats were identified by Tandem repeat finder (Benson et al. 1999) with parameters "2 7 7 80 10 50 2000 -f -d -m". Transposable elements were annotated with the EDTA pipeline (Ou et al. 2019). Gene annotation was performed by iteratively running Maker (version 2.31.8) (Campbell et al. 2014) three times with guidance of RNA-seq and ISO-seq data. More information for genome annotation is provided in Supplemental Methods.

Single molecule RNA sequencing (Iso-seq) experiment and data analysis

Total RNA was extracted from the adult full body separately for male and female individuals using RNeasy Plus Mini Kit (Cat No./ID: 74134). cDNA synthesis and library preparation was performed at UCI GHTF. One SMRT cell for each sex sample was then sequenced using PacBio Sequel I. ISO-seq data were processed using the *IsoSeq* v3 pipeline which is available at https://github.com/PacificBiosciences/IsoSeq.

Hi-C experiment

Hi-C library was prepared for *D. pseudoobscura* adult whole bodies with both male and female individuals. Hi-C experiments were performed by Arima Genomics (https://arimagenomics.com/) according to the Arima-HiC protocol described in the Arima-HiC kit (P/N : A510008) with minor modifications to the crosslinking protocol. First, flies were crosslinked as whole animals using 2% formaldehyde. After crosslinking, flies were pulverized on dry ice with mortar and pestle and then subject to the Arima-HiC protocol described in the Arima-HiC kit. Briefly, pulverized crosslinked fly tissue was digested using a cocktail or restriction enzymes recognizing the GATC and GANTC motifs. Next, digested ends were labelled, proximally ligated, and then proximally-ligated DNA was purified. After the Arima-HiC protocol, Illumina-compatible sequencing libraries were prepared by first shearing purified Arima-HiC proximally-ligated DNA and then size-selecting ~400bp DNA fragments using SPRI beads. The size-selected fragments containing ligation junctions were enriched using Enrichment Beads provided in the Arima-HiC kit, and converted into Illumina-compatible sequencing libraries using the KAPA Hyper Prep kit (P/N: KK8504) reagents. After unique dual index adapter ligation, DNA was PCR amplified and purified using SPRI beads. The purified DNA underwent standard QC (qPCR and Bioanalyzer) and sequenced on the HiSeq X following manufacturer's protocols. A total of ~397 millions clean paired end (2*150bp) reads were generated.

Hi-C data processing and TADs annotation

Hi-C raw reads were processed using Juicer (Durand et al. 2016) and HiCExploer (Ramírez et al, 2018) for filtering, mapping and constructing contact matrices. TAD annotation was performed using three callers including Arrowhead contained in the Juicer package, Armatus (Filippova et al. 2014) and HiCExplorer. Contact matrices from Juicer were used for the former two callers to predict TADs. TAD boundaries are defined as genomic regions 5kb upstream and downstream from the start or the end of each TAD domain for these two callers. HiCExplorer outputs both TAD domains and boundary locations *.bed* file. HiCPlotter

(Akdemir and Chin 2015)was used for visualization and manual inspection. The optimized parameter combinations for each tool are provided in Supplemental Table X. The detail commands can be found at https://github.com/yiliao1022/TAD_SV.

ChIP-seq data and analysis

The ChIP-seq reads of BEAF-32 and CTCF for *D. pseudoobscura* obtained from previous studies (Yang et al. 2012; Ni et al. 2012) were aligned to our *D. pseudoobscura* assembly using bowtie2 v2.2.7 (Langmead and Salzberg 2012), and peak calling was performed using MACS2 v2.0.10 (Zhang et al. 2008). To quantify the distribution of these two insulators at TADs, we calculated the average occupancy value within 40kb downstream and upstream of each boundary using a 1 kb window. A matrix in which rows represent TAD boundaries and columns represent bins along the TADs was generated. The sum of each column indicates the number of peak summits for corresponding bins and therefore the same relative location around TADs.

We also obtained the ATAC-seq data and ChIP-seq data of H3K4me3 and H3K27me3 for *D. pseudoobscura* from previous studies. The raw reads were again aligned to the current genome assembly using bowtie2 v2.2.7. We calculated the log2 ratio of ChIP versus input as the target signal, which was binned and normalized using deepTools2 (v. 3.2.1)(Ramírez et al. 2016) *bamCompare* using the default parameters. Quantification of the enrichment and creating a profile plot of the signal at the TADs were performed using deepTools2 *computeMatrix* and *plotProfile*.

For *D. melanogaster*, we used either the ChIP-chip or the ChIP-seq data sets for six insulators (BEAF-32, CTCF, CP190, Chromator, Su(Hw) and Trl) for three cell lines (Kc167, BG3, and S2) that were generated from and preprocessed by the modENCODE Consortium (http://www.modencode.org/). A full list of these ChIP-chip and ChIP-seq data is provided in Supplemental Table S11.

Identification of conserved TADs and boundaries

Within species, conservation of TAD boundaries between different cell lines are identified using bedtools(Quinlan and Hall 2010) *intersect* function. In the case of TADs, conserved TADs are defined with bedtools with the parameters: *intersect* -F 0.8 -f 0.8, which require at least 80% reciprocal overlap with each other. To obtain interspecies conservation of TADs and boundaries, the coordinates of TAD domains and boundaries were first lifted over between species using UCSC liftover tool with the custom chain files. To estimate the background noise, we performed a randomization test for each pairwise comparison. To do so, we simulated 10,000 sets of random TAD or boundary locations which match the same distribution of TAD domain or boundary lengths and their chromosomal occurrence as the actual TAD domain and boundaries. The mean overlap percent of the 10,000 simulated locations was used as the background.

Expression data for *D. melanogaster* and *D. pseudoobscura*

The preprocessed expression data for 8 tissues (AC, abdomen without digestive or reproductive system; DG, digestive plus excretory system; GE, genitalia; GO, gonad; HD, head; RE, reproductive system without gonad; TX, thorax without digestive system; WB, whole body) for both sex in *D. melanogaster* and *D. pseudoobscura* were obtained from the Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo) database (accession ID: GSE99574) submitted by a *Drosophila* genome re-annotation project (Yang et al. 2018).

We obtained a total of 13,638 ortholog pairs between *D. melanogaster* and *D. pseudoobscura* from FlyBase (https://flybase.org/) *D. melanogaster* Orthologs gene sets. Of these ortholog pairs, we retrieved 10,921 one-to-one ortholog pairs that are contained in the expression data described above. For each pair of orthologs we computed the correlation of expression values across matching tissues and sexes as Pearson's correlation coefficient.

Assembly-based structural variations detection

Discovery of structural variations between genome assemblies was performed using a custom pipeline (Liao et al. 2018; Kou et al. 2019) based on LASTZ/CHAIN/NET/NETSYNTENY tools (Schwartz et al. 2003; Harris 2007). Briefly, soft masked genomes (transposon repeats, simple repeats and centromeric repeats were masked out) were aligned by the LASTZ (Version 1.04) program with output of Axt format. The result alignments were used to build the larger fragment chains if two matching alignments next to each other are close enough using axtChain, and then the built chains were sorted and merged into a single file using chainMergeSort. Next, the low scoring chains were filtered out by chainPreNet and the remaining were used to build the net alignments using chainNet. Synteny information was added using netSyntenic. Finally, custom perl scripts were run on the final syntenic format file to annotate a simple catalogue of structural variants, including insertions, deletions, inversions, copy number variations (CNVs) and complex SVs with vague breakpoints. More details about the pipeline are available at https://github.com/yiliao1022/LASTZ_SV_pipeline.

Rearrangement breakpoints and their distribution at the TADs

We obtained highly contiguous genome assemblies of 15 *Drosophila* species from Miller et al (2018) and the genome assembly of *D. miranda* from Mahajan et al (2018). All these assemblies were generated using single-molecule sequencing technology and have an average contig N50 larger than 4Mb. We aligned all these assemblies to the *D. melanogaster* ISO1 reference assembly and our new *D. pseudoobscura* assembly, respectively. With the LASTZ/CHAIN/NET/NETSYNTENY tools (see above), we generated the "netSyntenic'' file for each assembly. We then used a custom perl script to extract the conserved syntenic blocks and identify the synteny breaks, each syntenic blocks requiring at least 10 kb in length for both reference and the query assemblies. Synteny breaks were classified into synteny breaks if the breaks obtained from "fills" whose syntenic information was annotated as "top", "syn" or "NonSyn", and inversion breaks if the breaks obtained from "fills" whose syntenic information was annotated as "inv". Because the assemblies are not

scaffolded into chromosomes (except for *D. miranda*), we only considered the synteny breaks within contigs. To do this, synteny breaks identified within the 10kb terminal regions of each contig were excluded from analysis due to these may come from erroneous assembly.

To quantify the number of synteny breaks at TADs or around TAD boundaries, we used the method as previously described (Krefting et al, 2018) with minor modification. Instead of the single base breakpoint, we used a 10 kb region which was obtained by extending 5Kb upstream and downstream for each breadpoint to represent the break region. We enlarged TAD domains by 50% of their length on each side and then subdivided this range into 20 equal sized bins. Next, we computed the number of overlaps of break regions and the 20 bins using Bedtools. We also generated a background control by simulating 100 times of the same number of synteny breaks as the same number, size and chromosomal distribution of the actual synteny breaks and computed the distribution of random synteny breaks around TADs in the same way as done for the actual synteny breaks.

Quantification of structural variations at the TAD boundaries

We applied two approaches to evaluate the relative abundance of structural variations at the TAD boundaries. SVs fall into the heterochromatin and centromeric regions were excluded from the analysis, as these regions may be more prone to variants artifacts. First, we calculated the observed/expected count and base coverage for each type of SVs using the formula: $(\sum_{i \in k} N_i)/(N_{total} \sum_{i \in k} \frac{S_i}{S_{total}})$ (Fudenberg and Pollard, 2019), where i indexes genomic regions annotated as TAD boundaries, S_{total} is the genome size, and N_{total} is the total number and base coverage of each type of SVs genomewide. Second, we performed permutation analyses comparing the number of SVs count and base coverage overlap with TAD boundaries to the number of overlaps of SVs count and base coverage overlap with 10,000 sets of random regions that had the same size and chromosomal distribution as the TAD boundaries.

Code availability

The code that reproduces analyses from the manuscript is available at https://github.com/yiliao1022/TAD_SV.

Data availability

All raw genomic data, Hi-C data and ISO-seq data have been deposited to NCBI under the BioProject PRJNA596268.

Disclosure declaration

The authors do not declare any conflict of interest.

Acknowledgements

This work was funded by the National Institutes of Health (R01GM123303 to J.J.E.) and National Science Foundation (IOS-1656260 to J.J.E.) and funding from the University of California, Irvine to J.J.E. We thank the Genomics High Throughput Facility at University of California, Irvine (UCI) and Arima Genomics, Inc. San Diego for expert service. We would also like to thank Luna Thanh Ngo for help with data collection and management.

Author contribution: J.J.E. and Y. L. conceived of the presented idea. M.C. contributed to the genomic data collection and genome assembly. Y.L., X.Z. and J.J.E. analyzed the data. Y.L. wrote the manuscript with support from X.Z., M.C. and J.J.E.

REFERENCE

Akdemir KC, Chin L. 2015. HiCPlotter integrates genomic data with interaction matrices. *Genome Biol* **16**: 198.

Akdemir KC, Le VT, Chandran S, Li Y, Verhaak RG, Beroukhim R, Campbell PJ, Chin L, Dixon JR, Futreal PA, et al. 2020. Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nat Genet* **52**: 294–305.

AlHaj Abed J, Erceg J, Goloborodko A, Nguyen SC, McCole RB, Saylor W, Fudenberg G, Lajoie BR, Dekker J, Mirny LA, et al. 2019. Highly structured homolog pairing reflects functional organization of the Drosophila genome. *Nat Commun* **10**: 4485.

Arzate-Mejía RG, Josué Cerecedo-Castillo A, Guerrero G, Furlan-Magaril M, Recillas-Targa F. 2020. In situ dissection of domain boundaries affect genome topology and gene transcription in Drosophila. *Nat Commun* **11**: 894.

Berthelot C, Muffato M, Abecassis J, Roest Crollius H. 2015. The 3D organization of chromatin explains evolutionary fragile genomic regions. *Cell Rep* **10**: 1913–1924.

Bintu B, Mateo LJ, Su J-H, Sinnott-Armstrong NA, Parker M, Kinrot S, Yamaya K, Boettiger AN, Zhuang X. 2018. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362**. http://dx.doi.org/10.1126/science.aau1783.

Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv X, Hugnot J-P, Tanay A, et al. 2017. Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**: 557–572.e24.

Bracewell R, Chatla K, Nalley MJ, Bachtrog D. 2019. Dynamic turnover of centromeres drives karyotype evolution in Drosophila. *Elife* **8**. http://dx.doi.org/10.7554/eLife.49002.

Campbell MS, Holt C, Moore B, Yandell M. 2014. Genome Annotation and Curation Using MAKER and MAKER-P. *Curr Protoc Bioinformatics* **48**: 188.

Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. 2016. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res* **44**: e147.

Chakraborty M, Chang C-H, Khost DE, Vedanayagam J, Adrion JR, Liao Y, Montooth K, Meiklejohn CD, Larracuente AM, Emerson JJ. 2020. Evolution of genome structure in the Drosophila simulans species complex. *bioRxiv* 2020.02.27.968743. https://www.biorxiv.org/content/10.1101/2020.02.27.968743v1.abstract (Accessed March 24, 2020).

Chakraborty M, Emerson JJ, Macdonald SJ, Long AD. 2019. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat Commun* **10**: 4872.

Chakraborty M, VanKuren NW, Zhao R, Zhang X, Kalsow S, Emerson JJ. 2018. Hidden genetic variation shapes the structure of functional elements in Drosophila. *Nat Genet* **50**: 20–25.

Chathoth KT, Zabet NR. 2019. Chromatin architecture reorganization during neuronal cell differentiation in *Drosophila* genome. *Genome Res* **29**: 613–625.

Crow JF, Kimura M, Others. 1970. An introduction to population genetics theory. *An introduction to population genetics theory*. https://www.cabdirect.org/cabdirect/abstract/19710105376.

Cubeñas-Potts C, Rowley MJ, Lyu X, Li G, Lei EP, Corces VG. 2017. Different enhancer classes in Drosophila bind distinct architectural proteins and mediate unique chromatin interactions and 3D architecture. *Nucleic Acids Res* **45**: 1714–1730.

Dali R, Blanchette M. 2017. A critical assessment of topologically associating domain prediction

tools. *Nucleic Acids Res* **45**: 2994–3005.

Despang A, Schöpflin R, Franke M, Ali S, Jerković I, Paliou C, Chan W-L, Timmermann B, Wittler L, Vingron M, et al. 2019. Functional dissection of the Sox9–Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nature Genetics* **51**: 1263–1271. http://dx.doi.org/10.1038/s41588-019-0466-z.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**: 376–380.

Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. 2016. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**: 95–98.

Eres IE, Luo K, Hsiao CJ, Blake LE, Gilad Y. 2019. Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates. *PLoS Genet* **15**: e1008278.

Filippova D, Patro R, Duggal G, Kingsford C. 2014. Identification of alternative topological domains in chromatin. *Algorithms Mol Biol* **9**: 14.

Finn EH, Misteli T. 2019. Molecular basis and biological function of variability in spatial genome organization. *Science* **365**: eaaw9498. http://dx.doi.org/10.1126/science.aaw9498.

Fishman V, Battulin N, Nuriddinov M, Maslova A, Zlotina A, Strunov A, Chervyakova D, Korablev A, Serov O, Krasikova A. 2019. 3D organization of chicken genome demonstrates evolutionary conservation of topologically associated domains and highlights unique architecture of erythrocytes' chromatin. *Nucleic Acids Res* **47**: 648–665.

Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S. 2017. Comparison of computational methods for Hi-C data analysis. *Nat Methods* **14**: 679–685.

Fudenberg G, Pollard KS. 2019. Chromatin features constrain structural variation across evolutionary timescales. *Proc Natl Acad Sci U S A* **116**: 2175–2180.

Ghavi-Helm Y, Jankowski A, Meiers S, Viales RR, Korbel JO, Furlong EEM. 2019. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nat Genet* **51**: 1272–1282.

Gong Y, Lazaris C, Sakellaropoulos T, Lozano A, Kambadur P, Ntziachristos P, Aifantis I, Tsirigos A. 2018. Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries. *Nature Communications* **9**. http://dx.doi.org/10.1038/s41467-018-03017-1.

Harmston N, Ing-Simmons E, Tan G, Perry M, Merkenschlager M, Lenhard B. 2017. Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. *Nat Commun* **8**: 441.

Harris RS. 2007. Improved pairwise Alignmnet of genomic DNA. https://etda.libraries.psu.edu/catalog/7971.

Hou C, Li L, Qin ZS, Corces VG. 2012. Gene density, transcription, and insulators contribute to the

partition of the Drosophila genome into physical domains. *Mol Cell* **48**: 471–484.

Hsieh T-HS, Weiner A, Lajoie B, Dekker J, Friedman N, Rando OJ. 2015. Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* **162**: 108–119.

Hug CB, Grimaldi AG, Kruse K, Vaquerizas JM. 2017. Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription. *Cell* **169**: 216–228.e19.

Jacobs J, Atkins M, Davie K, Imrichova H, Romanelli L, Christiaens V, Hulselmans G, Potier D, Wouters J, Taskiran II, et al. 2018. The transcription factor Grainy head primes epithelial enhancers for spatiotemporal activation by displacing nucleosomes. *Nat Genet* **50**: 1011–1020.

Kim K, Eom J, Jung I. 2019. Characterization of Structural Variations in the Context of 3D Chromatin Structure. *Mol Cells* **42**: 512–522.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. http://genome.cshlp.org/content/early/2017/03/15/gr.215087.116.abstract.

Kou Y, Liao Y, Toivainen T, Lv Y, Tian X, Emerson JJ. 2019. Evolutionary genomics of structural variation in Asian rice (Oryza sativa) and its wild progenitor (O. rufipogon). *bioRxiv*. https://www.biorxiv.org/content/10.1101/2019.12.19.883231v1.abstract.

Krefting J, Andrade-Navarro MA, Ibn-Salem J. 2018. Evolutionary stability of topologically associating domains is associated with conserved gene regulation. *BMC Biol* **16**: 87.

Lam K-K, LaButti K, Khalak A, Tse D. 2015. FinisherSC: a repeat-aware tool for upgrading de novo assembly using long reads. *Bioinformatics* **31**: 3207–3209.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Lazar NH, Nevonen KA, O'Connell B, McCann C, O'Neill RJ, Green RE, Meyer TJ, Okhovat M, Carbone L. 2018. Epigenetic maintenance of topological domains in the highly rearranged gibbon genome. *Genome Res* **28**: 983–997.

Le TBK, Imakaev MV, Mirny LA, Laub MT. 2013. High-resolution mapping of the spatial organization of a bacterial chromosome. *Science* **342**: 731–734.

Liao Y, Zhang X, Li B, Liu T, Chen J, Bai Z, Wang M, Shi J, Walling JG, Wing RA, et al. 2018. Comparison of Oryza sativa and Oryza brachyantha Genomes Reveals Selection-Driven Gene Escape from the Centromeric Regions. *Plant Cell* **30**: 1729–1744.

Li, Li L, Lyu X, Hou C, Takenaka N, Nguyen HQ, Ong C-T, Cubeñas-Potts C, Hu M, Lei EP, et al. 2015. Widespread Rearrangement of 3D Chromatin Organization Underlies Polycomb-Mediated Stress-Induced Silencing. *Molecular Cell* **58**: 216–231. http://dx.doi.org/10.1016/j.molcel.2015.02.023.

Liu C, Cheng Y-J, Wang J-W, Weigel D. 2017. Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis. *Nat Plants* **3**: 742–748.

Liu C, Wang C, Wang G, Becker C, Zaidem M, Weigel D. 2016. Genome-wide analysis of chromatin packing in Arabidopsis thaliana at single-gene resolution. *Genome Res* **26**: 1057–1068.

Ludwig MZ, Bergman C, Patel NH, Kreitman M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**: 564–567.

Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R, et al. 2015. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**: 1012–1025.

Mahajan S, Wei KH-C, Nalley MJ, Gibilisco L, Bachtrog D. 2018. De novo assembly of a young Drosophila Y chromosome using single-molecule sequencing and chromatin conformation capture. *PLoS Biol* **16**: e2006348.

Marchal C, Sima J, Gilbert DM. 2019. Control of DNA replication timing in the 3D genome. *Nat Rev Mol Cell Biol* **20**: 721–737.

Miller DE, Staber C, Zeitlinger J, Hawley RS. 2018. Highly Contiguous Genome Assemblies of 15 Drosophila Species Generated Using Nanopore Sequencing. *G3* **8**: 3131–3141.

Mizuguchi T, Fudenberg G, Mehta S, Belton J-M, Taneja N, Folco HD, FitzGerald P, Dekker J, Mirny L, Barrowman J, et al. 2014. Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe. *Nature* **516**: 432–435.

Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P. 2013. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**: 59–64.

Ni X, Zhang YE, Nègre N, Chen S, Long M, White KP. 2012. Adaptive evolution and the birth of CTCF binding sites in the Drosophila genome. *PLoS Biol* **10**: e1001420.

Ou S, Su W, Liao Y, Chougule K, Ware D, Peterson T, Jiang N, Hirsch CN, Hufford MB. 2019. Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline. *bioRxiv* 657890. https://www.biorxiv.org/content/10.1101/657890v1 (Accessed December 5, 2019).

Pal K, Forcato M, Jost D, Sexton T, Vaillant C, Salviato E, Mazza EMC, Lugli E, Cavalli G, Ferrari F. 2019. Global chromatin conformation differences in the Drosophila dosage compensated chromosome X. *Nat Commun* **10**: 5355.

Pope BD, Ryba T, Dileep V, Yue F, Wu W, Denas O, Vera DL, Wang Y, Hansen RS, Canfield TK, et al. 2014. Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**: 402–405.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, Habermann B, Akhtar A, Manke T. 2018. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun* **9**: 189.

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T.

2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**: W160–5.

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**: 1665–1680.

Renschler G, Richard G, Valsecchi CIK, Toscano S, Arrigoni L, Ramírez F, Akhtar A. 2019. Hi-C guided assemblies reveal conserved regulatory topologies on X and autosomes despite extensive genome shuffling. *Genes Dev* **33**: 1591–1612.

Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, Nielsen R, Thornton K, Hubisz MJ, Chen R, Meisel RP, et al. 2005. Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. *Genome Res* **15**: 1–18.

Rowley MJ, Corces VG. 2018. Organizational principles of 3D genome architecture. *Nat Rev Genet* **19**: 789–800.

Sadowski M, Kraft A, Szalaj P, Wlasnowolski M, Tang Z, Ruan Y, Plewczynski D. 2019. Spatial chromatin architecture alteration by structural variations in human genomes at the population scale. *Genome Biol* **20**: 148.

Schmitt AD, Hu M, Ren B. 2016. Genome-wide mapping and analysis of chromosome architecture. *Nature Reviews Molecular Cell Biology* **17**: 743–755. http://dx.doi.org/10.1038/nrm.2016.104.

Schoenfelder S, Fraser P. 2019. Long-range enhancer--promoter contacts in gene expression control. *Nat Rev Genet* 1.

Schuettengruber B, Oded Elkayam N, Sexton T, Entrevan M, Stern S, Thomas A, Yaffe E, Parrinello H, Tanay A, Cavalli G. 2014. Cooperativity, specificity, and evolutionary stability of Polycomb targeting in Drosophila. *Cell Rep* **9**: 219–233.

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. 2003. Human-mouse alignments with BLASTZ. *Genome Res* **13**: 103–107.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. 2012. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**: 458–472.

Shanta O, Noor A, Human Genome Structural Variation Consortium (HGSVC), Sebat J. 2020. The effects of common structural variants on 3D chromatin structure. *BMC Genomics* **21**: 95.

Spielmann M, Lupiáñez DG, Mundlos S. 2018. Structural variation in the 3D genome. *Nat Rev Genet* **19**: 453–467.

Szabo Q, Jost D, Chang J-M, Cattoni DI, Papadopoulos GL, Bonev B, Sexton T, Gurgo J, Jacquier C, Nollmann M, et al. 2018. TADs are 3D structural units of higher-order chromosome organization inDrosophila. *Science Advances* **4**: eaar8082. http://dx.doi.org/10.1126/sciadv.aar8082.

Tao J-F, Zhou J-Z, Xie T, Wang X-T, Yang Q-Y, Zhang H-Y. 2017. Influence of Chromatin 3D Organization on Structural Variations of the Arabidopsis thaliana Genome. *Mol Plant* **10**: 340–344.

Thomas G, Hahn M. 2017. Drosophila 25 species phylogeny. https://figshare.com/articles/Drosophila_25_species_phylogeny/5450602.

Ulianov SV, Khrameeva EE, Gavrilov AA, Flyamer IM, Kos P, Mikhaleva EA, Penin AA, Logacheva MD, Imakaev MV, Chertovich A, et al. 2016. Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Res* **26**: 70–84.

van Steensel B, Furlong EEM. 2019. The role of transcription in shaping the spatial organization of the genome. *Nat Rev Mol Cell Biol* **20**: 327–337.

Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, Hadjur S. 2015. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep* **10**: 1297–1309.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS One* **9**: e112963.

Wang Q, Sun Q, Czajkowsky DM, Shao Z. 2018. Sub-kb Hi-C in D. melanogaster reveals conserved characteristics of TADs between insect and mammalian cells. *Nat Commun* **9**: 188.

Xie T, Zhang F-G, Zhang H-Y, Wang X-T, Hu J-H, Wu X-M. 2019. Biased gene retention during diploidization in Brassica linked to three-dimensional genome organization. *Nat Plants* **5**: 822–832.

Yang H, Jaime M, Polihronakis M, Kanegawa K, Markow T, Kaneshiro K, Oliver B. 2018. Re-annotation of eight Drosophila genomes. *Life Sci Alliance* **1**: e201800156.

Yang J, Ramos E, Corces VG. 2012. The BEAF-32 insulator coordinates genome organization and function during the evolution of Drosophila species. *Genome Res* **22**: 2199–2207.

Yang Y, Zhang Y, Ren B, Dixon JR, Ma J. 2019. Comparing 3D Genome Organization in Multiple Species Using Phylo-HMRF. *Cell Syst* **8**: 494–505.e14.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Zhang Y, McCord RP, Ho Y-J, Lajoie BR, Hildebrand DG, Simon AC, Becker MS, Alt FW, Dekker J. 2012. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* **148**: 908–921.

Zheng H, Xie W. 2019. The role of 3D genome organization in development and cell differentiation. *Nat Rev Mol Cell Biol* **20**: 535–550.

Zufferey M, Tavernari D, Oricchio E, Ciriello G. 2018. Comparison of computational methods for the identification of topologically associating domains. *Genome Biol* **19**: 217.

Supplemental Materials For:

"Topologically associating domains and their role in the evolution of genome structure and function in *Drosophila*"

Yi Liao[1,*], Xinwen Zhang[1], Mahul Chakraborty[1] and J.J. Emerson[1,2,*]

[1]Department of Ecology and Evolutionary Biology, University of California, Irvine, CA, USA.

[2]Center for Complex Biological Systems, University of California, Irvine, CA, USA.

*Email to jje@uci.edu or liaoy12@uci.edu

Supplemental Methods

Genome Annotation

Simple Tandem repeats, centromeric and telomeric repeats identification

Tandem repeats were identified by Tandem repeat finder (Benson 1999) with parameters "2 7 7 80 10 50 2000 -f -d -m". The top 5 abundant tandemly repeated units of consensus pattern were used as a query for further searching the whole genome to look for their chromosomal distribution and the exact number present using RepeatMasker. We identified two centromere tandem repeats, with motif size 170bp and 21bp, respectively. The former is only detected at the X and dot chromosomes and is unique to *D. pseudoobscura* relative to *D. miranda*, suggesting the rapid turnover of centromeric repeats between species and between chromosomes within species in *Drosophila* genomes.

Transposable elements annotation

TE annotation was performed with the EDTA pipeline. But we also run other softwares for comparison. FIrstly, we used RepeatModeler version 1.0.11 (Smit, Hubley, and Green 2015) for *de novo* repeat sequence identification with the default parameters. The resulted raw repeat library was further filtered by excluding candidates that have blastn hits with any annotated genes. In total, RepeatModeler generated 924 repeat families totaling 1,149,638 bp. We used LTR_retriever pipeline (Ou and Jiang 2018) to annotate long terminal repeat (LTR) retrotransposons for both *D. miranda* and *D. pseudoobscura* genome. Potential LTR retrotransposons were first identified using LTRharvest (Ellinghaus, Kurtz, and Willhoeft 2008) with parameters: -similar 90 -vic 10 -seed 20 -seqids yes -minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1, and LTR_FINDER (Xu and Wang 2007) with parameters: -D 15000 -d 1000 -L 7000 -l 100 -p 20 -C -M 0. The output was then processed by LTR_retriever to identify the intact LTR-RT elements and construct a LTR sequence library. We identified a total of 470 and 1,927 (of them, 1,406 from the new Y chromosome) full LTR-RT elements for *D. pseudoobscura* and *D. miranda* , respectively. We also used MITE_Hunter(Han and Wessler 2010) and GRF https://github.com/bioinfolabmu/GenericRepeatFinder to find Miniature Inverted repeat Transposable Elements (MITE). These two programs generated a total of 149 families of MITEs sequences. We further annotated the 924 *de novo* repeat families from RepeatModeler by blasting these repeat sequences to Repbase, LTR retrotransposon lib and the MITE sequence. Finally, we classified 454 LTR elements and 343 DNA elements, leaving 123 elements as unknown elements. A secondary custom repeat sequence library was generated by integrating Repbase, LTR_retriver ,MITE annotation and RepeatModeler output.

Gene annotation

In this study, we performed *de novo* gene annotation for both *D. miranda* (Mahajan et al. 2018)and *Drosophila pseudoobscura* genome sequences using Maker (version 2.31.8)(Campbell et al. 2014) . For *D. miranda*, we first built transcriptome assembly from two publicly available RNA-seq datasets (Nozawa et al. 2016; Mahajan et al. 2018)-each including 12 and 21 samples,respectively-which covered a wide range of tissues and developmental stages in both male and female. RNA-seq reads from each sample were individually aligned to the genome assembly using HiSat2 (Kim, Langmead, and Salzberg 2015) and assembled by StringTie (Pertea et al. 2015) with default parameters. The resulting gtf files for each dataset were then merged to a single file using "stringtie --merge" command included in the StringTie package, respectively. To prepare transcript sequences for Maker, the merged gtf files were translated to the gff file using "gffread" and Fasta sequences were extracted using ReSeqTools (He et al. 2013) based on the gff file and reference genome sequence. We used the amino acid sequence from *D. melanogaster* (Flybase r6.26) and *D. pseudoobscura* (r3.04) as the protein evidence. For *D. pseudoobscura*, we directly used transcripts from Flybase (r3.04) and a recent re-annotation project (H. Yang et al. 2018), each containing 23,456 and 39,527 transcripts, respectively. In addition, we assembled the transcriptome using RNA-seq data from a previous study (Nozawa et al. 2016). Moreover, in order to get a more accurate gene models and identify high-confidential alternative splice transcripts, a total of 15,372 and 22,237 Full-length Isoforms obtained from our ISO-seq experiment in male and female sample, respectively, were also used to guide the gene models construction in Maker pipeline. Protein evidence included translated sequence from *D. melanogaster* (Flybase r6.26) and *D. miranda* annotated in this study. The Maker was run three times for each species.

Supplemental Tables

Supplemental Table S1: Whole genome sequencing data used for *de novo* assembling and verifying the *D. pseudoobscura* genome.

PacBio			Illumina paired-end			Hi-C		
Reads	Mean Length	Depth	Pairs	Length	Depth	Pairs	Length	Depth
3,170,985	14,493 bp	~280	154,932,139	150bp	~283	397,020,516	150bp	~726

* Depth is calculated based on the genome size of 164 MB.

Supplemental Table S2: Contig and chromosome assembly statics for the *D. pseudoobscura* genome.

Features	Scaffold	ConTig
Number	70	72
Total Bases	163,274,969	163,282,969
N10	68,158,638	35,530,770
N30	68,158,638	32,422,566
N50	32,422,566	30,706,867
N70	30,706,867	23,510,042
N80	30,706,867	22,902,128
N90	23,510,042	9,717,740
Min	1,117	1,117
Max	68,158,638	35,530,770

		Contig Number	NCBI accession
Chr2	32,422,566	1	CM020868.1
Chr3	23,510,042	1	CM020869.1
Chr4	30,706,867	1	CM020870.1
Chr5	1,881,070	1	CM020871.1
ChrX	68,158,638	3	CM020872.1
Mitochondrion	16,118	1	CM020873.2
Unplaced	6,587,668	64	-

Supplemental Table S3: Quality evaluation of the *D. pseudoobscura* genome assembly.

Quality Value (QV) = -10log$_{10}$(*Probability of error*)		52	
		Diptera_odb9	**Arthropoda_odb9**
BUSCO score	Complete BUSCOs (C)	2731 (97.6%)	1061 (99.5%)
	Complete and single-copy BUSCOs (S)	2700 (96.5%)	1047 (98.2%)
	Fragmented BUSCOs (F)	31 (1.1%)	14 (1.3%)
	Missing BUSCOs (M)	42 (1.5%)	1 (0.1%)
	Total BUSCO groups searched	26 (0.9%)	4 (0.4%)
		2799	1066

Supplemental Table S4: Sequence composition of the 64 unplaced contigs.

Overall TE occupancy	Centromeric repeats: 3639793 bp 54 % Simple repeats: 71045 bp 1 % Small RNA: 263447 bp 4 % Retroelements: 2211064 bp 33 % DNA transposon: 26375 bp 0.4% Unclassified: 453477bp 6.8 %		
	170bp	**21bp**	**Both**
Contigs contains centromere-specific tandem repeats	WVEN01000009.1 WVEN01000011.1 WVEN01000012.1 WVEN01000013.1 WVEN01000014.1 WVEN01000015.1 WVEN01000016.1 WVEN01000018.1 WVEN01000020.1 WVEN01000022.1 WVEN01000028.1 WVEN01000041.1 WVEN01000042.1 WVEN01000044.1 WVEN01000045.1 WVEN01000046.1 WVEN01000047.1 WVEN01000048.1 WVEN01000049.1 WVEN01000050.1 WVEN01000054.1 WVEN01000055.1 WVEN01000056.1 WVEN01000057.1 WVEN01000058.1 WVEN01000065.1 WVEN01000066.1	WVEN01000008.1 WVEN01000019.1 WVEN01000021.1 WVEN01000023.1 WVEN01000029.1 WVEN01000060.1 WVEN01000061.1	WVEN01000026.1 WVEN01000043.1 WVEN01000059.1

*Contig names are from NCBI assignment.

Supplemental Table S5: Gene and TE annotation of the *D. pseudoobscura* genome

Gene Models	**Number**	**BUSCO score (Diptera_odb9)**
Chr2	3,287	C:95.0%[S:62.8%,D:32.2%],F:2.5%,M:2.5%,n:2799
Chr3	2,545	2658 Complete BUSCOs (C)
Chr4	2,564	1757 Complete and single-copy BUSCOs (S)
Chr5	74	901 Complete and duplicated BUSCOs (D)
X	4,292	70 Fragmented BUSCOs (F)
Mitochondrion	13	71 Missing BUSCOs (M)
Unplaced	638	2799 Total BUSCO groups searched
Total	13,413	

Major Repetitive sequences	Centromeric repeats: 4,851,739 bp 2.97%
	Simple repeats: 7,656,035 bp 4.69%
	Retro-elements: 22,773,268 bp 13.95%
	DNA transposons: 1,985,898 bp 1.22%
	Unclassified 6,660,145 bp 4.08%

Supplemental Table S6: RNA sequencing data used for gene annotation using the MAKER pipeline.

Data	NCBI accession		Reference
Published RNA-seq	From DRR055234 To DRR055275	abdomens; Male accessory glands; abdomens without gonads; imaginal discs; heads; 3rd instar larvae; 3rd instar larvae without imaginal discs; Female ovaries; pupae; thoraxes; Male testes; adult;	(Nozawa et al. 2016)
Published transcripts			
39,527 transcripts	GSE99574	Details see GSE99574	(H. Yang et al. 2018)
23,456 transcripts	NA	NA	FlyBase r3.04
ISO-seq			
15,372 transcripts	PRJNA596268	Male adult full body	This study
22,237 transcripts	PRJNA596268	Female adult full body	This study

Supplemental Table S7: Quality control on Hi-C data.

	HiCExplorer		**3dDNA**
Pairs considered	397,020,516	**Sequenced Read Pairs**	397,020,516
Pairs mappable, unique and high quality	304,513,515 (76.70%)	Alignable (Normal+Chimeric Paired)	337,748,332 (85.1%)
Pairs used (Hi-C Contacts)	194,589,768 (**49.01%**)	Paired used (Hi-C Contacts)	204,080,597 (**51.4%**)
Inter chromosomal	9,275,009 (2.3%)	Inter chromosomal	7,746,309 (2.0%)
Intra chromosomal		Intra chromosomal	
Short range (<20kb)	74,173,026 (18.7%)	Short range (<20kb)	99,160,171 (25.0%)
Long range	111,141,733 (28.0%)	Long range	97,172,702 (24.5%)

Supplemental Table 8: Combinations of parameter values tested for each TAD annotation tool.

HiCExplorer	**hicFindTADs --correctForMultipleTesting fdr** --**m** hic_corrected.h5 | hic_corrected_5K.h5 | hic_corrected_10K.h5 | hic_corrected_20K.h5 **--delta** 0.01 | 0.04 **--minBoundaryDistance** 5000 | 20000 **--thresholdComparisons** 0.05 | 0.01 | 0.005 | 0.001 **hicFindTADs -m hic_corrected_5K.h5 --correctForMultipleTesting fdr --delta 0.04 --minBoundaryDistance 20000 --thresholdComparisons 0.001**
Juicer Arrowhead	**java -Xmx20g -jar /data/apps/juicer/1.5.6/scripts/juicer_tools.jar arrowhead -k NONE --ignore_sparsity inter_30.hic dpse_arrowhead** **-r** 5000 | 10000 **ava -Xmx20g -jar /data/apps/juicer/1.5.6/scripts/juicer_tools.jar arrowhead -k NONE -r 5000 --ignore_sparsity inter_30.hic dpse_arrowhead**
Armatus	**Armatus -m -S** **-r** 5000 | 10000 | 20000 | 25000 | 40000 | 50000 **-s** 0.1 | 0.05 **-g** 1.0 | 2.0 | 0.9 **armatus -S -g 0.9 -r 5000 -s 0.1 -i 5K.matrix -o 5K.out**

Supplemental Table S9: Synteny information between *D. melanogaster* and *D. pseudoobscura*.

A. *D. melanogaster to D. pseudoobscura*

	Within homologous chromosomes		Outside of homologous chromosomes		
	Number of Syntenic blocks	Syntenic coverage	Number of Syntenic blocks	Syntenic coverage	Overall Base Coverage
2L	163	19,489,042	1	81,462	16,062,093
2R	157	17,105,335	1	16,682	14,630,878
3L	152	18,854,583	5	69,854	16,580,574
3R	210	24,793,351	0	0	21,277,942
4	29	895,323	0	0	406,947
X	217	15,483,353	50	3,243,056	12,146,805
Total	**928**	**96,620,987**	**57**	**3,411,054**	**81,105,239**

B. *D. pseudoobscura to D. melanogaster*

	Within homologous chromosomes		Outside of homologous chromosomes		
	Number of Syntenic blocks	Syntenic coverage	Number of Syntenic blocks	Syntenic coverage	Overall Base Coverage
Chr2	210	26,818,840	6	129,583	22,275,738
Chr3	157	17,119,684	0	0	15,090,175
Chr4	163	23,133,097	1	4,662	17,083,083
Chr5	29	846379	0	0	456,074
XR	152	20,096,070	50	4,026,659	19,857,440
XL	217	17,063,934	0	0	10,747,531
Total	**928**	**105,078,004**	**57**	**4,160,904**	**85,510,041**

Supplemental Table S10: TADs identified in different cell lines/tissues and tools.

Species/ cell lines	Armatus			HiCExporer			Juicer/Arrowhead			Data from
	Num.	**Syn.**	**Cov.**	**Num.**	**Syn.**	**Cov.**	**Num.**	**Syn.**	**Cov.**	
Dmel										
Kc167	914	786	95 Mb	933	641	122 Mb	699	574	68 Mb	1
BG3	979	878	95 Mb	964	688	121 Mb	610	531	63 Mb	1
S2	1,005	858	82 Mb	1,107	791	120 Mb	625	516	61 Mb	2
Dpse										
Full body	1,136	875	71 Mb	1013	689	101 Mb	926	680	66 Mb	This study

Num., Number of TADs ; Syn., Synteny in genome of the other species, Cov., genome coverage. 1(Chathoth and Zabet 2019); 2(Wang et al. 2018).

Supplementary Table S11: Published ChIP-seq or ChIP-chip data sources used in this study.

Species	Cell lines /tissue	Architectural proteins	GEO accession	Reference
D. melanogaster	S2	BEAF32	GSE20760	(Schwartz et al. 2012)
D. melanogaster	S2	CP190	GSE20766	(Schwartz et al. 2012)
D. melanogaster	S2	CTCF	GSE32818	(Riddle et al. 2011)
D. melanogaster	S2	Chromator	GSE20765	modENCODE
D. melanogaster	S2	Su(Hw)	GSE32813	(Riddle et al. 2011)
D. melanogaster	S2	Trl	GSE32822	(Riddle et al. 2011)
D. melanogaster	Kc167	BEAF32	GSM762845	(Wood et al. 2011)
D. melanogaster	Kc167	CP190	GSM762836	(Wood et al. 2011)
D. melanogaster	Kc167	CTCF	GSM1535983	(Wood et al. 2011)
D. melanogaster	Kc167	Chromator	GSM1318357/GSE20763	(Wood et al. 2011)
D. melanogaster	Kc167	Su(Hw)	GSM762839/GSE51964	(Wood et al. 2011)
D. melanogaster	Kc167	Trl	modEncode_3245	modENCODE
D. melanogaster	BG3	BEAF32	GSE20811	(Schwartz et al. 2012)
D. melanogaster	BG3	CP190	GSE20814	(Schwartz et al. 2012)
D. melanogaster	BG3	CTCF	GSE20767	(Schwartz et al. 2012)
D. melanogaster	BG3	Chromator	GSE20761	modENCODE
D. melanogaster	BG3	Su(Hw)	GSE20833	(Schwartz et al. 2012)
D. melanogaster	BG3	Trl	GSE23466	modENCODE
D. pseudoobscura	Embryos	BEAF-32	-	(J. Yang, Ramos, and
D. pseudoobscura	white pre-pupae (WPP)	CTCF	-	Corces 2012)
				(Ni et al. 2012)

Supplementary Table S12: Conservation of TAD boundaries between *D. melanogaster* and *D. pseudoobscura* in comparisons between different categories (Supplement to Figure3).

A Comparison of TAD boundaries that overlapped with insulators.

A1: Kc167

	Insulator	Conserved boundaries that overlapped the insulator	Total number that boundaries that overlapped with the insulator	Percent	Conserved boundaries that don't overlapped the insulator	Total number that boundaries That don't overlapped with the insulator	Percent
HiCExplorer	**BEAF**	278	456	0.61	93	196	0.47
	CTCF	163	273	0.60	208	379	0.55
	Chro	325	532	0.61	46	120	0.38
	SuHw	227	392	0.58	144	260	0.55
	GAF	124	210	0.59	247	442	0.564
	CP190	304	519	0.59	67	133	0.50
Juicer	**BEAF**	252	401	0.63	367	662	0.55
	CTCF	167	258	0.65	452	805	0.56
	Chro	335	520	0.64	284	534	0.53
	SuHw	272	452	0.60	347	611	0.57
	GAF	119	221	0.54	500	842	0.59
	CP190	329	525	0.63	290	538	0.542
Armatus	**BEAF**	534	747	0.71	315	635	0.50
	CTCF	320	438	0.73	529	944	0.56
	Chro	654	918	0.71	195	464	0.42
	SuHw	481	727	0.66	368	655	0.56
	GAF	270	404	0.67	579	978	0.59
	CP190	622	901	0.69	227	481	0.47

A.2 BG3

HiCExplorer	**BEAF**	60	98	0.61	332	623	0.537

	CTCF	78	154	0.51	314	567	0.55
	Chro	349	602	0.58	43	119	0.36
	SuHw	101	194	0.527	291	527	0.55
	GAF	178	320	0.56	214	401	0.53
	CP190	254	428	0.59	138	293	0.47
Juicer	**BEAF**	36	68	0.53	514	933	0.55
	CTCF	102	169	0.60	448	832	0.54
	Chro	320	546	0.59	230	455	0.51
	SuHw	158	290	0.54	392	711	0.55
	GAF	170	313	0.54	380	688	0.55
	CP190	228	389	0.59	322	612	0.53
Armatus	**BEAF**	102	149	0.68	772	1352	0.57
	CTCF	162	280	0.58	712	1221	0.58
	Chro	704	1061	0.66	170	440	0.39
	SuHw	240	475	0.51	634	1026	0.62
	GAF	366	609	0.60	508	892	0.57
	CP190	497	778	0.64	377	723	0.52

A3. S2

HiCExplorer	**BEAF**	328	580	0.57	77	215	0.36
	CTCF	49	103	0.48	356	692	0.51
	Chro	367	648	0.57	38	147	0.26
	SuHw	68	143	0.48	337	652	0.52
	GAF	190	361	0.53	215	434	0.50
	CP190	336	620	0.54	69	175	0.39
Juicer	**BEAF**	262	459	0.57	234	525	0.457
	CTCF	51	101	0.50	445	883	0.50
	Chro	325	566	0.57	171	418	0.41
	SuHw	90	183	0.49	406	801	0.51
	GAF	137	279	0.49	359	705	0.51
	CP190	294	516	0.57	202	468	0.43

Armatus	**BEAF**	594	904	0.66	215	563	0.38
	CTCF	100	175	0.57	709	1292	0.55
	Chro	662	1021	0.65	147	446	0.33
	SuHw	159	345	0.46	650	1122	0.58
	GAF	301	516	0.58	508	951	0.53
	CP190	629	1036	0.611	180	431	0.42

B. Comparison of strong and weak TAD boundaries.

	Total number of strong boundaries	Number of conserved strong boundaries	Percent	Total number of weak boundaries	Number of conserved weak boundaries	Percent
Dmel to Dpse						
Kc167	292	196	0.67	393	186	0.47
BG3	347	220	0.63	396	181	0.46
S2	430	261	0.61	394	153	0.39
Dpse to Dmel						
Kc167						
BG3	367	235	0.64	326	130	0.40
S2	367	225	0.61	326	132	0.40
	367	243	0.66	326	142	0.44

C. Conservation degree of TAD boundaries that are shared across cell lines or specific to certain cell lines.

Methods	Total of shared boundaries	Conserved of shared boundaries	Percent	Total of cell specific boundaries	Conserved of specific boundaries	Percent
Armatus	1757	1180	0.67	943	405	0.43
Hicexplorer	672	392	0.58	483	165	0.34
Juicer	1229	753	0.61	795	299	0.38

D. Conservation degree of TAD boundaries that have different flanking chromatin modifications.

Chromatin combinations	**Total number of the category**	**Conserved number of the category**	**Percent of the conserved**
Active_Active	659	262	0.40
Active_PcG	342	152	0.44
Active_Inactive	665	287	0.43
Inactive_Inactive	233	73	0.31
Inactive_PcG	166	62	0.37
PcG_PcG	141	33	0.23
Total	2245	881	0.39
Randomization test	2245	476	**0.21**

Supplementary Table S13: 129 long genes concincied with full TADs in *D. melanogaster*.

Dmel_geneID	Gene_name	Chromosome	Start	End	Dpse_ortholog	Dpse_new_geneID	Chromosome	Start	End
FBgn0000479									
FBgn0003256									
FBgn0004198									
FBgn0052580									
FBgn0085638									
FBgn0266199									
FBgn0267033									
FBgn0267336									
FBgn0284236									
FBgn0003963	ush	2L	476220	540560	FBgn0075472	LOC4817316	NC_046681.1	18886568	18962765
FBgn0041111	lilli	2L	2885929	2954406	FBgn0081331	LOC4816631	NC_046681.1	5205103	5303786
FBgn0015600	toc	2L	3068345	3144786	FBgn0246841	LOC6903567	NC_046681.1	10160973	10211507
FBgn0051774	fred	2L	3902928	3991713	FBgn0076481	LOC4816819	NC_046681.1	19828341	19951473
FBgn0000547	ed	2L	4031377	4115749	FBgn0071799	LOC4817240	NC_046681.1	19639096	19758672
FBgn0003716	tkv	2L	5219000	5271384	FBgn0072755	LOC4817227	NC_046681.1	24852925	24919964
FBgn0085450	Snoo	2L	7891187	7984196	FBgn0080091	LOC4817376	NC_046681.1	23141739	23255965
FBgn0011259	Sema1a	2L	8542147	8672041	FBgn0074948	LOC4816208	NC_046681.1	14892768	15012903
FBgn0041092	tai	2L	9166778	9250211	FBgn0072098	LOC4816259	NC_046681.1	15388231	15481384
FBgn0032151	nAChRalpha6	2L	9793317	9886250	FBgn0077986	LOC4817310	NC_046681.1	18110288	18208112
FBgn0051721	Trim9	2L	10544883	10628054	FBgn0079380	LOC4816752	NC_046681.1	6696732	6769563
FBgn0259176	bun	2L	12455540	12546611	FBgn0247162	LOC6902896	NC_046681.1	20992639	21111231
FBgn0263354	CG42784	2L	13415431	13500427	FBgn0246640	LOC6903438	NC_046681.1	12668130	12693940
FBgn0259984	kuz	2L	13550139	13639411	FBgn0080133	LOC4818148	NC_046681.1	12755253	12871144
FBgn0028509	CenG1A	2L	13835564	13898712	FBgn0076510	LOC4816443	NC_046681.1	3648729	3710205
FBgn0261563	wb	2L	14263027	14328262	FBgn0073665	LOC4816591	NC_046681.1	4185195	4261530
FBgn0003016	osp	2L	14599196	14689340	FBgn0077483	LOC4818069	NC_046681.1	11488589	11613073
FBgn0028863	CG4587	2L	15644183	15718272	FBgn0078284	LOC4816419	NC_046681.1	5020033	5066453
FBgn0259735	mtgo	2L	16545016	16663218	FBgn0247420	LOC6903125	NC_046681.1	26136170	26267233
FBgn0261804	CG42750	2L	18159811	18275770	FBgn0080155	LOC4817320	NC_046681.1	22488603	22544365
FBgn0058006	CG40006	2L	22622417	22756349	FBgn0265242	LOC6902757	NC_046681.1	2952011	2974803
FBgn0263780	CG17684	2R	1866080	2262115	FBgn0246910	LOC6903318	NC_046681.1	17564336	17570127
FBgn0260995	dpr21	2R	3066499	3191011	FBgn0272884	LOC26533897	NC_046681.1	17771326	17783592
FBgn0040849	Ir41a	2R	4917519	5021955	FBgn0245983	LOC6898024	NC_046680.1	3767355	3769951
FBgn0053554	Nipped-A	2R	5177965	5251022	FBgn0245413	LOC26533571	NC_046680.1	1398559	1548443
FBgn0000546	EcR	2R	6087873	6169087	FBgn0243551	LOC6899016	NC_046680.1	21585943	21671815
FBgn0086655	jing	2R	6502006	6621792	FBgn0263814	LOC6899052	NC_046680.1	21772166	21902528

FBgn0003090	pk	2R	7150703	7223378	FBgn0070804	LOC4803477	NC_046680.1	4966792	5020419
FBgn0033438	Mmp2	2R	9611139	9683858	FBgn0245614	LOC6898791	NC_046680.1	17885883	17960741
FBgn0261698	SLO2	2R	10322195	10416996	FBgn0071943	LOC4804232	NC_046680.1	10892329	10984572
FBgn0263102	psq	2R	10557888	10617280	FBgn0264580	LOC6898873	NC_046680.1	19855085	19902889
FBgn0053144	CG33144	2R	10740054	10798709	FBgn0077331	LOC4805321	NC_046680.1	19716315	19755708
FBgn0000633	CG17716	2R	13623044	13749367	FBgn0074649	LOC4805164	NC_046680.1	18229569	18358516
FBgn0265974	ttv	2R	14526267	14587917	FBgn0070143	LOC4803961	NC_046680.1	8788651	8853115
FBgn0264089	sli	2R	15869495	15922118	FBgn0263810	LOC4805561	NC_046680.1	21507934	21562752
FBgn0285917	sbb	2R	18278198	18357296	FBgn0245672	LOC6898650	NC_046680.1	15255540	15329626
FBgn0003435	sm	2R	19517864	19631508	FBgn0264769	LOC4804945	NC_046680.1	16339804	16441091
FBgn0010415	Sdc	2R	21392402	21481176	FBgn0070413	LOC4805278	NC_046680.1	19298122	19391643
FBgn0266129	lov	2R	25085780	25139627	FBgn0265027	LOC4804618	NC_046680.1	13839121	13849504
FBgn0035094	CG9380	2R	25183628	25255318	FBgn0081730	LOC4804620	NC_046680.1	13921798	13927626
FBgn0001316	klar	3L	434083	540614	FBgn0074321	LOC117184056	NC_046679.1	17232815	17233267
FBgn0267487	Ptp61F	3L	1342493	1475257	FBgn0081581	LOC4812118	NC_046683.1	64577786	64585987
FBgn0266696	Svil	3L	2377655	2499226	FBgn0076834	LOC4812804	NC_046683.1	55075034	55211476
FBgn0263392	Tet	3L	2786207	2879158	FBgn0249681	LOC6900088	NC_046683.1	65557238	65659194
FBgn0262593	Shab	3L	2895017	2959579	FBgn0249911	LOC6900092	NC_046683.1	65471782	65527489
FBgn0262870	axo	3L	4629604	4687317	FBgn0074906	LOC4812589	NC_046683.1	59089541	59153919
FBgn0035574	RhoGEF64C	3L	4692783	4796253	FBgn0249899	LOC6900063	NC_046683.1	59158565	59284682
FBgn0085447	sif	3L	5657913	5749599	FBgn0078858	LOC4813607	NC_046683.1	50614964	50710357
FBgn0020251	sfl	3L	6495680	6550104	FBgn0080991	LOC4814130	NC_046683.1	66608453	66672946
FBgn0261788	Ank2	3L	7655389	7718395	FBgn0245004	LOC6900405	NC_046683.1	44181432	44250541
FBgn0016694	Pdp1	3L	7811458	7867369	FBgn0249858	LOC4812101	NC_046683.1	57972383	58017080
FBgn0011817	nmo	3L	7979046	8054147	FBgn0249744	LOC6899982	NC_046683.1	57785259	57853064
FBgn0085385	bma	3L	9908120	9967298	FBgn0244878	LOC6900645	NC_046683.1	50839836	50897100
FBgn0041096	rols	3L	12008225	12064859	FBgn0076690	LOC4812739	NC_046683.1	60074493	60128691
FBgn0260941	app	3L	12206238	12266841	FBgn0079009	LOC4813857	NC_046683.1	45062249	45128414
FBgn0264001	bru3	3L	13521907	13803400	FBgn0262083	LOC6900229	NC_046683.1	62237844	62554451
FBgn0087007	bbg	3L	14412828	14536276	FBgn0081889	LOC4813512	NC_046683.1	49738548	49857900
FBgn0259175	ome	3L	14672839	14747868	FBgn0076726	LOC4813899	NC_046683.1	43509178	43577286
FBgn0261090	Sytbeta	3L	15014695	15064977	FBgn0075740	LOC4813079	NC_046683.1	43181391	43225404
FBgn0260943	Rbp6	3L	17058774	17238212	FBgn0244751	LOC6900901	NC_046683.1	55683847	55875214
FBgn0016797	fz2	3L	19140975	19235373	FBgn0081984	LOC4812483	NC_046683.1	61926741	61929000
FBgn0262737	mub	3L	21844788	21932702	FBgn0080346	LOC4813190	NC_046683.1	47711136	47763777
FBgn0004449	Ten-m	3L	22293044	22407863	FBgn0246553	LOC6903600	NC_046683.1	56595766	56737063
FBgn0266347	nAChRalpha4	3L	23206669	23287448	FBgn0071670	LOC4811728	NC_046679.1	18978998	19036653
FBgn0287185	nvd	3L	24324866	24401655	FBgn0272278	LOC26533291	NC_046679.1	19042190	19046707
FBgn0267429	lovit	3L	25236940	25670980	FBgn0263149	LOC6897476	NC_046679.1	19205188	19215147

FBgn0287183	l(3)80Fg	3L	25798441	26014112	FBgn0270973	LOC26531986	NC_046679.1	18926361	18949519
FBgn0267430	Pzl	3R	2554162	3263582	FBgn0246990	LOC6899548	NC_046679.1	26030363	26054844
FBgn0263346	smash	3R	4659579	4712193	FBgn0263296	LOC6897751	NC_046679.1	25020967	25036713
FBgn0013576	mtd	3R	5270116	5350443	FBgn0248772	LOC6897619	NC_046679.1	22529751	22600472
FBgn0083949	side-III	3R	5862865	5959555	FBgn0248644	LOC6897387	NC_046679.1	16350190	16436589
FBgn0083963	Nlg3	3R	7570281	7639434	FBgn0247584	LOC6897868	NC_046679.1	27586270	27676356
FBgn0266801	CG45263	3R	8364328	8477617	FBgn0076231	LOC4802637	NC_046679.1	24064667	24177705
FBgn0261929	5-HT2B	3R	8578205	8629996	FBgn0080752	LOC4802031	NC_046679.1	17586939	17634157
FBgn0262614	pyd	3R	8827068	8931896	FBgn0263173	LOC4802625	NC_046679.1	23926197	24041582
FBgn0003165	pum	3R	9066343	9237682	FBgn0263302	LOC6899331	NC_046679.1	25150068	25361033
FBgn0001235	hth	3R	10507561	10639568	FBgn0074358	LOC4801015	NC_046679.1	7419554	7526298
FBgn0037963	Cad87A	3R	11893512	11949129	FBgn0248851	LOC6897810	NC_046679.1	26173815	26206576
FBgn0264493	rdx	3R	13967061	14032057	FBgn0082107	LOC4801646	NC_046679.1	13899676	13970122
FBgn0285955	cv-c	3R	14391683	14481713	FBgn0076192	LOC4801949	NC_046679.1	16632194	16724764
FBgn0263929	jvl	3R	14741582	14795868	FBgn0077532	LOC4801060	NC_046679.1	7994919	8070854
FBgn0266756	btsz	3R	14804092	14875488	FBgn0248064	LOC6897024	NC_046679.1	8084537	8136265
FBgn0259244	CG42342	3R	16502770	16571363	FBgn0262050	LOC6896920	NC_046679.1	5650062	5712218
FBgn0263995	cpo	3R	17919832	18018892	FBgn0076133	LOC4803162	NC_046679.1	29609529	29684944
FBgn0004652	fru	3R	18414273	18545586	FBgn0262676	LOC4801442	NC_046679.1	12163419	12258033
FBgn0263974	qin	3R	18591027	18650137	FBgn0072931	LOC4801434	NC_046679.1	11990290	12061241
FBgn0003118	pnt	3R	23290231	23346167	FBgn0262662	LOC6897112	NC_046679.1	9813046	9884417
FBgn0051145	CG31145	3R	23605751	23669659	FBgn0248188	LOC6896811	NC_046679.1	3196002	3278400
FBgn0004509	Fur1	3R	25347431	25473058	FBgn0263364	LOC6897901	NC_046679.1	28168848	28337718
FBgn0011666	msi	3R	25516118	25608423	FBgn0248760	LOC6897597	NC_046679.1	21943710	22050910
FBgn0039431	plum	3R	26470078	26535217	FBgn0079629	LOC4801081	NC_046679.1	8225629	8289170
FBgn0004842	RYa-R	3R	26996980	27057935	FBgn0262494	LOC13036374	NC_046679.1	8807721	8868130
FBgn0004369	Ptp99A	3R	29377646	29487131	FBgn0263238	LOC4800475	NC_046679.1	2268638	2383280
FBgn0266411	sima	3R	30058311	30121798	FBgn0080705	LOC4800293	NC_046679.1	729199	793552
FBgn0010113	hdc	3R	30277932	30372382	FBgn0262041	LOC6896726	NC_046679.1	361603	444621
FBgn0261988	Gprk2	3R	31405245	31457873	FBgn0074786	LOC4800727	NC_046679.1	4750576	4790127
FBgn0284435	tyn	X	142208	200663	FBgn0244270	LOC6901851	NC_046683.1	22322415	22367122
FBgn0052816	CG32816	X	316576	470726	FBgn0077175	LOC4815282	NC_046683.1	14038292	14061888
FBgn0264449	CG43867	X	808217	927975	FBgn0243670	LOC6901139	NC_046683.1	39399697	39543032
FBgn0025390	Mur2B	X	1523554	1668788	FBgn0244397	LOC6902079	NC_046683.1	40512867	40519529
FBgn0283741	prage	X	1773719	1853667	FBgn0073294	LOC4815930	NC_046683.1	4167796	4202646
FBgn0052791	DIP-alpha	X	3507245	3571649	FBgn0244620	LOC6901640	NC_046683.1	9241195	9268292
FBgn0266429	AstA-R1	X	3574536	3666724	FBgn0075512	LOC4814982	NC_046683.1	9393504	9418311
FBgn0283657	Tlk	X	3720007	3789770	FBgn0247156	LOC6901632	NC_046683.1	9488755	9586869
FBgn0000635	Fas2	X	4132887	4206093	FBgn0077610	LOC4814416	NC_046683.1	39288156	39368197

FBgn0029881	pigs	X	6600062	6653015	FBgn0077823	LOC4815731	NC_046683.1	22102264	22159609
FBgn0261873	sdt	X	8178549	8240474	FBgn0077114	LOC4814721	NC_046683.1	14168679	14235710
FBgn0261260	mgl	X	9358496	9500129	FBgn0071481	LOC4816002	NC_046683.1	2983510	3010620
FBgn0052698	CARPB	X	9783223	9855449	FBgn0243764	LOC6901062	NC_046683.1	1115681	1163504
FBgn0259170	alpha-Man-Ia	X	10266792	10325958	FBgn0077085	LOC4811663	NC_046683.1	9116064	9182650
FBgn0085443	spri	X	10490069	10585422	FBgn0247809	LOC6901478	NC_046683.1	12339033	12451658
FBgn0267001	Ten-a	X	12043698	12335214	FBgn0077061	LOC4815803	NC_046683.1	42182177	42397715
FBgn0259171	Pde9	X	12785563	12897327	FBgn0077055	LOC4814257	NC_046683.1	38534763	38676799
FBgn0264078	Flo2	X	14839376	14933946	FBgn0244429	LOC6902140	NC_046683.1	4923878	5025517
FBgn0000535	eag	X	14955477	15009989	FBgn0244430	LOC6902148	NC_046683.1	4848589	4889314
FBgn0285944	para	X	16455027	16533096	FBgn0082095	LOC4814623	NC_046683.1	15180603	15252555
FBgn0266354	CG45002	X	16969069	17066732	FBgn0248789	LOC6901200	NC_046683.1	17251805	17361961
FBgn0003380	Sh	X	17924307	18063247	FBgn0247929	LOC6901448	NC_046683.1	12898779	13029681
FBgn0265598	Bx	X	18515428	18572809	FBgn0247290	LOC6902855	NC_046681.1	20210853	20212694
FBgn0264090	CG43759	X	18930335	19028629	FBgn0076989	LOC4814891	NC_046683.1	7320033	7439102

Supplementary Table S14: 73 long genes concincied with full TADs in *D. pseudoobscura*.

Dpse_ID	Chromosome	Start	End	Flybase_geneID	Dmel_ortholog	Dmel_gene_name	Chromosome	Start	End
LOC6896811	NC_046679.1	3196002	3278400	FBgn0248188	FBgn0051145	CG31145	3R	23605751	23669659
LOC4800781	NC_046679.1	5386959	5438514	FBgn0077383	FBgn0053208	Mical	3R	10001554	10042895
LOC6896916	NC_046679.1	5461116	5578166						
LOC6896920	NC_046679.1	5650062	5712218	FBgn0262050	FBgn0259244	CG42342	3R	16502770	16571363
LOC4800948	NC_046679.1	6763020	6834216	FBgn0248101	FBgn0037546	mAChR-B	3R	8059284	8087049
LOC4801003	NC_046679.1	7189395	7246437	FBgn0079305	FBgn0004876	cdi	3R	19044882	19094270
LOC6897007	NC_046679.1	7732765	7790727	FBgn0248077	FBgn0262617	Nuak1	3R	10264306	10308467
LOC4801081	NC_046679.1	8225629	8289170	FBgn0079629	FBgn0039431	plum	3R	26470078	26535217
LOC6897112	NC_046679.1	9813046	9884417	FBgn0262662	FBgn0003118	pnt	3R	23290231	23346167
LOC4801646	NC_046679.1	13899676	13970122	FBgn0082107	FBgn0264493	rdx	3R	13967061	14032057
LOC4801764	NC_046679.1	14784567	14835704	FBgn0247893	FBgn0053100	eIF4EHP	3R	24063077	24108960
LOC4801843	NC_046679.1	15359555	15533548	FBgn0075943	FBgn0011224	heph	3R	31846987	32015520
LOC6897387	NC_046679.1	16350190	16436589	FBgn0248644	FBgn0083949	side-III	3R	5862865	5959555
LOC4801949	NC_046679.1	16632194	16724764	FBgn0076192	FBgn0285955	cv-c	3R	14391683	14481713
LOC4802074	NC_046679.1	17817447	17894364	FBgn0243536	FBgn0003944	Ubx	3R	16656623	16734426
LOC4802304	NC_046679.1	21082351	21136455	FBgn0261650	FBgn0085386	CG34357	3R	4491943	4558067
LOC4802625	NC_046679.1	23926197	24041582	FBgn0263173	FBgn0262614	pyd	3R	8827068	8931896
LOC6899331	NC_046679.1	25150068	25361033	FBgn0263302	FBgn0003165	pum	3R	9066343	9237682
LOC26534216	NC_046679.1	26295215	26353745	FBgn0247625	FBgn0264357	SNF4Agamma	3R	21140739	21214269
LOC6897868	NC_046679.1	27586270	27676356	FBgn0247584	FBgn0083963	Nlg3	3R	7570281	7639434
LOC6897901	NC_046679.1	28168848	28337718	FBgn0263364	FBgn0004509	Fur1	3R	25347431	25473058
LOC4803056	NC_046679.1	28349791	28404616	FBgn0263363	FBgn0262582	cic	3R	20252770	20303942
LOC6899355	NC_046679.1	28625913	28734691	FBgn0249366	FBgn0038755	Hs6st	3R	19929277	20008512
LOC4803264	NC_046679.1	30602446	30780262	FBgn0263373	FBgn0260003	Dys	3R	19461085	19597288
LOC4803646	NC_046680.1	6343075	6410100	FBgn0075887	FBgn0040752	Prosap	2R	14060456	14140902
LOC4804232	NC_046680.1	10892329	10984572	FBgn0071943	FBgn0261698	SLO2	2R	10322195	10416996
LOC6898462	NC_046680.1	11786375	11861598	FBgn0245753	FBgn0013733	shot	2R	13864237	13942110
LOC4804516	NC_046680.1	12998411	13067112	FBgn0072699	FBgn0033405	CG13954	2R	9309296	9389467
LOC6898564	NC_046680.1	13825579	13905008	FBgn0246243	FBgn0054038	CG34038	2R	25167659	25169855
LOC6898642	NC_046680.1	15005173	15069436	FBgn0245675	FBgn0002643	mam	2R	13991203	14060356
LOC6898647	NC_046680.1	15163374	15249936	FBgn0245674	FBgn0053958	CG33958	2R	18184382	18192851
LOC4804945	NC_046680.1	16339804	16441091	FBgn0264769	FBgn0003435	sm	2R	19517864	19631508
LOC6898791	NC_046680.1	17885883	17960741	FBgn0245614	FBgn0033438	Mmp2	2R	9611139	9683858
LOC4805263	NC_046680.1	19176666	19251051	FBgn0075833	FBgn0041239	Gr58a	2R	21979860	21981099
LOC4805278	NC_046680.1	19298122	19391643	FBgn0070413	FBgn0010415	Sdc	2R	21392402	21481176
LOC6898874	NC_046680.1	19914625	19976308	FBgn0264579	FBgn0283521	lola	2R	10481894	10543291
LOC4805422	NC_046680.1	20762914	20815098	FBgn0072148	FBgn0033667	reb	2R	11904323	11912779
LOC4805561	NC_046680.1	21507934	21562752	FBgn0263810	FBgn0264089	sli	2R	15869495	15922118
LOC6899052	NC_046680.1	21772166	21902528	FBgn0263814	FBgn0086655	jing	2R	6502006	6621792
LOC4805739	NC_046680.1	23319395	23407484	FBgn0080886	FBgn0023441	fus	2R	15657318	15676669
LOC4816443	NC_046681.1	3648729	3710205	FBgn0076510	FBgn0028509	CenG1A	2L	13835564	13898712
LOC4816591	NC_046681.1	4185195	4261530	FBgn0073665	FBgn0261563	wb	2L	14263027	14328262
LOC6903561	NC_046681.1	10275293	10366559	FBgn0246567	FBgn0016977	spen	2L	159032	203397
LOC4818069	NC_046681.1	11488589	11613073	FBgn0077483	FBgn0003016	osp	2L	14599196	14689340
LOC4817896	NC_046681.1	12699950	12749976	FBgn0081619	FBgn0023407	B4	2L	13500624	13549328

LOC4818148	NC_046681.1	12755253	12871144	FBgn0080133	FBgn0259984	kuz	2L	13550139	13639411
LOC26533542	NC_046681.1	14293071	14349942						
LOC4816208	NC_046681.1	14892768	15012903	FBgn0074948	FBgn0011259	Sema1a	2L	8542147	8672041
LOC4816259	NC_046681.1	15388231	15481384	FBgn0072098	FBgn0041092	tai	2L	9166778	9250211
LOC6902405	NC_046681.1	15579131	15732776	FBgn0249459	FBgn0000464	Lar	2L	19586623	19732069
LOC4817865	NC_046681.1	16393823	16474646	FBgn0076373	FBgn0266521	stai	2L	6100377	6124653
LOC4817310	NC_046681.1	18110288	18208112	FBgn0077986	FBgn0032151	nAChRalpha6	2L	9793317	9886250
LOC4816903	NC_046681.1	18235141	18286585	FBgn0075064	FBgn0028704	Nckx30C	2L	9711512	9746495
LOC4817316	NC_046681.1	18886568	18962765	FBgn0075472	FBgn0003963	ush	2L	476220	540560
LOC4817453	NC_046681.1	19120954	19177483	FBgn0078274	FBgn0004611	Plc21C	2L	305935	355566
LOC6902868	NC_046681.1	20531988	20656719	FBgn0272756					
LOC6902896	NC_046681.1	20992639	21111231	FBgn0247162	FBgn0259176	bun	2L	12455540	12546611
LOC4817376	NC_046681.1	23141739	23255965	FBgn0080091	FBgn0085450	Snoo	2L	7891187	7984196
LOC6903125	NC_046681.1	26136170	26267233	FBgn0247420	FBgn0259735	mtgo	2L	16545016	16663218
LOC4817542	NC_046681.1	26739952	26812700	FBgn0073723	FBgn0053516	dpr3	2L	2058790	2109878
LOC6903295	NC_046681.1	29801894	29908866	FBgn0249104	FBgn0034106	CG9068	2R	16283435	16287450
LOC4811795	NC_046682.1	784082	839864	FBgn0074773	FBgn0085432	pan	4	69326	114270
LOC4815782	NC_046683.1	3474993	3530448	FBgn0074848	FBgn0004370	Ptp10D	X	11622015	11677338
LOC6901756	NC_046683.1	6654611	6736259	FBgn0249613	FBgn0029946	CG15034	X	7346259	7347506
LOC6901632	NC_046683.1	9488755	9586869	FBgn0247156	FBgn0283657	Tlk	X	3720007	3789770
LOC6901448	NC_046683.1	12898779	13029681	FBgn0247929	FBgn0003380	Sh	X	17924307	18063247
LOC26532449	NC_046683.1	25938122	26018820						
LOC6901093	NC_046683.1	38284146	38417666	FBgn0243698	FBgn0265597	rad	X	12988615	13079070
LOC6901139	NC_046683.1	39399697	39543032	FBgn0243670	FBgn0264449	CG43867	X	808217	927975
LOC6902018	NC_046683.1	41866256	41951315	FBgn0244155	FBgn0264542	hwt	X	12516634	12568452
LOC4813557	NC_046683.1	45301101	45446946	FBgn0076932	FBgn0262509	nrm	3L	22998372	23040300
LOC6900681	NC_046683.1	51484002	51610780	FBgn0245227	FBgn0052062	Rbfox1	3L	10481412	10594012
LOC4812408	NC_046683.1	61629954	61730871	FBgn0076771	FBgn0052206	CG32206	3L	19340321	19412912

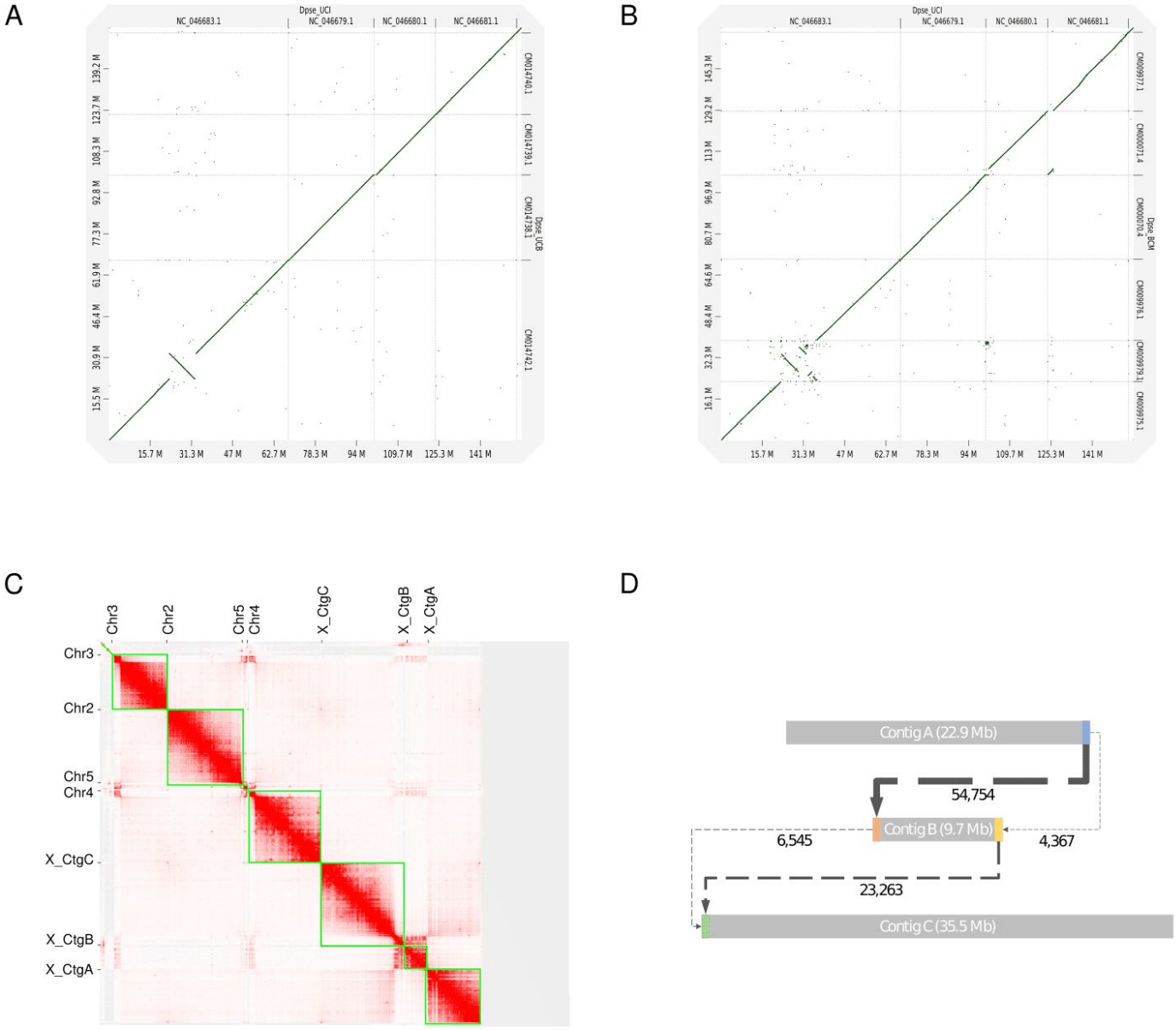
Supplemental Table S15 : Number of different categories of syntenic blocks (>10Kb) identified by aligning 16 *Drosophila* genome assemblies to *D. melanogaster* and *D. pseudoobscura*, respectively.

	Divergence Time (Mys)	Top fills	Nonsyn fills	Syntenic fills	Inv fills	Total syntenic breaks	Total Inversion breaks
D. melanogaster versus							
D. simulans	3.53	101	9	3	13	121	13
D. mauritiana	3.53	144	12	3	14	167	14
D. sechellia	3.53	123	8	4	10	143	10
D. erecta	7.25	90	6	4	22	108	22
D. yakuba	7.25	112	10	5	16	135	16
D. eugracilis	14.27	178	55	4	19	245	19
D. biarmipes	15.24	85	158	18	38	269	38
D. triauraria	Unknown	627	115	9	36	759	36
D. ananassae	34.37	189	237	59	106	493	106
D. bipectinata	34.37	540	219	7	22	774	22
D. pseudoobscura	49	**169**	**9**	**246**	**314**	**432**	**314**
D. persimilis	49	385	281	35	101	709	101
D. miranada	49	**187**	**30**	**223**	**308**	**448**	**308**
D. willistoni	62.1	1001	163	8	63	1180	63
D. mojavensis	72.63	879	123	16	58	1026	58
D. virilis	72.63	729	184	41	91	962	91
D. pseudoobscura versus							
D. simulans	49	232	324	94	144	678	172
D. mauritiana	49	265	286	101	164	681	193
D. sechellia	49	197	323	112	159	661	188
D. melanogaster	49	**181**	**4**	**256**	**331**	**469**	**359**
D. erecta	49	260	221	113	190	624	220
D. yakuba	49	352	317	49	115	748	145
D. eugracilis	49	555	290	16	54	889	82
D. biarmipes	49	353	392	44	105	819	135
D. triauraria	49	722	265	15	36	1032	66
D. ananassae	49	302	307	80	162	717	190
D. bipectinata	49	769	234	3	33	1033	60
D. persimilis	1.96	292	80	24	10	446	60
D. miranada	2.76	**60**	**120**	**29**	**54**	**259**	**104**
D. willistoni	62.1	1072	135	10	66	1242	91
D. mojavensis	72.63	941	95	21	50	1085	78
D. virilis	72.63	723	217	43	77	1010	104

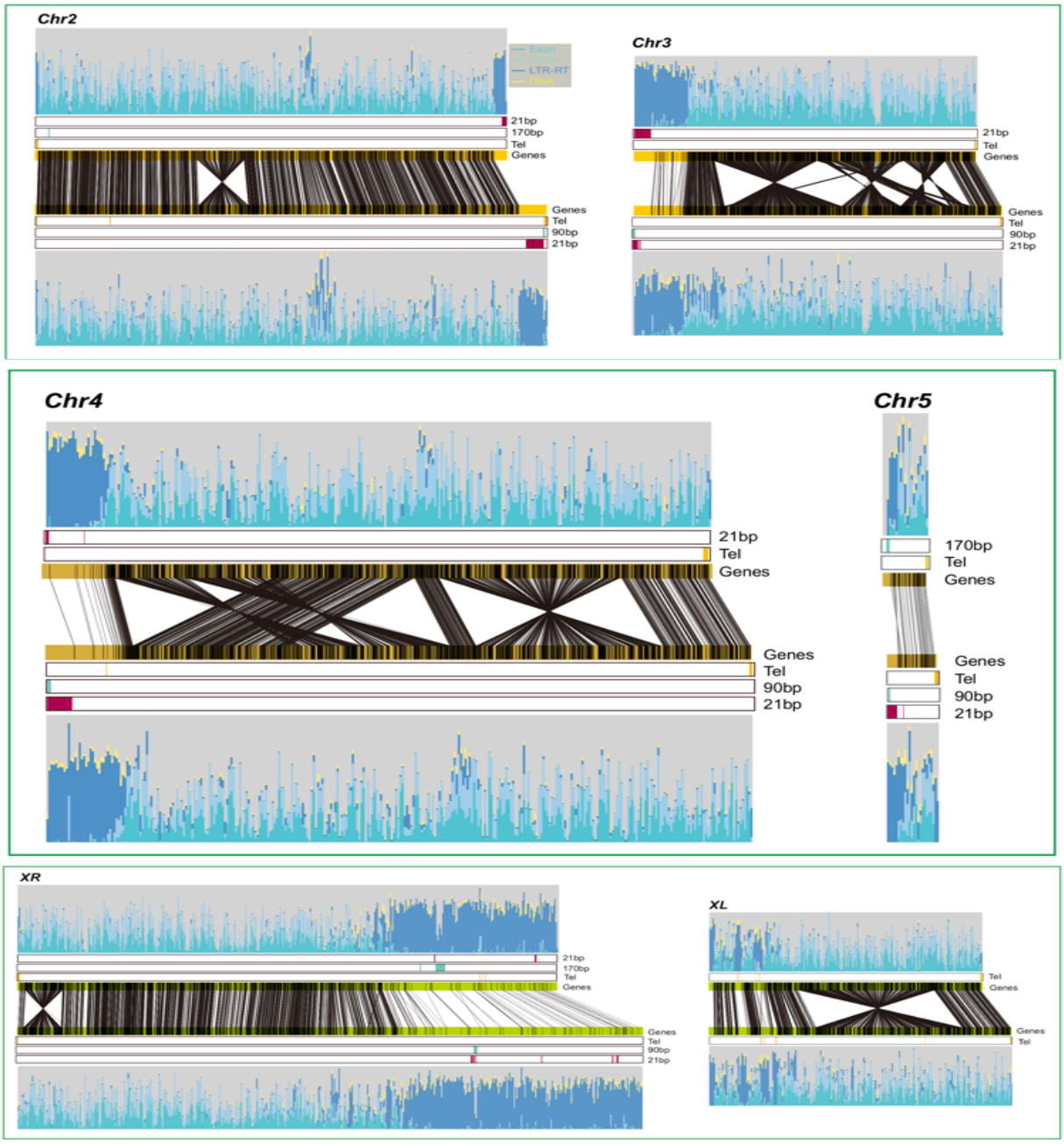
Supplementary Table S16: SVs detected within the euchromatic region were used for analysis

Chromosome	**Coordinates**
2L	1..22200000
2R	5000000..25479258
3L	1..23400000
3R	4000000..31815305
X	1..21900000

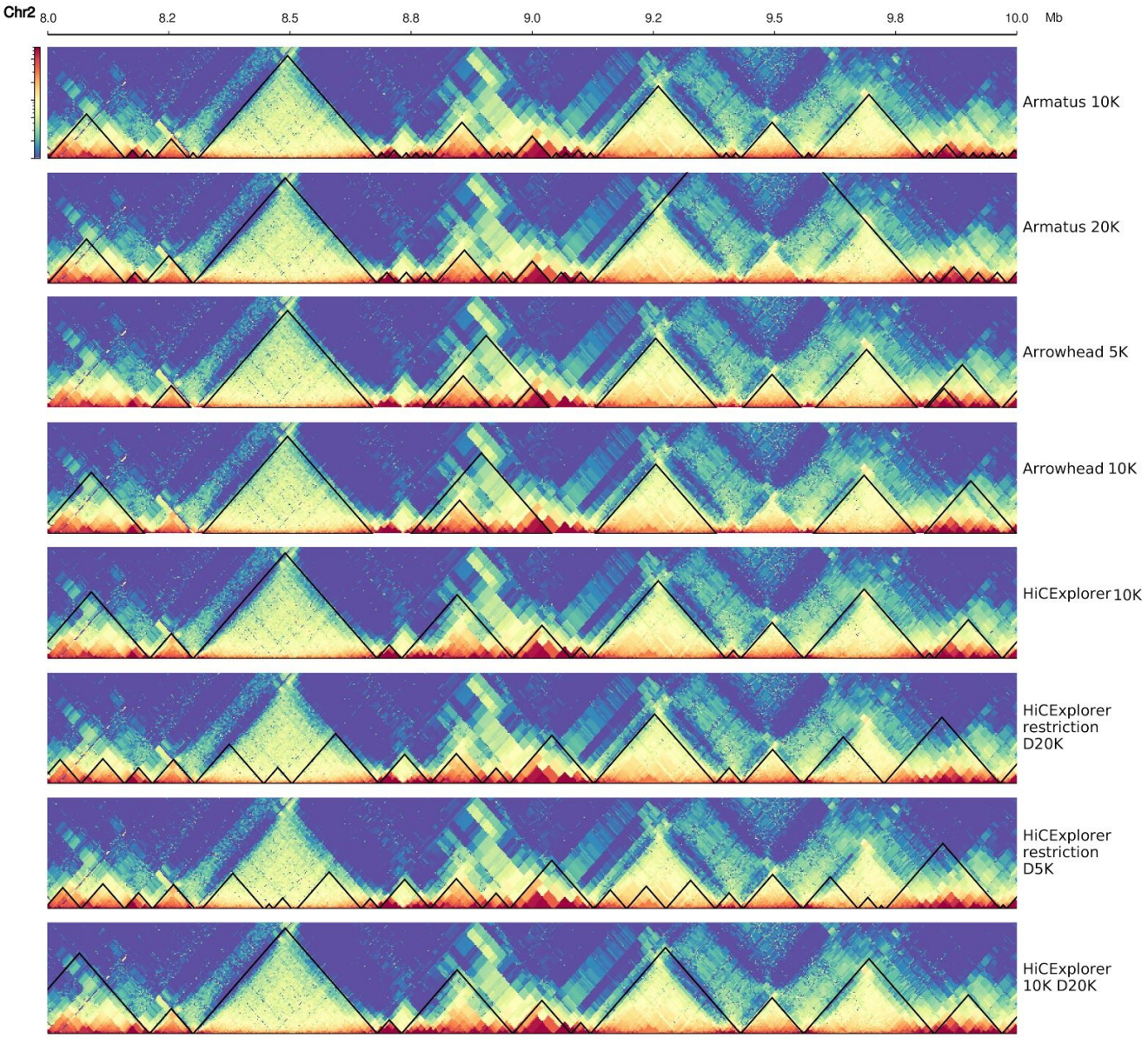
Supplemental Figures



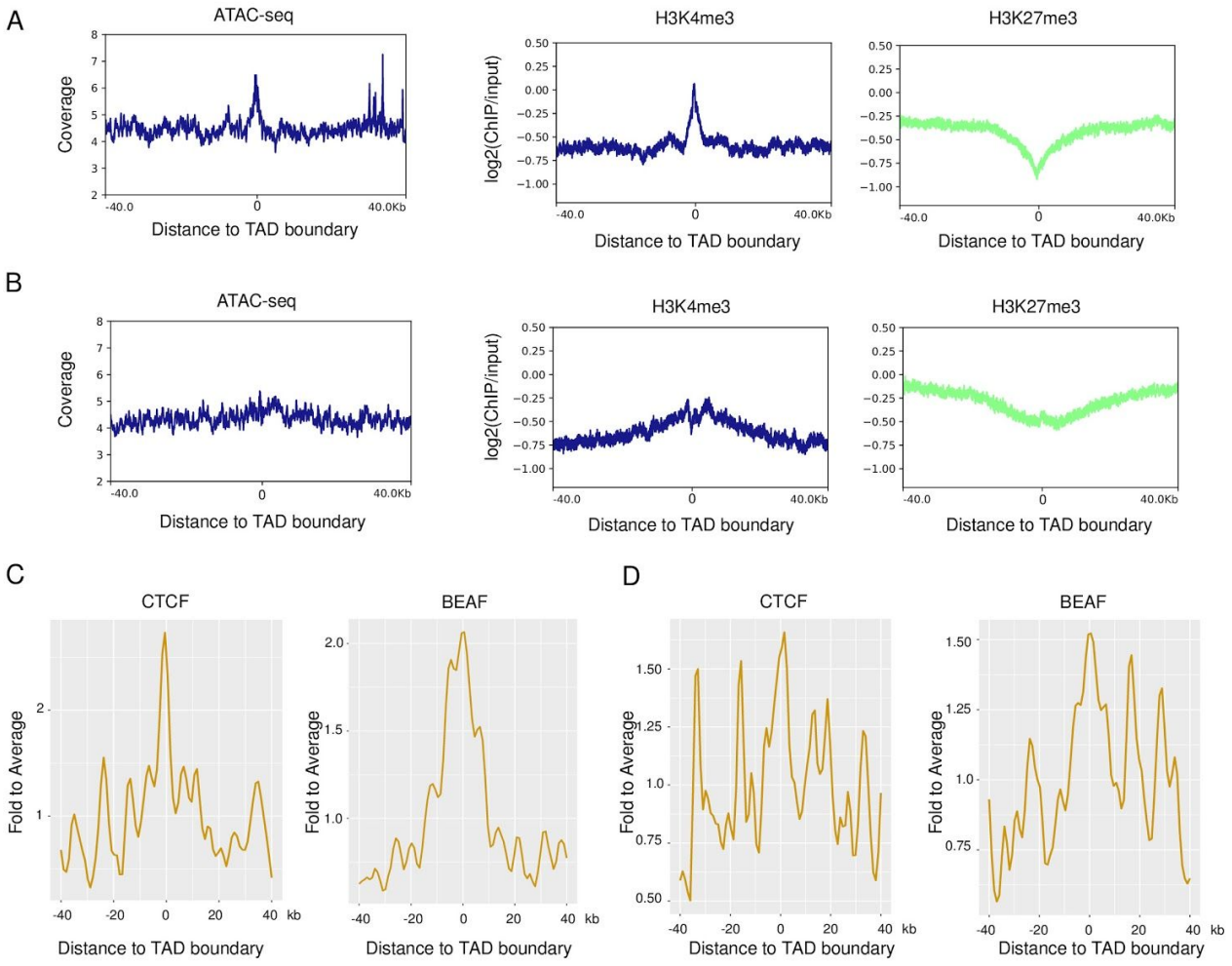
Supplemental Fig. S1: A large pericentromeric inversion (9.7Mb) on the X chromosome of *D. pseudoobscura* between the current assembly and the previous assemblies of this species. (A) Whole genome alignment plot between the current assembly and the UCB assembly (Mahajan et al. 2018)**.** (B) Whole genome alignment plot between the current assembly and the FlyBase Dpse_4.0 assembly. (C) Hi-C contact map verifying the highly continuous contigs and supported the scaffolding of the three contigs of the X chromosome. (D) Local Hi-C contact signals further supported the scaffolding.



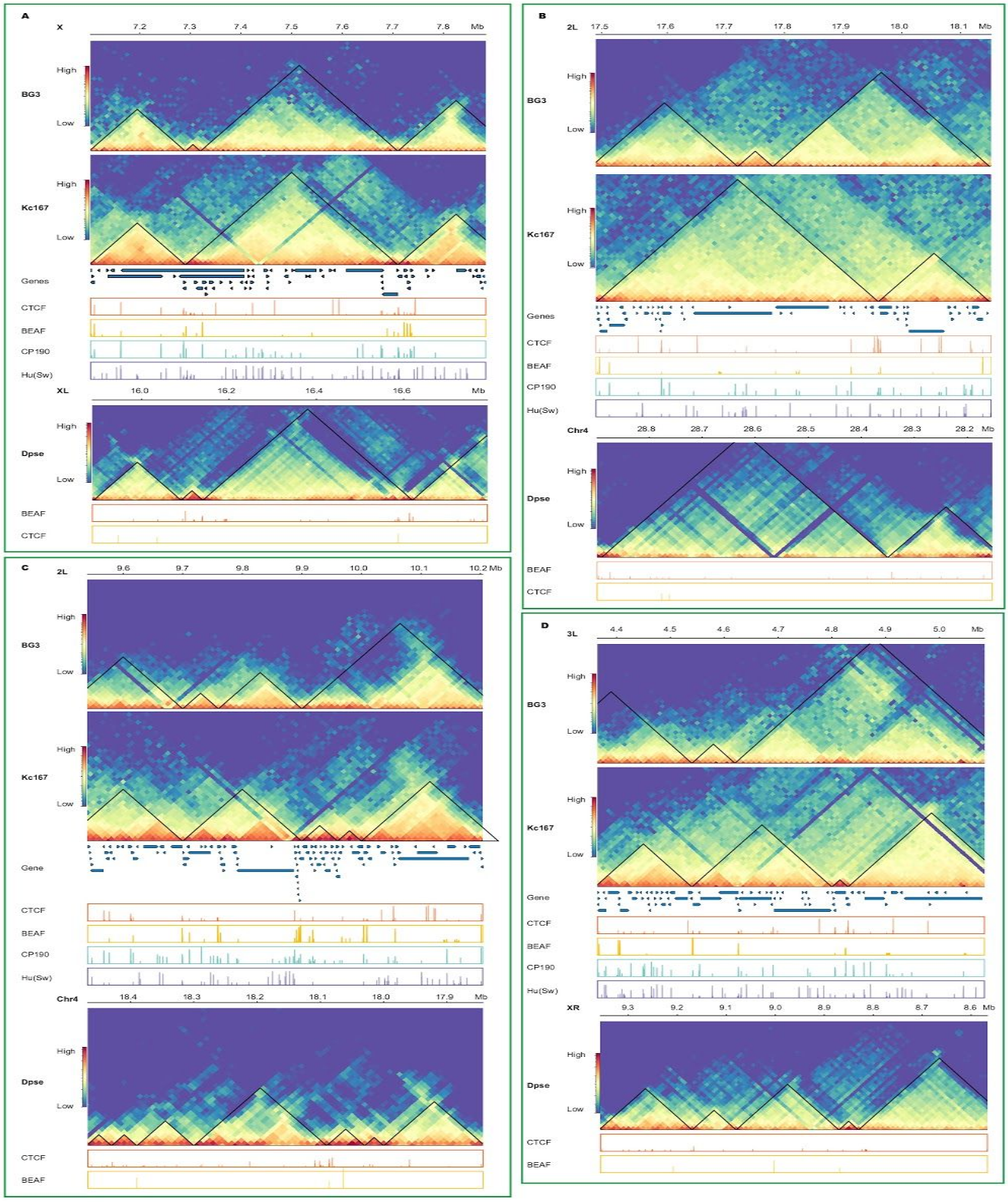
Supplemental Fig. S2: Comparative genomic features between *D. pseudoobscura* and *D. miranda*. Although extensive genome arrangements have occurred between these two species, they still showed similar distribution of genomic features (i.e. repeat and gene content). LTR retrotransposons were accumulated at the pericentromeric regions, which were poorly assembled in previous assemblies. Interestingly, the two species shared a common centromeric repeat (21bp), while each also acquired a species-specific centromeric repeat.



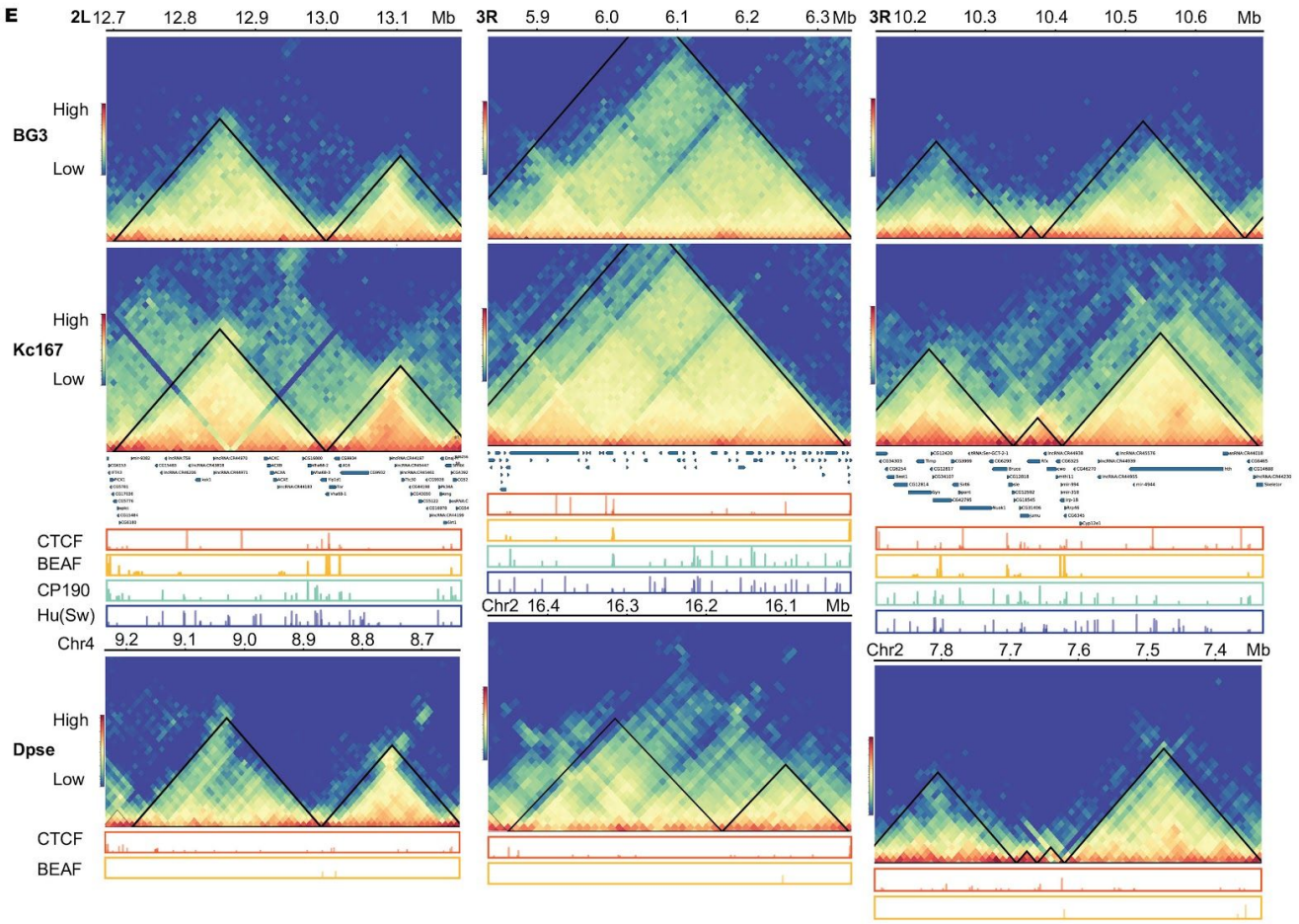
Supplemental Fig. S3: Comparison of TAD annotation using Armatus, Arrowhead in the Juicer package and HiCExplorer with different Hi-C contact map resolutions. An example region on chromosome 2 from 8,000,000 - 10, 000,000 Mb was shown.



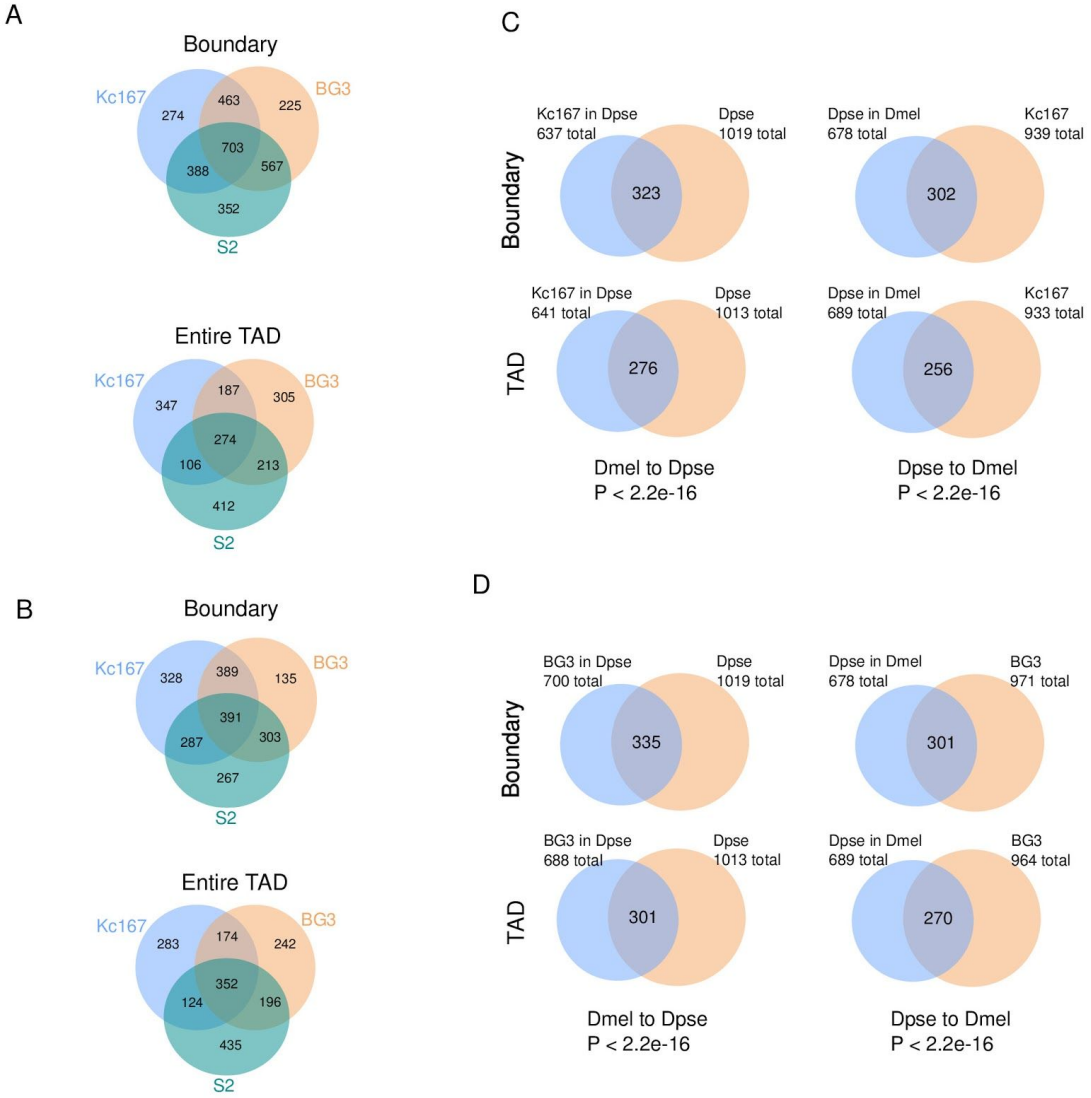
Supplemental Fig. S4: Enrichment analysis of two insulator proteins (BEAF-32 and CTCF), two epigenetic marks (H3K4me3 and H3K27me3) and the ATAC-seq for TAD boundaries identified by different methods. A,C for TAD boundaries identified by Aramtus; B,D for TAD boundaries identified by Arrowhead in the Juicer package.



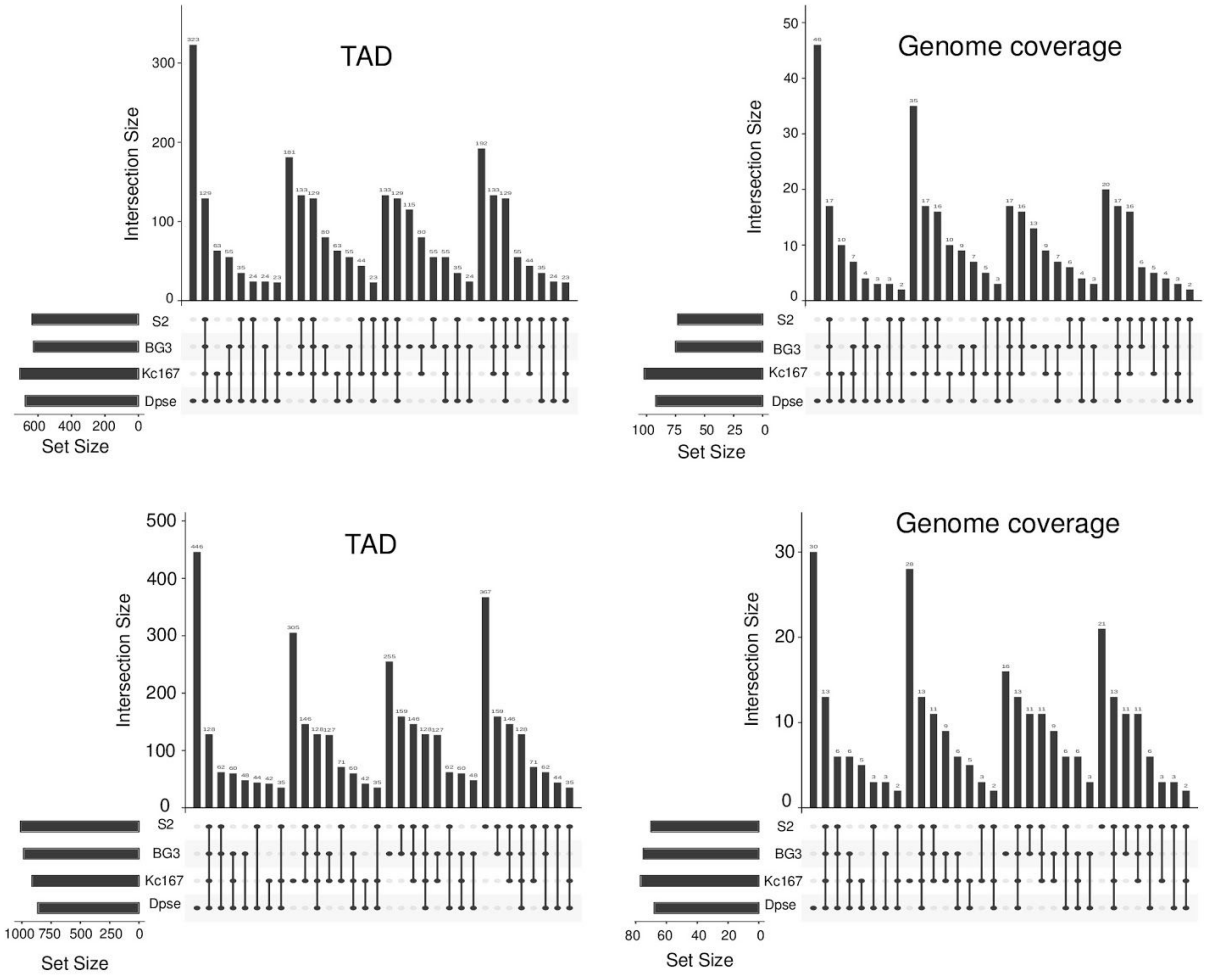
Continue on next page



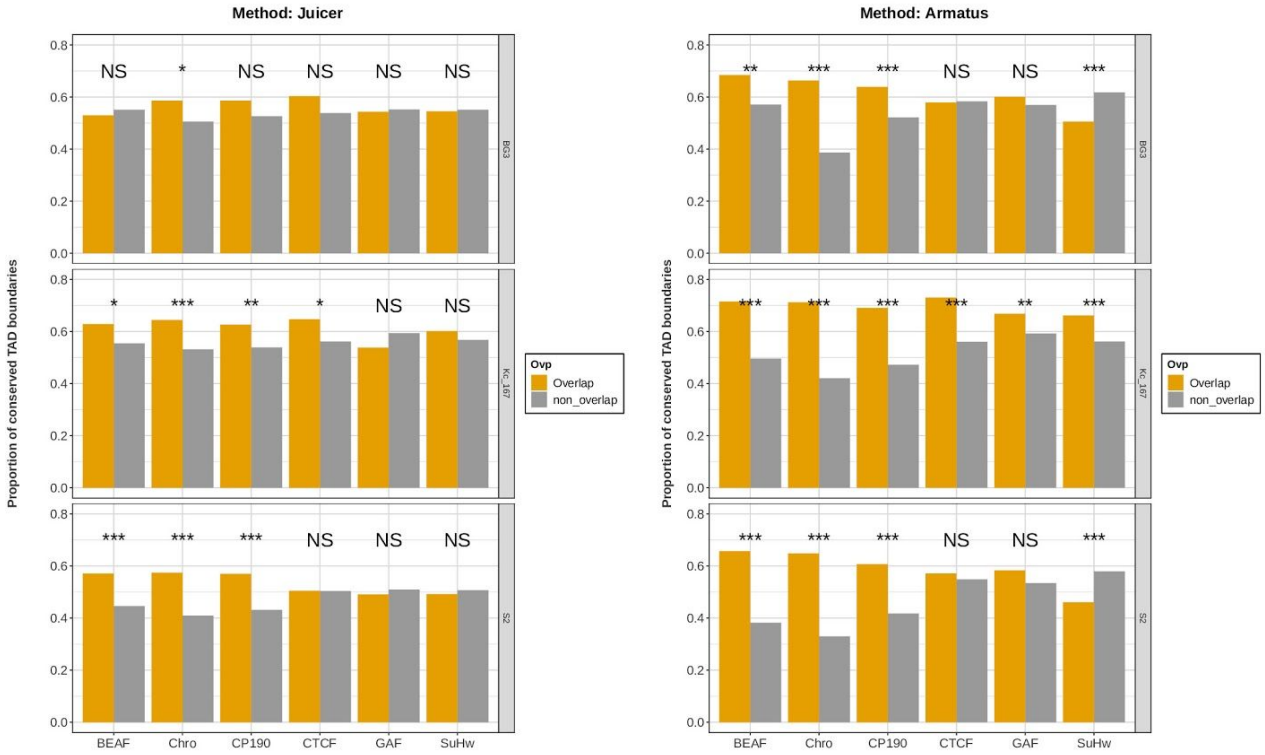
Supplemental Fig. S5: Examples of syntenic genomic regions between *D. melanogaster* and *D. pseudoobscura* remain conserved TAD structures.



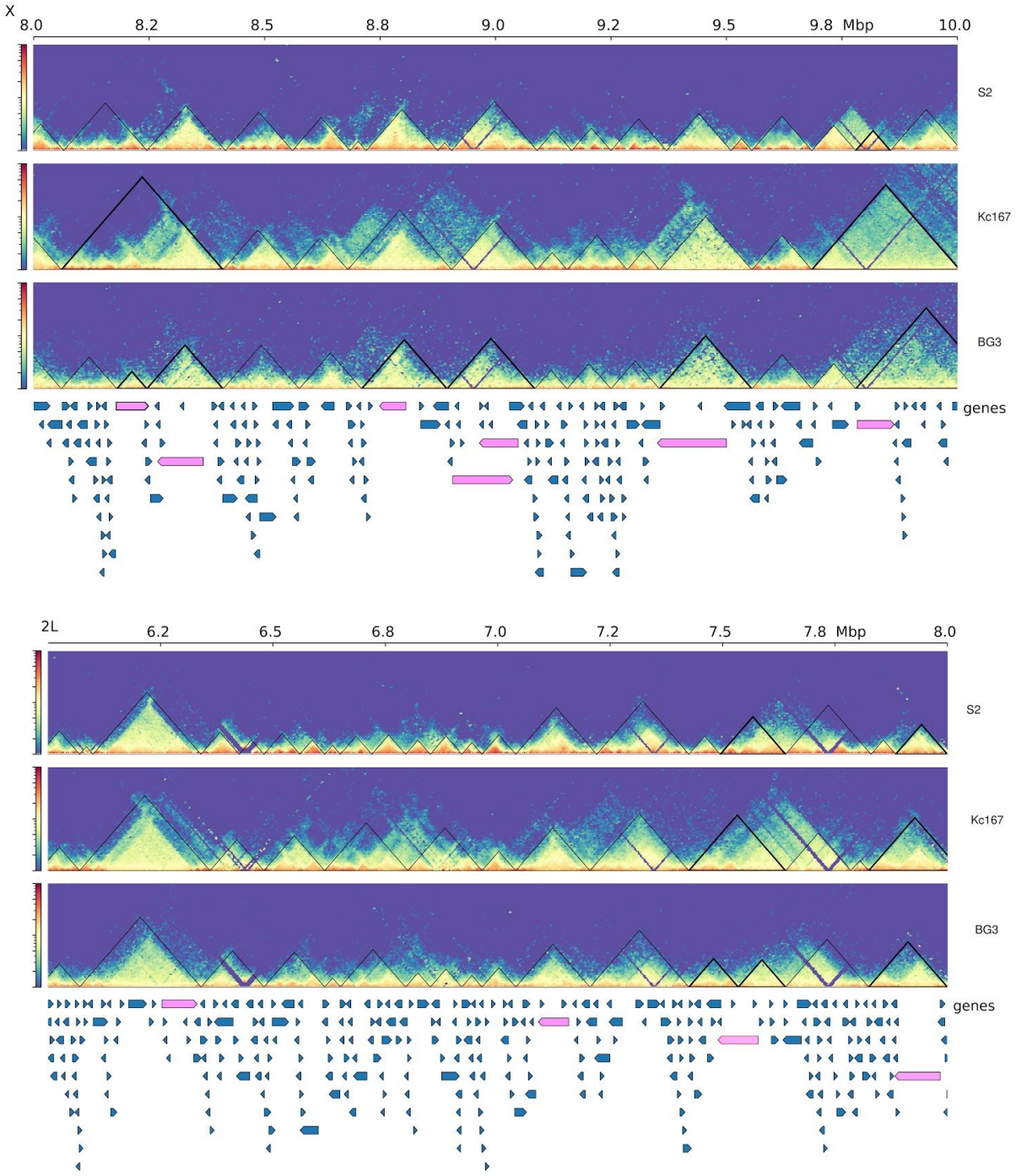
Supplemental Fig. S6: Conservation of TAD structures among cell lines in *D. melanogaster* and between *D. melanogaster* and *D. pseudoobscura*. (A) TAD identified by Armutas with comparison among three cell lines (Kc167, BG3 and S2) in boundaries (top) and TAD body (bottom). (B) TAD identified by Arrowhead. (C) Comparison of TAD structures between *D. melanogaster* Kc167 and *D. pseudoobscura* (full body) in boundary (top) and TAD body (bottom). Coordinates are liftovered between each other using the UCSC liftOver tool. (D) Comparison of TAD structures between *D. melanogaster* BG3 and *D. pseudoobscura* (full body).



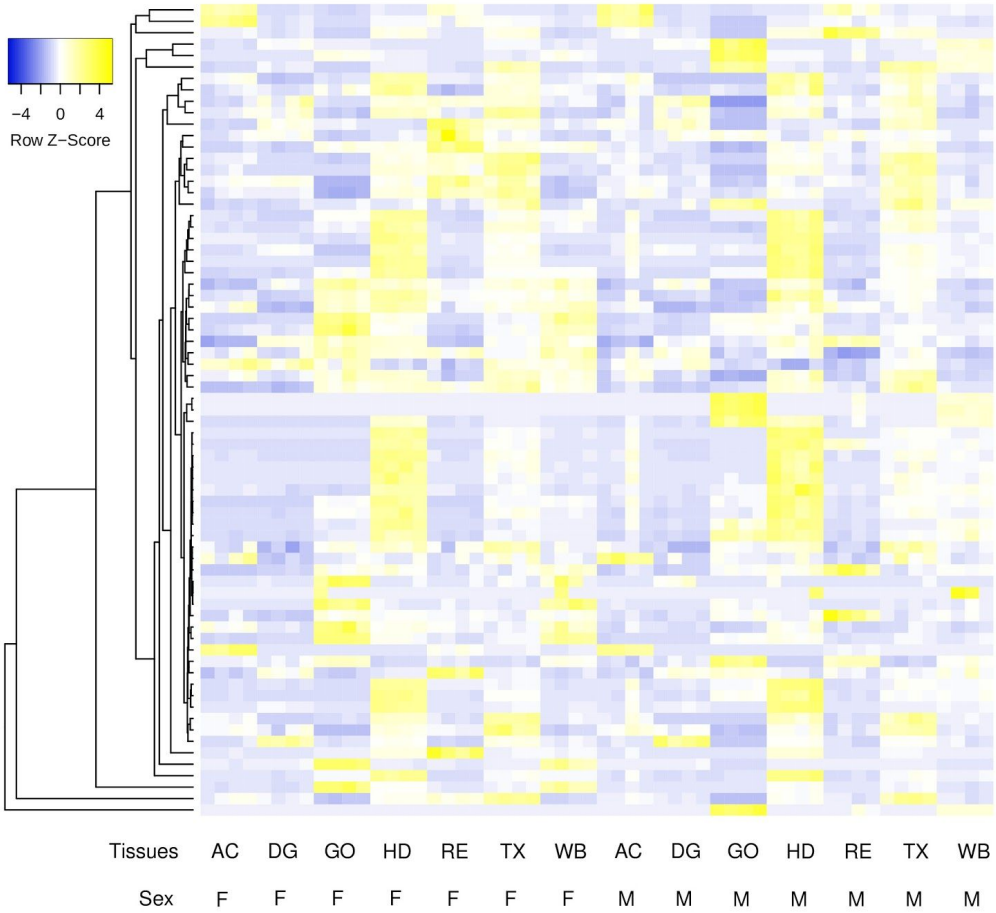
Supplemental Fig. S7: Upset plot showing conservation of TAD structures among three cell lines in *D. melanogaster* and *D. pseudoobscura*.Top are TAD identified by Arrowhead with comparison in terms of TAD body (left) and genome coverage (right). Below is for TAD identified by Aramtus.



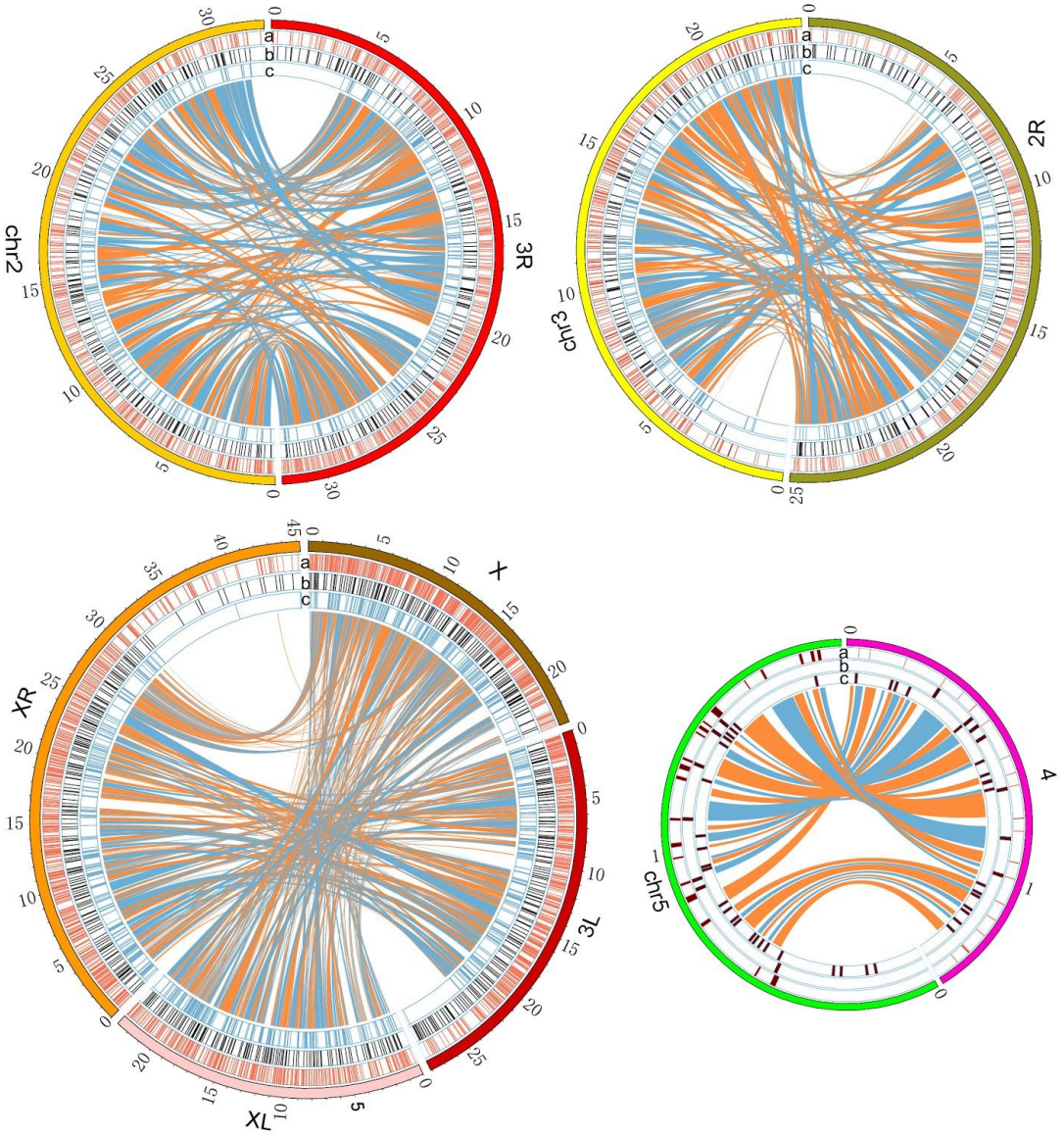
Supplemental Fig. S8: Comparison of conservation degree for TAD boundaries demarcated/overlapped with different insulator proteins (BEAF, Chro, CP190, CTCF, GAF (Trl) and Su(Hw)).The coordinates were liftovered from *D. melanogaster* to *D. pseudoobscura* and then assessed their overlap with the boundary coordinates in *D. pseudoobscura*. Left is analysis with TAD boundaries identified by Arrowhead and left is analysis with TAD boundaries identified by Armatus.



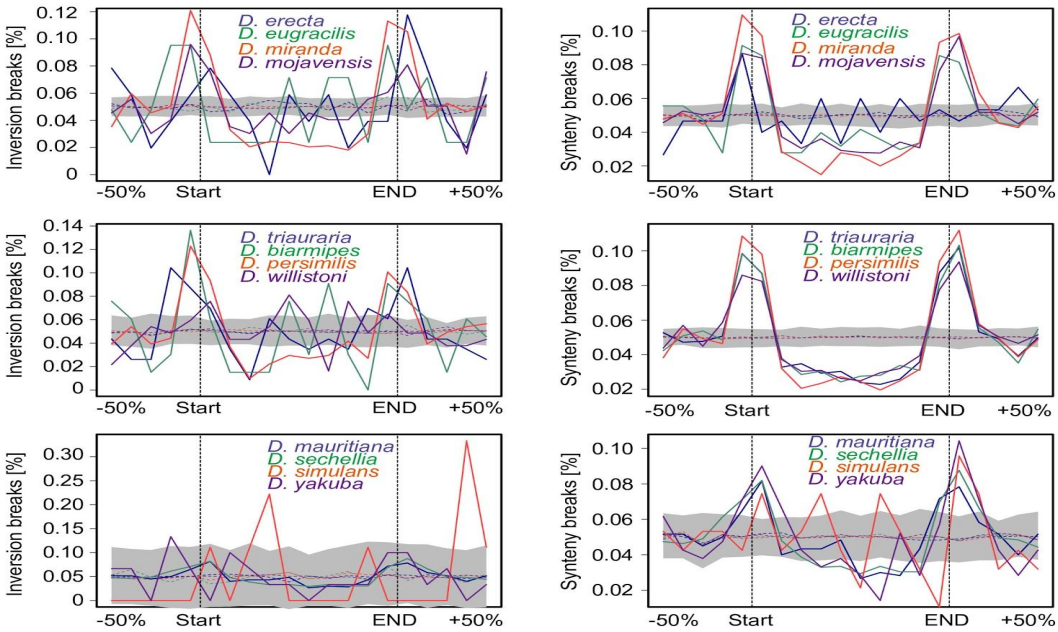
Supplemental Fig. S9: Examples of distribution of physical overlap between long genes and annotated TADs spanning them in *D. melanogaster*. Long genes are shown in pink color.



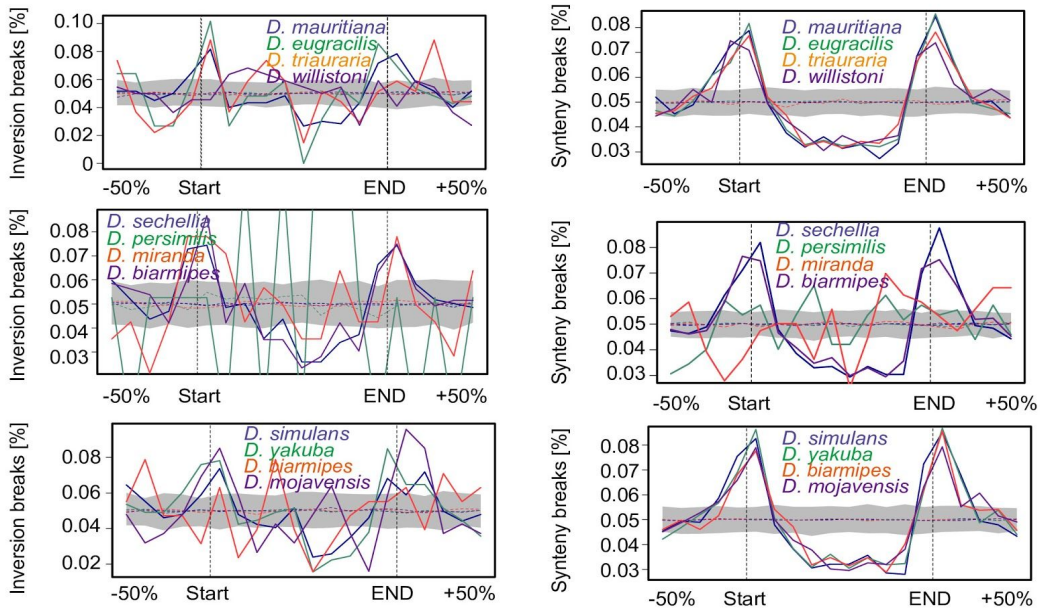
Supplemental Fig. S10 Expression profile of 71 long genes that overlap with complete TADs in both sex of *D. pseudoobscura* across 7 tissues. AC, abdomen without digestive or reproductive system; DG, digestive plus excretory system; GO, gonad; HD, head; RE, reproductive system without gonad; TX, thorax without digestive system; WB, whole body. F, female; M, male.



Supplemental Fig. S11: Synteny map between *D. melanogaster* and *D. pseudoobscura* (supplement for Figure 5A). Tracks a: TAD boundaries annotated by HiCExperor at restriction fragment resolution; b: 10 kb resolution; c: synteny breakpoints.

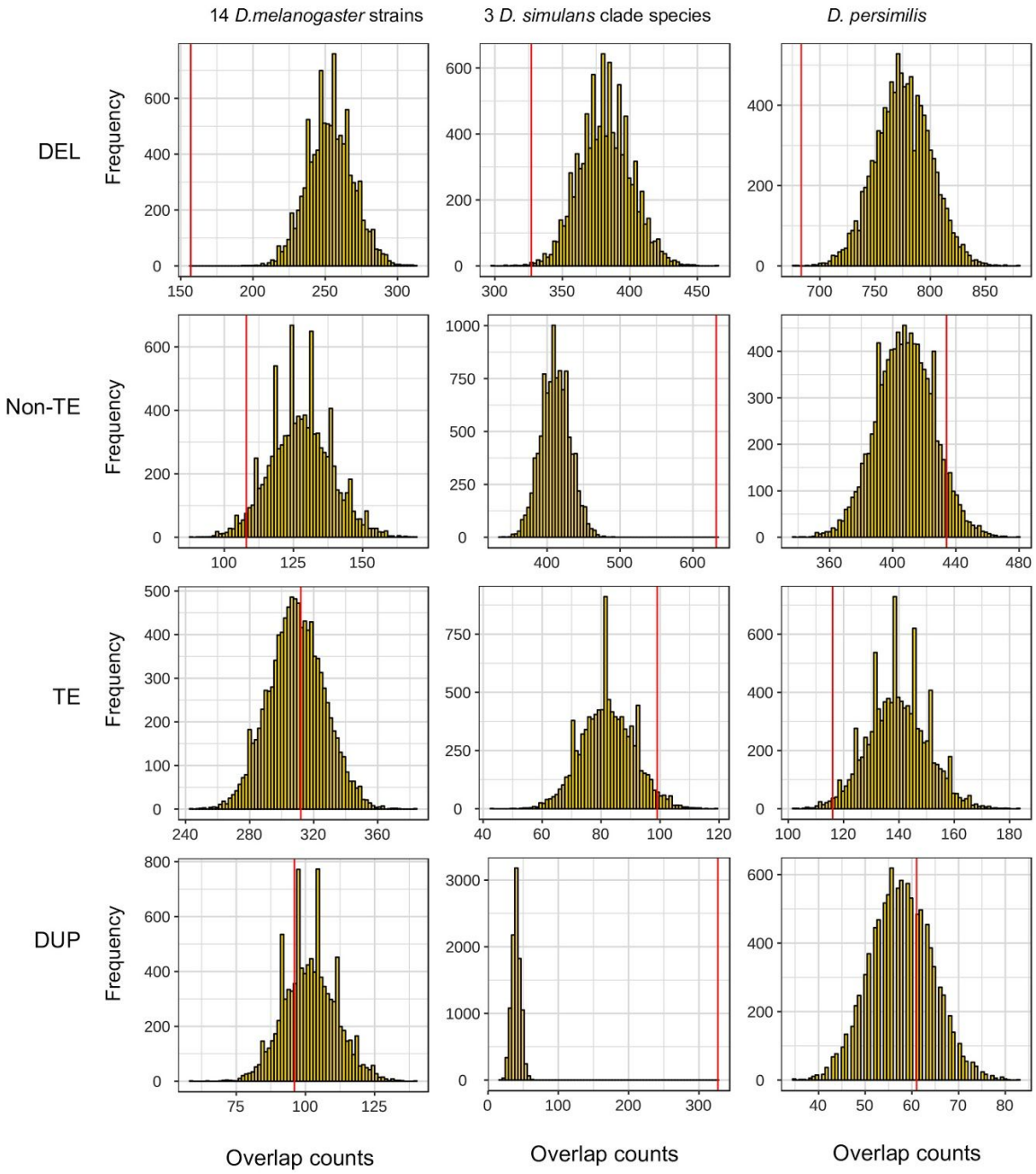


***D. melanogaster* TAD**

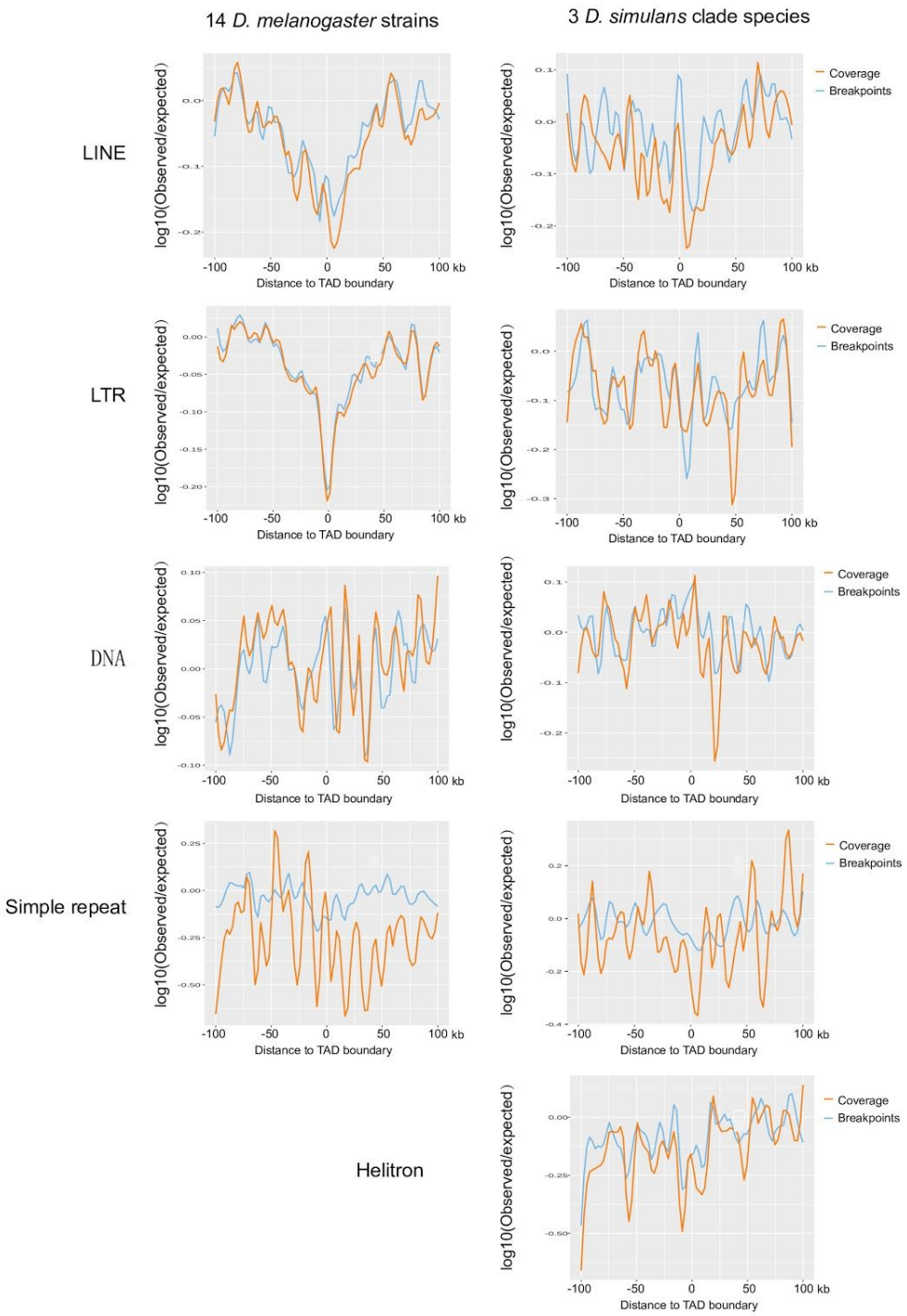


***D. pseudoobscura* TAD**

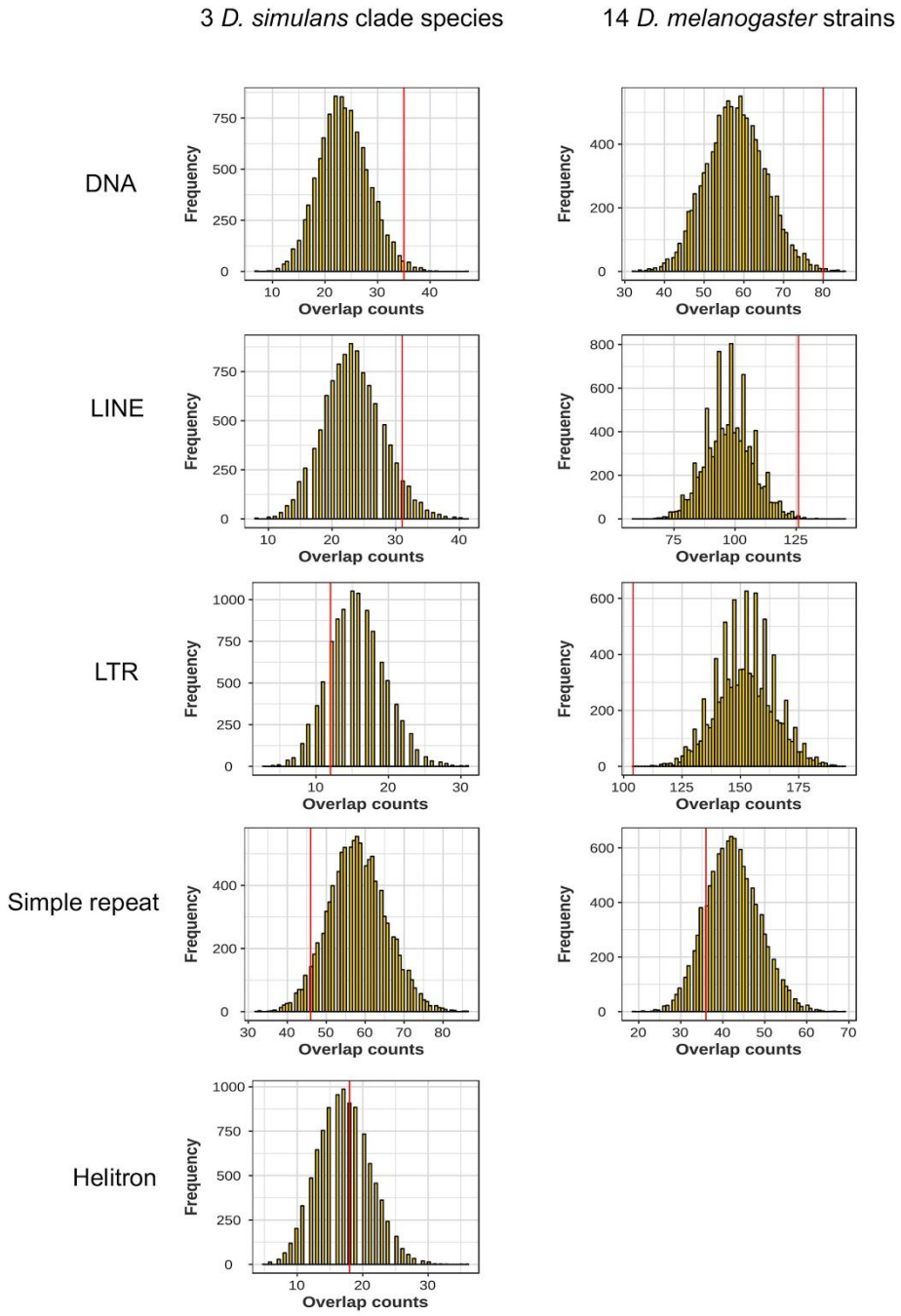
Supplemental Fig. S12: Distribution of evolutionary synteny/inversion breakpoints around TAD boundaries (supplement to Figure 5C,D). Top for comparison between *D. melanogaster* and 12 divergent *Drosophila* species. Below is for *D. pseudoobscura*.



Supplemental Fig. S13: Permutation test for evaluating the significant enrichment or depletion of four major SV types around the TAD boundaries in three SVs data sets representing different evolutionary time scales.We simulate 10,000 sets of random regions that had the same size and chromosomal distribution as the TAD boundaries and plot the frequency of the number of SVs that overlap with the random data set. The red line represents the real observed value.



Supplemental Fig. S14: Plot of the frequency of different TE types around the TAD boundaries (100 kb each for downstream and upstream).



Supplemental Fig. S15: Permutation test for evaluating the significant enrichment or depletion of different TE types around the TAD boundaries in two data sets representing different evolutionary time scales. We simulate 10,000 sets of random regions that had the same size and chromosomal distribution as the TAD boundaries and plot the frequency of the number of SVs that overlap with the random data set. The red line represents the real observed value.

Supplemental References

Benson, G. 1999. “Tandem Repeats Finder: A Program to Analyze DNA Sequences.” *Nucleic Acids Research* 27 (2): 573–80.

Campbell, Michael S., Carson Holt, Barry Moore, and Mark Yandell. 2014. “Genome Annotation and Curation Using MAKER and MAKER-P.” *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 48 (1): 188.

Chathoth, Keerthi T., and Nicolae Radu Zabet. 2019. “Chromatin Architecture Reorganization during Neuronal Cell Differentiation in Drosophila Genome.” *Genome Research* 29 (4): 613–25.

Ellinghaus, David, Stefan Kurtz, and Ute Willhoeft. 2008. “LTRharvest, an Efficient and Flexible Software for de Novo Detection of LTR Retrotransposons.” *BMC Bioinformatics* 9 (January): 18.

Han, Yujun, and Susan R. Wessler. 2010. “MITE-Hunter: A Program for Discovering Miniature Inverted-Repeat Transposable Elements from Genomic Sequences.” *Nucleic Acids Research* 38 (22): e199.

He, W., S. Zhao, X. Liu, S. Dong, J. Lv, D. Liu, J. Wang, and Z. Meng. 2013. “ReSeqTools: An Integrated Toolkit for Large-Scale next-Generation Sequencing Based Resequencing Analysis.” *Genetics and Molecular Research: GMR* 12 (4): 6275–83.

Kim, Daehwan, B. Langmead, and S. L. Salzberg. 2015. “hisat2.” *Nature Methods*. https://ccb.jhu.edu/software/hisat2/data/HISAT2-first_release-Sept_8_2015.pdf.

Mahajan, Shivani, Kevin H-C Wei, Matthew J. Nalley, Lauren Gibilisco, and Doris Bachtrog. 2018. “De Novo Assembly of a Young Drosophila Y Chromosome Using Single-Molecule Sequencing and Chromatin Conformation Capture.” *PLoS Biology* 16 (7): e2006348.

Ni, Xiaochun, Yong E. Zhang, Nicolas Nègre, Sidi Chen, Manyuan Long, and Kevin P. White. 2012. “Adaptive Evolution and the Birth of CTCF Binding Sites in the Drosophila Genome.” *PLoS Biology* 10 (11): e1001420.

Nozawa, Masafumi, Kanako Onizuka, Mai Fujimi, Kazuho Ikeo, and Takashi Gojobori. 2016. “Accelerated Pseudogenization on the Neo-X Chromosome in Drosophila Miranda.” *Nature Communications* 7 (November): 13659.

Ou, Shujun, and Ning Jiang. 2018. “LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons.” *Plant Physiology* 176 (2): 1410–22.

Pertea, Mihaela, Geo M. Pertea, Corina M. Antonescu, Tsung-Cheng Chang, Joshua T. Mendell, and Steven L. Salzberg. 2015. “StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads.” *Nature Biotechnology* 33 (3): 290–95.

Riddle, Nicole C., Aki Minoda, Peter V. Kharchenko, Artyom A. Alekseyenko, Yuri B. Schwartz, Michael Y. Tolstorukov, Andrey A. Gorchakov, et al. 2011. “Plasticity in Patterns of Histone Modifications and Chromosomal Proteins in Drosophila Heterochromatin.” *Genome Research* 21 (2): 147–63.

Schwartz, Yuri B., Daniela Linder-Basso, Peter V. Kharchenko, Michael Y. Tolstorukov, Maria Kim, Hua-Bing Li, Andrey A. Gorchakov, et al. 2012. “Nature and Function of Insulator Protein Binding Sites in the Drosophila Genome.” *Genome Research* 22 (11): 2188–98.

Smit, A. F. A., R. Hubley, and P. Green. 2015. “RepeatModeler Open-1.0. 2008--2015.” *Seattle, USA: Institute for Systems Biology. Available from: Httpwww. Repeatmasker. Org, Last Accessed May* 1: 2018.

Wang, Qi, Qiu Sun, Daniel M. Czajkowsky, and Zhifeng Shao. 2018. “Sub-Kb Hi-C in D. Melanogaster Reveals Conserved Characteristics of TADs between Insect and Mammalian Cells.” *Nature Communications* 9 (1): 188.

Wood, Ashley M., Kevin Van Bortle, Edward Ramos, Naomi Takenaka, Margaret Rohrbaugh, Brian C.

Jones, Keith C. Jones, and Victor G. Corces. 2011. “Regulation of Chromatin Organization and Inducible Gene Expression by a Drosophila Insulator.” *Molecular Cell* 44 (1): 29–38.

Xu, Zhao, and Hao Wang. 2007. “LTR_FINDER: An Efficient Tool for the Prediction of Full-Length LTR Retrotransposons.” *Nucleic Acids Research* 35 (Web Server issue): W265–68.

Yang, Haiwang, Maria Jaime, Maxi Polihronakis, Kelvin Kanegawa, Therese Markow, Kenneth Kaneshiro, and Brian Oliver. 2018. “Re-Annotation of Eight Drosophila Genomes.” *Life Science Alliance* 1 (6): e201800156.

Yang, Jingping, Edward Ramos, and Victor G. Corces. 2012. “The BEAF-32 Insulator Coordinates Genome Organization and Function during the Evolution of Drosophila Species.” *Genome Research* 22 (11): 2199–2207.