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 rearrangement 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, 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=163Mb) 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 163Mb assembly being represented by 6 contigs that are 9.7 Mb or larger (i.e. N90 = 9.7 Mb and N50= 30.7 Mb). 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)(see Methods; Supplemental Table S2,3). 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 (Supplemental Fig. S1) and identified a large (~9.7 Mb) pericentromeric inversion (Supplemental Fig. S2) 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 30.3% of the *D. pseudoobscura* genome is annotated as repetitive sequences, including 11.8% in LTR retrotransposons, 5.4% in LINE and 2% in the DNA-type TE (Supplemental Table S6). 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 (see Methods). 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 sequenced, of which half are retained after filtering and are valid for constructing the Hi-C contact map (see Methods; 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 highly 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 enriched for the active chromatin marker, H3K4me3, but depleted for the repressive chromatin marker, 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.

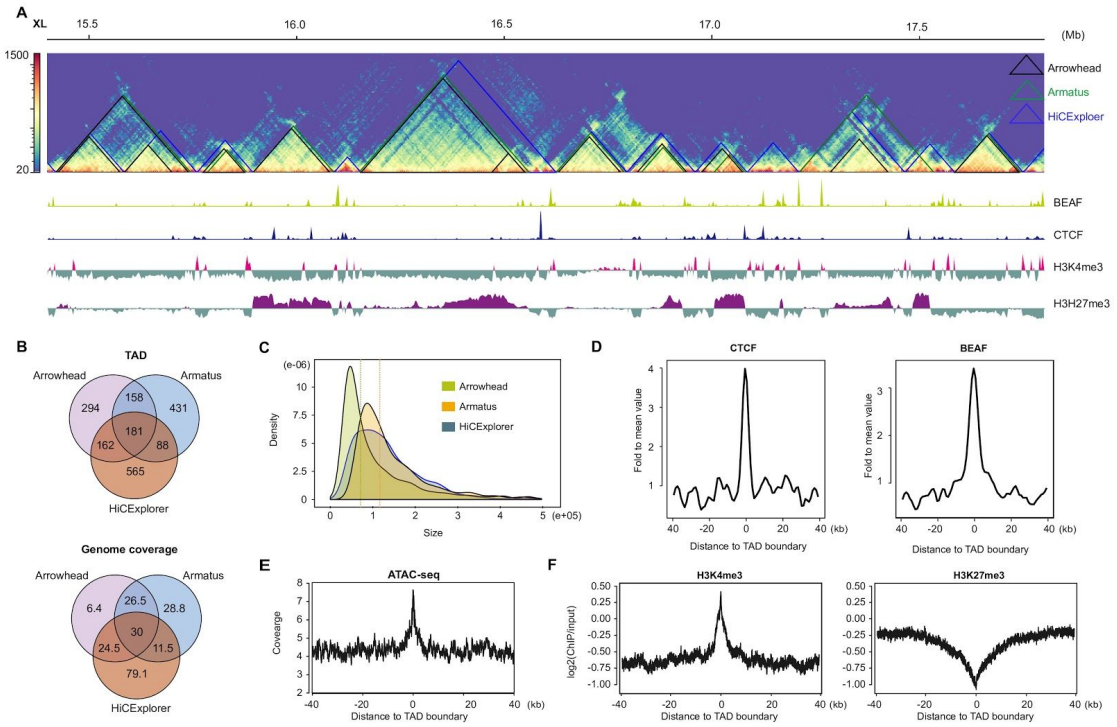

Figure 1: In situ Hi-C map and topologically associating domains (TADs) from the *D. pseudoobscura* adult full body. (**A**) Top: Hi-C contact map from 14.8 Mb to 17.8 Mb on X chromosome with 5kb bin size and TADs identified by three methods, Arrowhead, Armatus and HiCExplore ; Bottom: ChIP-seq for two insulator proteins, BEAF-32 and CTCF, and two epigenetic markers, H3K4me3 and H3H27me3. (**B**) TAD annotation consistency among three methods in terms of number (top) and genome coverage (bottom). (**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 markers and at 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 and identified conserved syntenic regions. The resulting dataset 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 5 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 comparisons 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* WB TADs, which is 2.4 fold enrichment relative to random expectation (~18.5 %; Fisher's exact test p-value < 2.2e-16; Supplemental Table S11). Correspondingly, 48% (321/685) of *D. pseudoobscura* TAD boundaries (Fig. 2D) are also detected in the S2 cell line, compared to random expectation of 20% (Fisher's exact test p-value < 2.2e-16; Supplemental Table S11). Similar trends are also shown in the analysis with KC167 and BG3 cell lines and with TADs identified in different tools (Supplemental Fig. S7). Our results show that at least 40% of the TADs spanning 30% of the genome (Fig. 2E) 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.

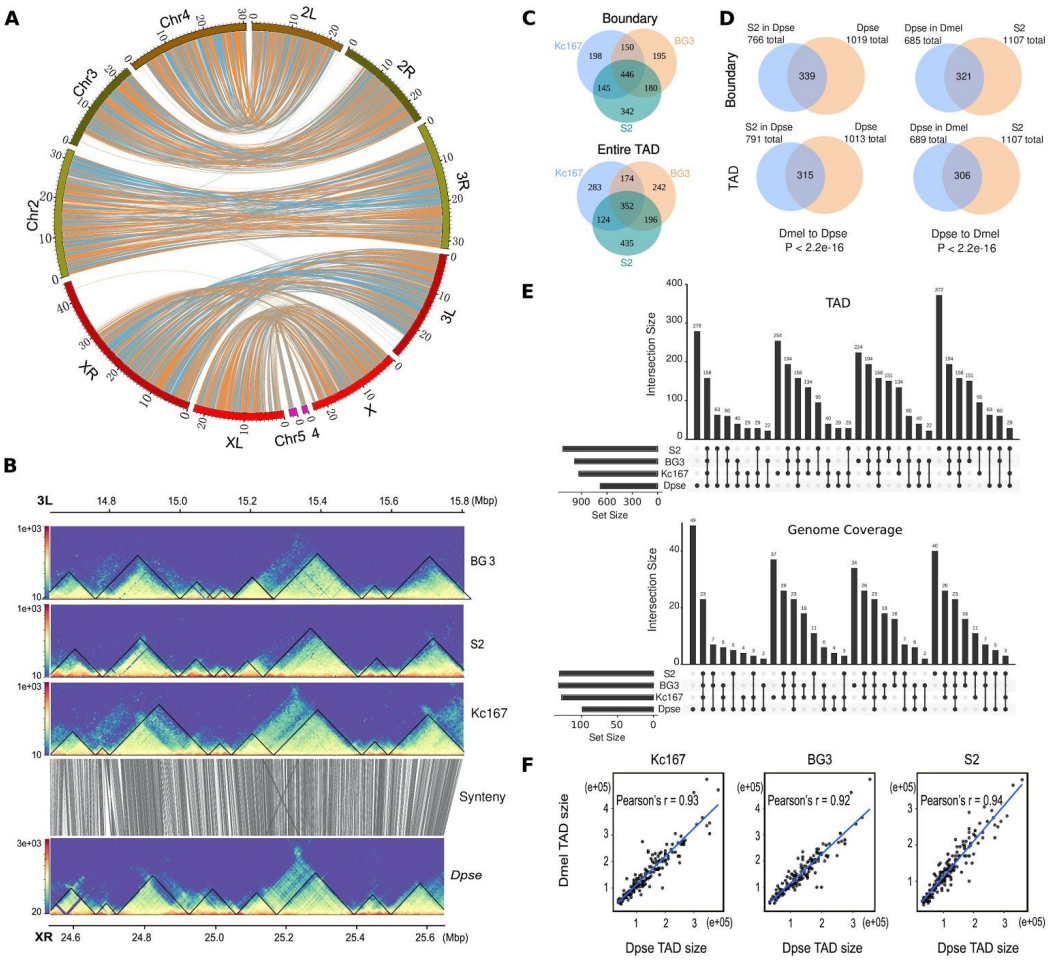

Figure 2: Evolutionary conservation of TAD structure between *D. melanogaster* and *D. pseudoobscura.* (**A**). Genome-wide synteny map between *D. melanogaster* and *D. pseudoobscura* constructed using 985 syntenic blocks which are larger than 10kb. (**B**) Hi-C contact maps and TAD structures from 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; Bottom: 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 S12). We found that boundaries (annotated with HiCExplorer) that overlap with BEAF-32, CP190 and Chromator are substantially more conserved than those that do not (Fig. 3A), while boundaries overlap with CTCF, Su(Hw) and Trl don't show this pattern (Fig. 3A). Analogous results are obtained from TAD boundaries annotated with two other methods (Supplemental Fig. S8). For *D. pseudoobscura*, we obtained ChiP-seq data for BEAF-32 and CTCF and found that TAD boundaries overlap with BEAF-32 are more conserved than those that are not, but not for CTCF (Supplemental Fig. S9). 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) (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 the most conserved TAD boundaries are usually those adjacent to at least one active TAD (Fig. 3B), which were also classified as the strongest TAD boundaries (Ramírez et al. 2018). While, TAD boundaries between

inactive-inactive TADs, inactive-PcG TADs or PcG-PcG TADs are less conserved than those between active TADs.

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 a total of 921 TAD boundaries that are shared in at least two cell types. Some boundaries appeared in only one cell type dataset: 198 were Kc167 specific, 195 were BG3 specific, and 342 were S2 specific(Supplemental Table S13). We found that 58% (391/672) of boundaries shared by two or more cell types are also TAD boundaries in *D. pseudoobscura* (only 23% is expected at random), 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 or development stages are more likely to be conserved than those that are cell type specific. 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 random expectation), while only 40% (130/326) of weak boundaries are found to be conserved (Fig. 3B). Correspondingly, this pattern is also observed in the reciprocal comparison focusing on *D. melanogaster* TAD boundaries. This result suggests that stronger TAD boundaries should be more conserved in the course of evolution.

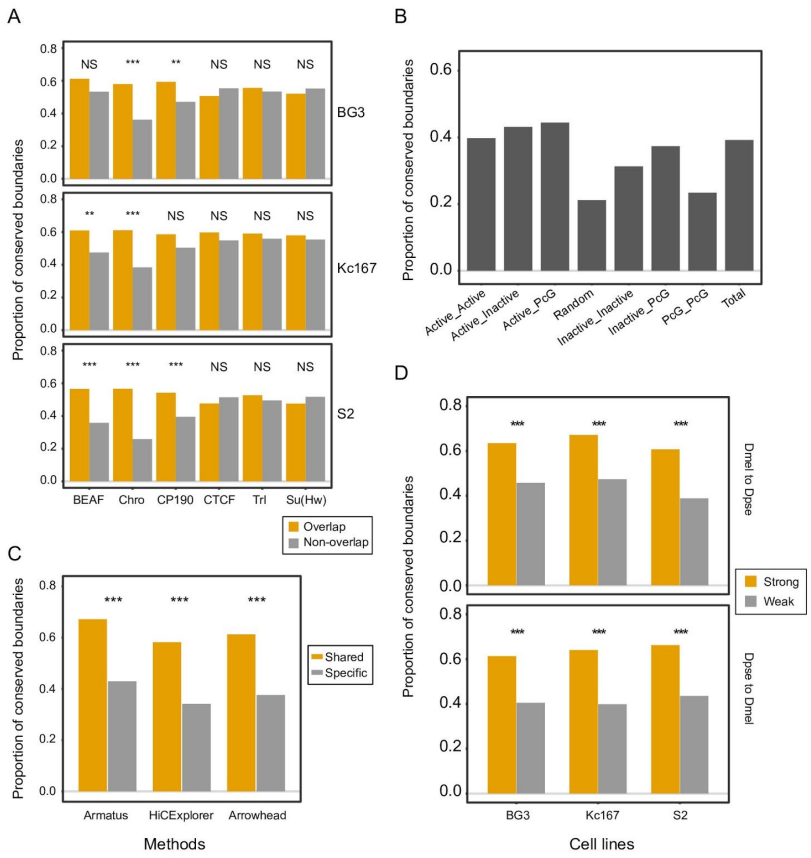

Figure 3: Conservation of TAD boundaries between *D. melanogaster* and *D. pseudoobscura*. (A) Conservation of TAD boundaries overlapped with architectural proteins. (B) Conservations of boundaries from TAD with different chromatin modifications. (C) Comparison between across cell types boundaries (share at least in two cell lines) and cell specific boundaries. (D) Comparison between strong and weak boundaries. Significance is determined by Fisher exact test (***P < 0.001; **<0.01; NS: no significance).

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 8 tissues from both males and females (Yang et al. 2018). Krefting et al. (Krefting et al. 2018) suggest 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: genes inside TADs (i.e. genes fall in the genomic regions where TAD bodies are annotated using Juicer); or 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-5, Fig. 4B). These results suggest that TAD arrangements are associated with variance in conservation of gene expression during *Drosophila* genome evolution.

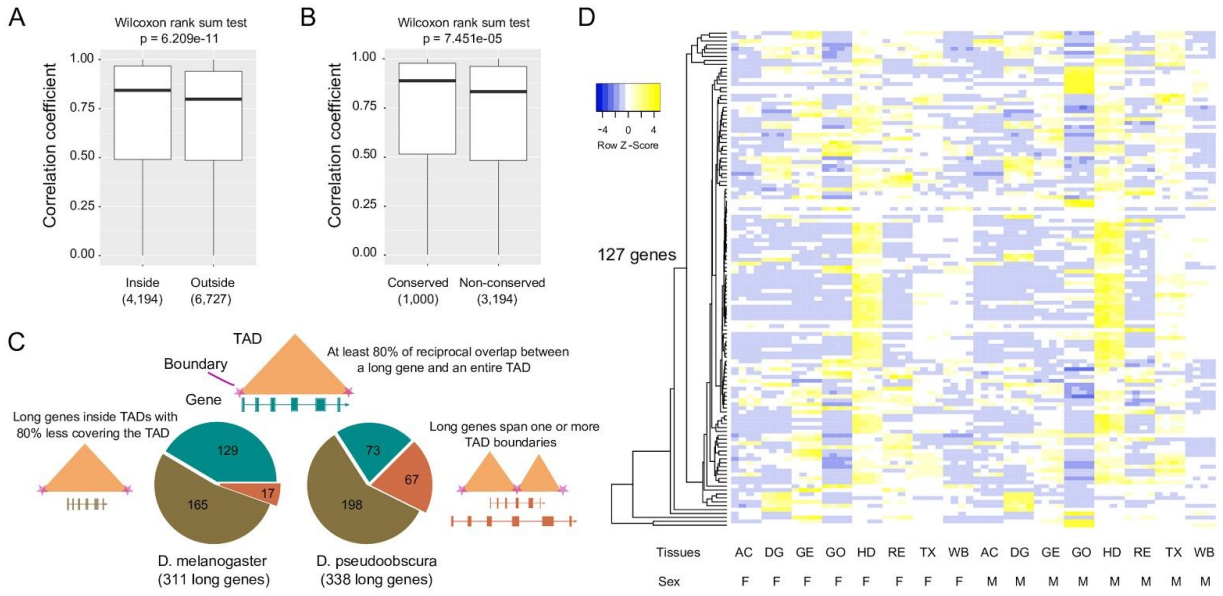

Figure 4: **The role of TADs in gene expression evolution and regulation of long genes.** (A) Expression correlation of one-to-one orthologs across 8 tissues and both sexes in *D. melanogaster*

and *D. pseudoobscura* for *D. melanogaster* genes within or outside the TADs. (B) Expression correlation of one-to-one orthologs across 8 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, $P < 10e\text{-}23$), 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 S14) 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, 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. Of these 73 genes, 64 have orthologs in *D. melanogaster* and 43 still have conserved TAD structure between these two species. 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). 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; see other chromosomes in Supplemental Fig. S10). 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(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 11), 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 11).

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.

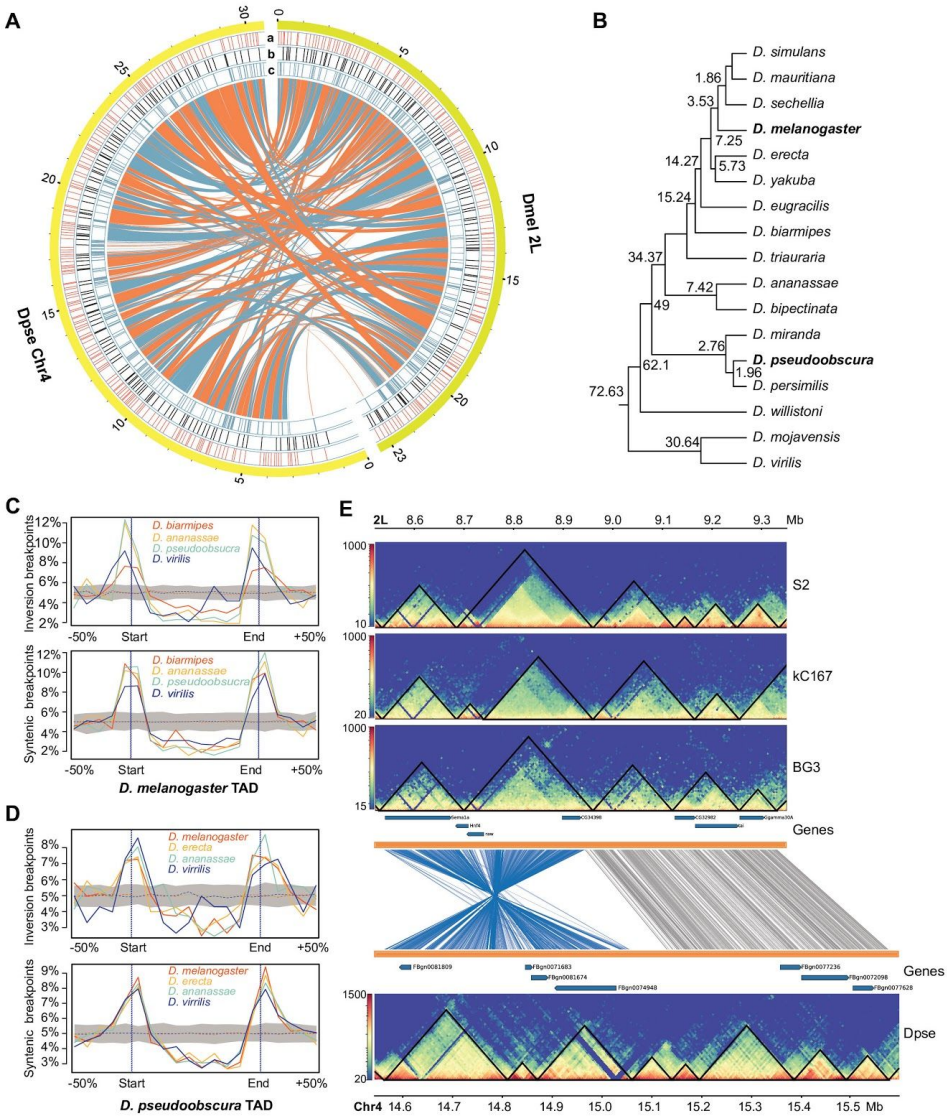

Figure 5: Evolutionary synteny breaks are enriched with TAD boundaries in D*rosophila* genomes. (A) Synteny map between *D. melanogaster* 2L and *D. pseudoobscura* Chr4. Other pairs of homologous chromosomes are shown in Supplementary Fig. S10. Tracks a: TAD boundaries annotated by HiCExperor at restriction 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. S11. (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. S11. (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); 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); 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). Insertions show a relatively complex pattern. Although the overall insertions are found to be depleted at the TAD boundaries, some special types of insertions are not. For example, among TE insertions, LTR and LINE families are strongly depleted at TAD boundaries, but not DNA-type TEs (Supplemental Fig S12). The same trend was also observed in large insertions from *D. simulans* clade species. It is worth noting that deletions and insertions from the *D. melanogaster* population are more depleted than those from *D. simulans* clade species.

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 a sample of *D. melanogaster* strains are not, suggesting. 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 in *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.

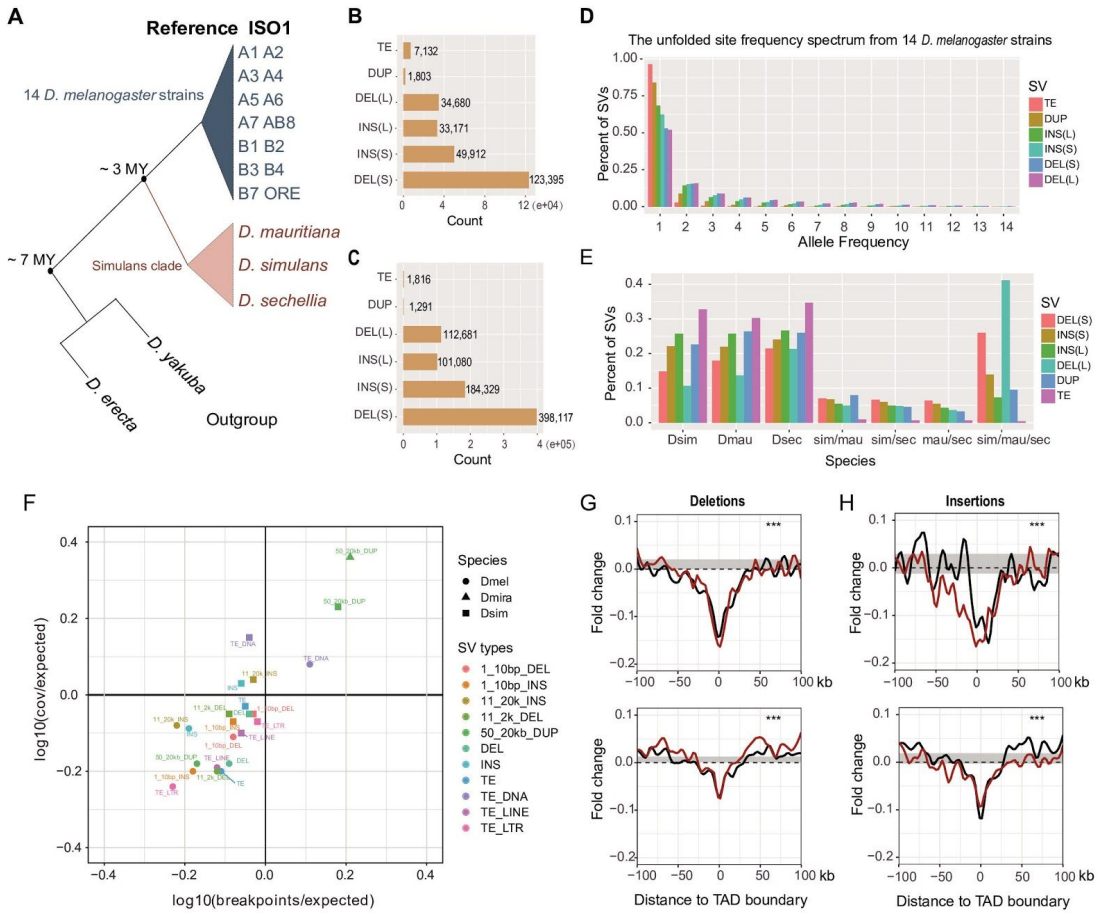

Figure 6: Structural variants are under purifying selection at *Drosophila* TAD boundaries. (A) Highly continuous genome assemblies from 14 *D. melanogaster* strains and three *D. simulans* clade

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 30% 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. 4A). 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 79 (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 ~283 X (G =160Mb) 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 S12.

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 three 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 simply plot the count of each type of SVs (i.e. deletions, Non-TE insertions, TE insertions and CNVs) in a 5kb sliding window for 100 kb upstream and downstream from the midpoint of TAD boundaries for all TAD boundaries. Count was normalized by dividing the mean count of that type of SVs in all windows. The count was then plotted in a heatmap. Second, 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. Third, 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.