The Complex Landscape of Structural Divergence Between the *Drosophila pseudoobscura* and *D. persimilis* Genomes

Javier Carpinteyro-Ponce and Carlos A. Machado *

Department of Biology, University of Maryland, College Park, MD, USA

*Corresponding author: E-mail: machado@umd.edu.

Accepted: March 07, 2024

Abstract

Structural genomic variants are key drivers of phenotypic evolution. They can span hundreds to millions of base pairs and can thus affect large numbers of genetic elements. Although structural variation is quite common within and between species, its characterization depends upon the quality of genome assemblies and the proportion of repetitive elements. Using new high-quality genome assemblies, we report a complex and previously hidden landscape of structural divergence between the genomes of *Drosophila persimilis* and *D. pseudoobscura,* two classic species in speciation research, and study the relationships among structural variants, transposable elements, and gene expression divergence. The new assemblies confirm the already known fixed inversion differences between these species. Consistent with previous studies showing higher levels of nucleotide divergence between fixed inversions relative to collinear regions of the genome, we also find a significant overrepresentation of INDELs inside the inversions. We find that transposable elements accumulate in regions with low levels of recombination, and spatial correlation analyses reveal a strong association between transposable elements and structural variants. We also report a strong association between differentially expressed (DE) genes and structural variants and an overrepresentation of DE genes inside the fixed chromosomal inversions that separate this species pair. Interestingly, species-specific structural variants are overrepresented in DE genes involved in neural development, spermatogenesis, and oocyte-to-embryo transition. Overall, our results highlight the association of transposable elements with structural variants and their importance in driving evolutionary divergence.

Key words: Drosophila, species divergence, structural variant, transposable element.

Significance

A full understanding of genomic divergence between species requires a detailed characterization of both nucleotide and structural variation (i.e. indels, inversions, and translocations), which can only be achieved by comparing high-quality genome assemblies. Here, we conduct a comprehensive characterization of patterns of structural divergence between the genomes of *Drosophila pseudoobscura* and *D. persimilis,* a pair of closely related model fly species for the study of the genetics of speciation, and find a complex and previously unexplored pattern of genomic divergence that underlies the importance of Transposable Elements (TEs) in the generation of genomic and gene expression divergence. Our results highlight the association of TEs with structural variation and their importance in driving evolutionary divergence between closely related species.

Introduction

The rapid development of sequencing technologies has revolutionized the field of evolutionary genomics. With the recent emergence of long-read sequencing it is now possible to generate highly contiguous de-novo genome assemblies with fewer computational resources (Chakraborty et al. 2016; Jain et al. 2018; Wenger et al. 2019; Hon et al. 2020; Logsdon et al. 2020; Shafin et al. 2020; Nurk et al. 2022). Improvements to sequencing technologies and scaffolding methods, such as the PacBio HiFi and Hi-C methods, are also enabling new approaches to generating high-quality genome assemblies using even fewer computational resources (Hon et al. 2020). The availability of high-quality genomes allows the characterization of regions harboring a high proportion of transposable elements (TEs), which, given their repetitive nature, often present major challenges during the assembly process (O'Neill et al. 2020). High-quality genome assemblies have also revolutionized the identification and analysis of structural variants (SVs) such as inversions, duplications, insertions, and deletions. Improvements in genome assembly, therefore, have increased our understanding of how structural variation contributes to phenotypic differences between species (Chakraborty et al. 2018; Kronenberg et al. 2018; Wellenreuther et al. 2019; Logsdon et al. 2020; O'Neill et al. 2020; Weissensteiner et al. 2020; Nurk et al. 2022).

SVs can originate through a variety of DNA repair mechanisms, errors during meiotic recombination, and the transposition activity of mobile elements (Hastings et al. 2009; Weckselblatt and Rude 2015; Scully et al. 2019). An association of INDELs with TEs is inevitable given that recent transposition events represent recent insertions. Evidence from structural variation studies in *Drosophila* species has suggested a significant association between TEs and the genesis of large SVs such as inversions or tandem duplications (Richards et al. 2005; Bracewell et al. 2019). Furthermore, the effects of both SVs and TE activity on gene expression, through the alteration of gene structure, modification of associated regulatory regions, or epigenetic silencing of neighboring regions have been studied in several species (Zichner et al. 2013; Chiang et al. 2017; Kronenberg et al. 2018; Choi and Lee 2020; Weissensteiner et al. 2020; Huang et al. 2022). Extensive empirical evidence on SV-TE associations and TE proliferation shows that TEs tend to accumulate in genomic regions with suppressed recombination (Brennecke et al. 2007; Yang and Xi 2017; Ozata et al. 2019; Gebert et al. 2021). However, there is little agreement on the nature of the evolutionary forces shaping TE abundance levels (Dolgin and Charlesworth 2008). Ultimately, better genome assemblies will increase our understanding of how different evolutionary forces shape genome structure and TE content.

The genus Drosophila has been a model for studying eukaryotic genome evolution (Richards et al. 2005; Drosophila 12 Genomes et al. 2007; Bracewell et al. 2019). Although new genome assemblies based on long-read sequencing have emerged for several species in this genus (Allen et al. 2017; Mahajan et al. 2018; Miller et al. 2018; Liao et al. 2021), evolutionary inferences about the role of structural variation on species divergence are still limited. Genome assemblies for several Drosophila species that were first sequenced using either Sanger or short-read sequencing (Drosophila 12 Genomes et al. 2007) have yet to be updated. It is therefore important to improve the quality of those genome assemblies using the latest sequencing technologies.

D. pseudoobscura and *D. persimilis* are recently diverged species (<1 Mya) that represent a classic species pair widely used in speciation genetics research (Dobzhansky 1944; Orr 1987; Noor et al. 2001a, 2001b; Machado et al. 2002; Kulathinal et al. 2009; Fuller et al. 2018; Korunes et al. 2021). *D. pseudoobscura* is distributed across the western half of North America inhabiting environments that range from temperate forests to deserts. *D. persimilis* occurs in sympatry with *D. pseudoobscura* in a restricted range in the western Pacific coast states and mostly inhabits temperate forests (Dobzhansky and Epling 1944). The genomes of these species is organized in four telocentric chromosomes (2nd, 3rd, 4th, 5th), and the metacentric X chromosome. The karyotypes of the two species differ by fixed paracentric inversions in chromosomes 2 and in the left arm of chromosome X (XL) (Tan 1935; Anderson et al. 1977; Schaeffer et al. 2008). Furthermore, a large inversion in the right arm of chromosome X (XR) is fixed among *D. pseudoobscura* and non Sex-Ratio (SR) XR *D. persimilis* strains (Policansky and Zouros 1977). In addition, chromosome 3 harbors a diverse suite of inversions that are polymorphic in each species, with one shared arrangement (Standard or "ST") (Dobzhansky 1944; Fuller et al. 2019).

Genome assemblies for *D. pseudoobscura* (Richards et al. 2005) and *D. persimilis* (Drosophila 12 Genomes et al. 2007) were first published more than a decade ago. Recent sequencing projects have reported more contiguous genome assemblies for *D. pseudoobscura* based on long reads, or a combination of long reads and Hi-C (Miller et al. 2018; Bracewell et al. 2019; Liao et al. 2021), resulting in a new high-quality reference genome assembly (Liao et al. 2021). For *D. persimilis*, two assemblies based on Nanopore long reads were recently published (Miller et al. 2018; Kim et al. 2021), but their utility for studying SVs and their divergence is limited due to their highly fragmented nature. Recent work that reported genome assemblies for other *D. pseudoobscura* group species has provided evidence that centromere evolution in this group is driven by TEs, although *D. persimilis* was not included (Bracewell et al. 2019). The lack of a high-quality contiguous

genome assembly for *D. persimilis* hampers our ability to address questions about the effect of SVs on genome and gene expression divergence between this species and *D. pseudoobscura*.

Here we present the most highly contiguous genome assembly and annotation available for *D. persimilis*, together with a new high-quality genome assembly for *D. pseudoobscura*. We selected strains that have not yet been sequenced to facilitate future studies addressing intraspecific variability in SVs and TE content in these species. We present the first characterization of genome-wide patterns of structural divergence between these species, testing the hypothesis that TEs are directly involved in the generation of structural variation between species. Further, we assess the overall differences in gene content and structure between the genomes characterizing correlations between SVs, TE content, and recombination rate. Finally, we assess the association of SVs with protein coding genes and their effects on differential gene expression, focusing on genes located inside the fixed chromosomal inversions that separate these two species.

Results

Highly Contiguous Genome Assemblies for *D. persimilis* and *D. pseudoobscura*

We report the first highly contiguous genome assembly for *D. persimilis* (Strain: *Mather 40* (Machado et al. 2002)) and a high-quality genome assembly for a new strain of *D. pseudoobscura (Strain: Dpse\wild-type, 14011-0121.41)*. Our genome assembly approach resulted in the capture of all Muller elements in 11 contigs for *D. pseudoobscura* and 13 contigs for *D. persimilis* (Fig. 1a). We were able to assemble chromosomes 2, 3, and 5 (Muller elements E, C, F) in single contigs in *D. persimilis*. Chromosomes 4 and 5 (Muller element B and F) were assembled in single contigs in *D. pseudoobscura* (Fig. 1a). The third chromosome karyotype for the *D. pseudoobscura* line is AR based on full collinearity with the MV225 reference genome (supplementary fig. S22, Supplementary Material online). The third chromosome karyotype for *D. persimilis* Mather 40 is unknown but is not ST based on the complex rearrangement pattern shown in the comparison with the MV225 genome (supplementary fig. S23, Supplementary Material online).

The genome assembly for *D. pseudoobscura* appears to be less fragmented (lower number of contigs) but the N50 value is higher in *D. persimilis* (Fig. 1b). A summary of genome assembly statistics for both species can be found in Table 1. Genome sizes estimated using the k-mer count distribution in the Illumina reads (Vurture et al. 2017) were 134.6 Mb (26.4% repetitive) for *D. pseudoobscura*, and 145.5 Mb (33.3% repetitive) for *D. persimilis*. The genome assembly for *D. pseudoobscura* covered a total of 162.6 Mb with a GC content of 45.25%. For *D. persimilis*, the genome assembly covered 160.6 Mb with a GC content of 45.08% (Table 1). Completeness assessment using Benchmarkingsets of Universal Single-Copy Orthologs (BUSCO) showed a single-copy ortholog coverage of 98.4% and 98.8% for *D. pseudoobscura* and *D. persimilis,* respectively (Table 1). Overall, our genome assemblies are more contiguous than other assemblies publicly available (Fig. 1b) and add up to the vast repertoire of genomic resources of Drosophila species.

We provide confirmation of the fixed chromosomal inversions in chromosomes 2, XL and XR, plus the different arrangement between the two strains on chromosome 3, (Figs. 1a and 2). We also confirm a pattern where all derived inversions between the two species appear to have arisen in *D. persimilis* (Tan 1935; Machado et al. 2007). Moreover, we report 9 additional micro inversions (10 to 148 Kb) that also appear to be derived in *D. persimilis* (supplementary table S1, Supplementary Material online).

Conserved Gene Collinearity but Increased Transcript Length in *D. persimilis*

More protein coding genes were annotated in *D. pseudoobscura* (14,503 vs. 13,888), but the number of annotated ncRNAs and transcript lengths are significantly higher in *D. persimilis* (Fig. 1, supplementary table S2, Supplementary Material online). Because the overall transcript length is longer in *D. persimilis* (Fig. 1c), a significantly higher proportion of bp is annotated as mRNA ($X^2 = 4{,}011.4$, P-value < 2.2e-16) and ncRNA in this species ($X^2 = 88{,}2357$, df = 1, P-value < 2.2e-16, Fig. 1d). However, the difference in transcript length is due to UTR length and not to intron size (supplementary fig. S1, Supplementary Material online). When 3′ and 5′ UTRs are included, whole gene span is significantly longer in *D. persimilis* for chromosomes 2, 4 and XL (Fig. 1c). The number of annotated mRNAs is higher in *D. pseudoobscura* only in the XR chromosome, and the number of annotated ncRNAs is higher in *D. persimilis* for all chromosomes except chromosome 5. There is also a strong positive correlation between the two species for both transcript length and intron size (supplementary fig. S1, Supplementary Material online). Even though *D. persimilis* has a higher proportion of longer transcripts, a general linear model (GLM) predicts longer transcripts in *D. pseudoobscura* for genes that are longer than ~15 Kb (supplementary fig. S1a, Supplementary Material online). Similar results are observed for intron size where *D. persimilis* still has more genes with longer introns, but the GLM still predicts longer introns for *D. pseudoobscura* in long genes (supplementary fig. S1b, Supplementary Material online).

Using the longest isoforms for each protein-coding gene, OrthoFinder found a total of 11,322 single-copy orthologs between *D. pseudoobscura* and *D.* persimilis. About 90% of the genes have a transcript length between

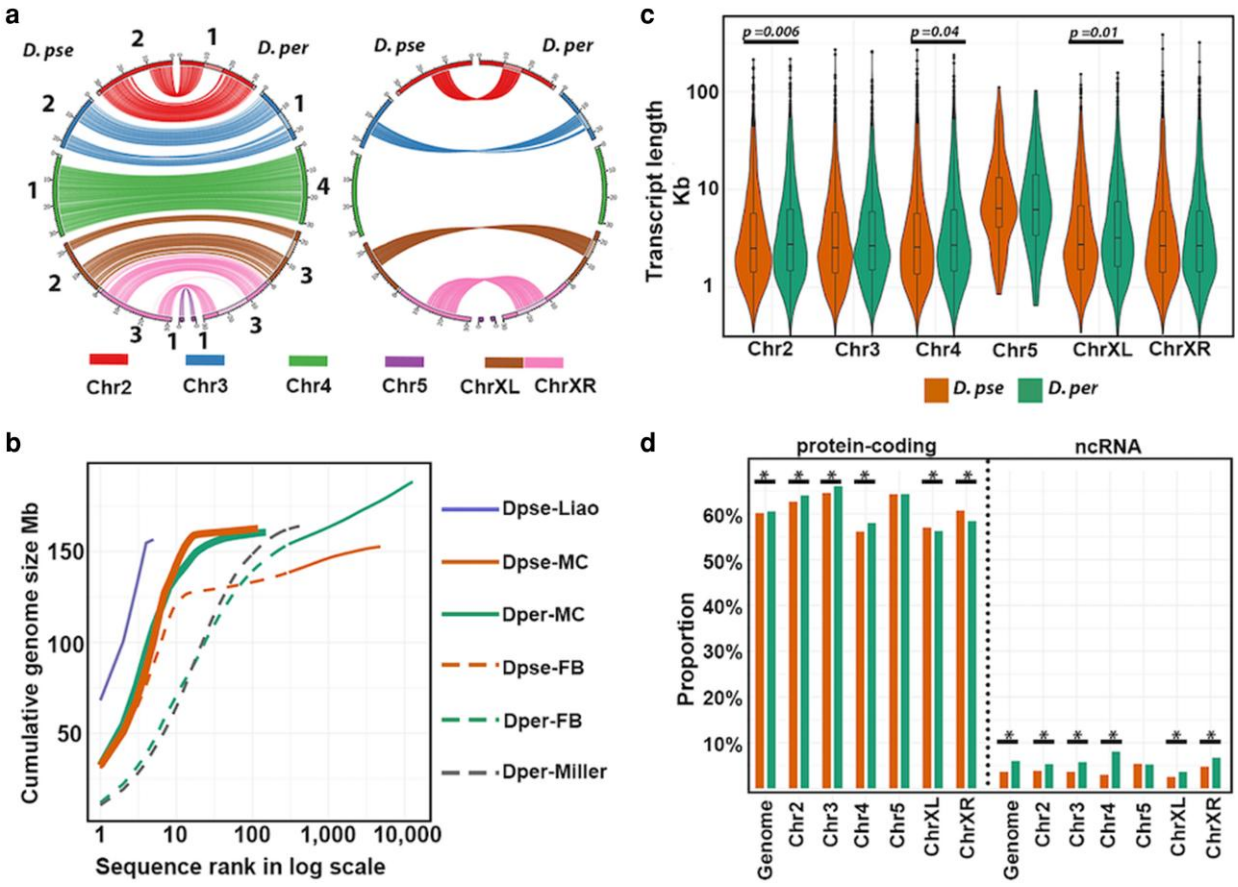

Fig. 1—. a) Circos plots showing collinear blocks (left) and inverted regions between *D. pseudoobscura* and *D. persimilis* (right). The number of contigs for each chromosome is indicated on the left plot. b) Comparison of assembly contiguity with previously published assemblies. MC: our study using CLR and short reads; Miller: assembly based on ONT reads (Miller et al. 2018); FB: FlyBase genome assemblies (Thurmond et al. 2019); Liao: most recent *D. pseudoobscura* reference genome based on CLR reads and Hi-C data (Liao et al. 2021). c) Comparison of transcript length for homologous genes. d) Comparison of the proportion of base pairs annotated as mRNA or ncRNA. * Significant difference between species $P < 2.2e\text{-}16$.

Table 1
Assembly statistics for *D. pseudoobscura* and *D. persimilis*

	D. pseudoobscura	*D. persimilis*
Total number of contigs	118	137
Contigs assigned to a chromosome	11	13
Chromosome-assigned coverage (Mb)	143.6	140.5
Maximum contig length (Mb)	32.1	32.2
Unassigned contigs (Mb)	19	20.1
Total assembly coverage (Mb)	162.6	160.6
N50 (Mb)	17.3	18.7
GC content	45.25%	45.08%
Complete BUSCOs (Insecta)	1,632 (98.4%)	1,639 (98.8%)

300 and 15,000 bp for both species, and the remaining 10% includes genes having a length between 15 and 400 kb (supplementary table S2, Supplementary Material online). Although some conservation in transcript length is observed between the two species, 54% (6,112) of the genes are longer in *D. persimilis*, 39% (4,444) longer in *D. pseudoobscura* and only 6.8% (779) of ortholog genes have the exact same transcript length in both species. These proportions change when considering amino acid length: 46% (5,277) of the genes have the same amino acid sequence length, 29% (3,254) are longer in *D. persimilis* and 25% (2,804) are longer in *D. pseudoobscura*, and amino acid length can differ between species up to 20%.

We observe strong conservation of gene collinearity with a small number of species-specific gene translocation events. Using genome assemblies from *D. miranda* and *D. lowei* we found that 11,628 out of 14,547 genes annotated in the outgroup *D. lowei* are collinear across all species of the pseudoobscura subgroup ("cl.cl.cl" code; see methods). When the other three species are taken as a query, the number of collinear genes ranges from 11,627 (*D. persimilis*) to 12,361 (*D. miranda*). This range in the number of genes reflects the existence of potential gene

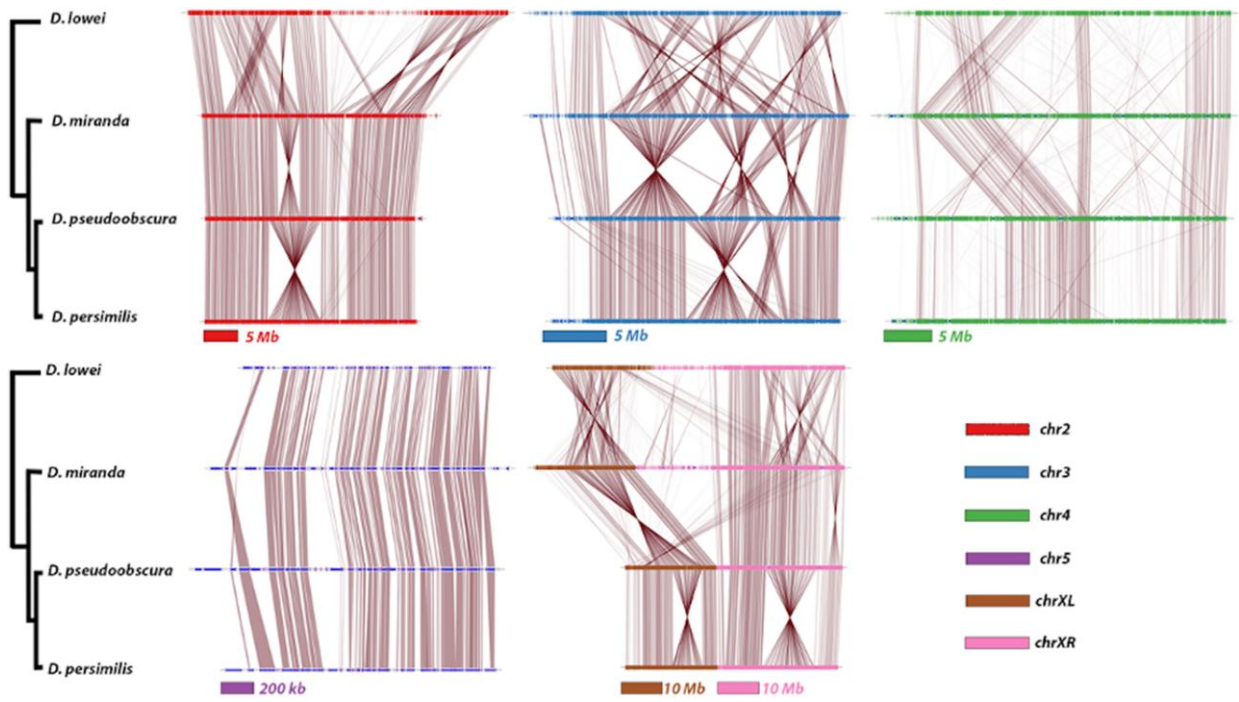

Fig. 2.—Gene collinearity plots for 1-to-1 single-copy orthologous genes across the pseudoobscura subgroup. Chromosomes are color-coded as in Fig. 1a. Vertical brown lines represent single-copy orthologs identified by OrthoFinder.

duplications or contractions occurring in each species. We also counted the number of potential lineage-specific translocations ("tr.tr.tr" code) and found a total of 125 and 159 translocations for *D. persimilis* and *D. pseudoobscura*, respectively. Of those, 54 interchromosomal translocations have happened in *D. pseudoobscura* and 40 in *D. persimilis*.

We further analyzed the position of 8,247 single-copy orthologs assigned by OrthoFinder to determine changes in collinearity among the four species of the pseudoobscura subgroup. As expected, collinearity among single-copy orthologs is highly conserved between *D. persimilis*, *D. pseudoobscura*, and *D. miranda* (Fig. 2). Although collinearity can be disrupted by chromosomal rearrangements such as inversions, we still detected strong collinearity within the large inversions from chromosomes 2, XL, XR, and 3 (Fig. 2). Only 39 single-copy ortholog pairs were found annotated in different chromosomes between *D. pseudoobscura* and *D. persimilis*.

Structural Variants Spatially Associated With Genes are More Frequent Inside Chromosomal Inversions

We characterized all structural differences between the genomes of *D. pseudoobscura* and *D. persimilis*. Using *D. pseudoobscura* as a reference we called a total of 7,941 INDELs (3,181 INSertions and 4,760 DELetions) (Fig. 3a). We also called a total of 551 and 322 Copy-number variants (CNVs) for *D. pseudoobscura* and *D. persimilis*, respectively (Fig. 3a). Our analyses reveal a greater accumulation of INS in *D. persimilis* (Fig. 3a). Nevertheless, the size distribution of INDELs suggest that INS in *D. pseudoobscura* are larger than in *D. persimilis* (Fig. 3a; Mann–Whitney *U* test, $P = 2.69e\text{-}10$). Further, the number of identified CNVs is greater in *D. pseudoobscura* but they have a similar size distribution in *D. persimilis* (Fig. 3a; Mann–Whitney *U* test, $P = 0.388$).

Close to 40% of all genes are spatially associated with an SV in *D. pseudoobscura* and *D. persimilis*. Although our results show that the overlap between SVs and the complete transcript span of annotated genes is lower than expected by chance (supplementary table S3, Supplementary Material online), correspondence analyses show that there is a significant association between SVs and the 10Kb upstream sequences of annotated genes in chromosomes 2, 4, 5, and XL in both species (Fig. 3b).

Interestingly, we found that genes located inside the major fixed inverted regions (INV) are more likely to be associated with SVs than genes in collinear regions (COL). For both species we found a significantly higher proportion of INDELs associated with genes within inversions than in COL for chromosomes 2 and XL, but not in chromosome XR (supplementary fig. S2, Supplementary Material online). The proportion of CNVs is higher in the INV of chromosome XR only for *D. pseudoobscura* and for chromosome XL in the two species (Fig. 3b, supplementary fig. S2, Supplementary Material online).

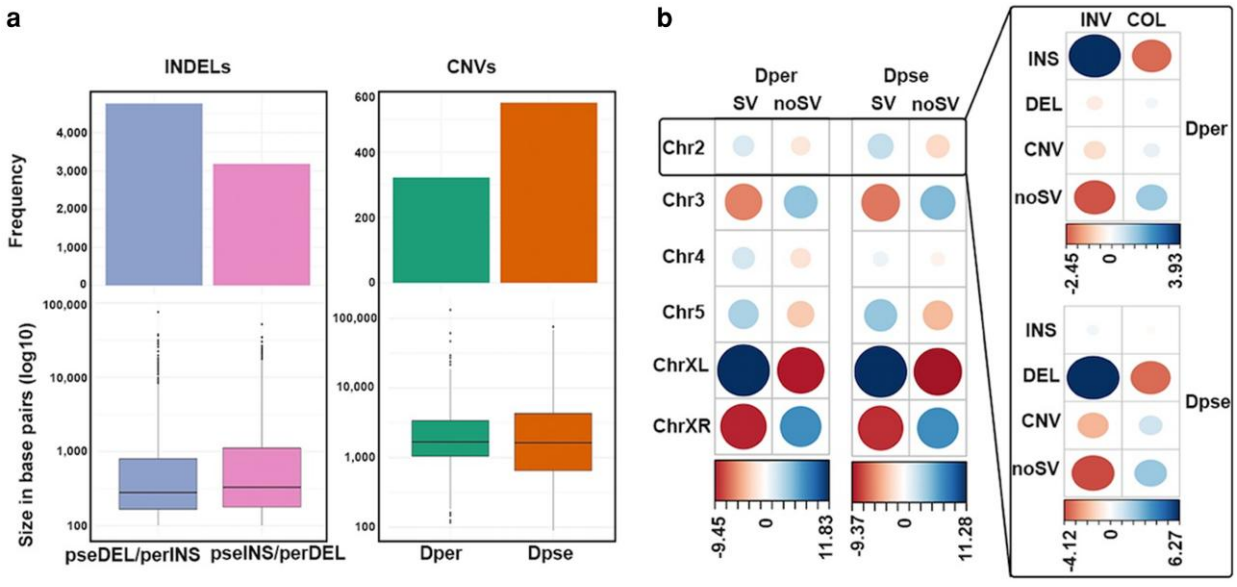

Fig. 3—. a) number of INDELs and CNVs (barplots) and size distribution for each SV type (boxplots). b) Correspondence analysis showing the association between genes (including the 10 kb upstream region) and SVs, for each chromosome. Circle sizes depict the number of genes, and color depicts correlation values. The inset for chromosome 2 shows a more detailed analysis comparing the fixed INV versus COL, for each variant type. INS, insertions; DEL, deletions; CNV, copy-number variants; noSV: genes not associated with SVs. See supplementary fig. S2, Supplementary Material online for detailed analyses for chromosomes 3, XL, and XR.

In addition to the previously identified large inversions occurring on chromosomes 2, 3, and X (Fig. 1a, supplementary table S1, Supplementary Material online), we identified and confirmed (using long reads) a total of 9 new micro inversion differences ranging between 10 and 148 kb in size. All these micro inversions are found in COL outside the major inversions, in all chromosomes except chromosome 5. Three of them are found on chromosome 4, a chromosome where no inversions between the two species have been previously reported (supplementary table S1, Supplementary Material online). Comparisons with *D. miranda* indicate that 8 of those 9 microinversions are derived in *D. persimilis* (no homologous regions were found for one of the inversions in the *D. miranda* genome). Analyses of genome assemblies from other strains will help determine if these newly identified micro inversions correspond to fixed inversions between this species pair.

Transposable Elements are Associated With SVs in *D. pseudoobscura* and *D. persimilis*

Transposable elements (TEs) are often associated with the generation of structural variation between species (Mérel et al. 2020). We investigated whether TE content is spatially correlated with all the called SVs between *D. pseudoobscura* and *D. persimilis*. Repeat masker annotations show a 25.5% and 21.7% repetitive sequence content in the *D. pseudoobscura* and *D. persimilis* genome assemblies, respectively (supplementary table S4, Supplementary Material online), although TE content is slightly higher in *D. persimilis* (17% vs. 16%). In addition, we ran the RepeatMasker annotation pipeline on the genomes of *D. miranda* (Mahajan et al. 2018) and *D. lowei* (Bracewell et al. 2019) finding that these genomes have 26.4% and 28.8% of total repetitive sequence content, respectively (supplementary table S4, Supplementary Material online), and that the four species share the most abundant TE classes and families (Fig. 4a and b). Although a considerable proportion of TEs were annotated as "unknown" (Fig. 4a), most TE annotations fall in four TE classes and 10 TE families (Fig. 4b). Further, *D. persimilis* has significantly more copies of all the most abundant TE families (except Tc1/mariner and Maverick) than *D. pseudoobscura* (supplementary fig. S36, Supplementary Material online), suggesting a higher level of activity of these TE families in *D. persimilis*. TE divergence plots (supplementary figs. S37 and S38, Supplementary Material online) and estimates of the average percent divergence for each TE family (supplementary table S15, Supplementary Material online) show that the Maverick elements have recently invaded these genomes. Interestingly, Helitron family elements show a very distinctive burst of proliferation at Kimura distance of 0.15 only in *D. pseudoobscura*, consistent with a previous analysis of the reference genome of this species (Petersen et al 2019).

Almost every TE family is significantly associated with INDELs in both species suggesting that TEs are a primary source of INDEL generation (Fig. 4c; supplementary fig. S3, Supplementary Material online). Interestingly, only in

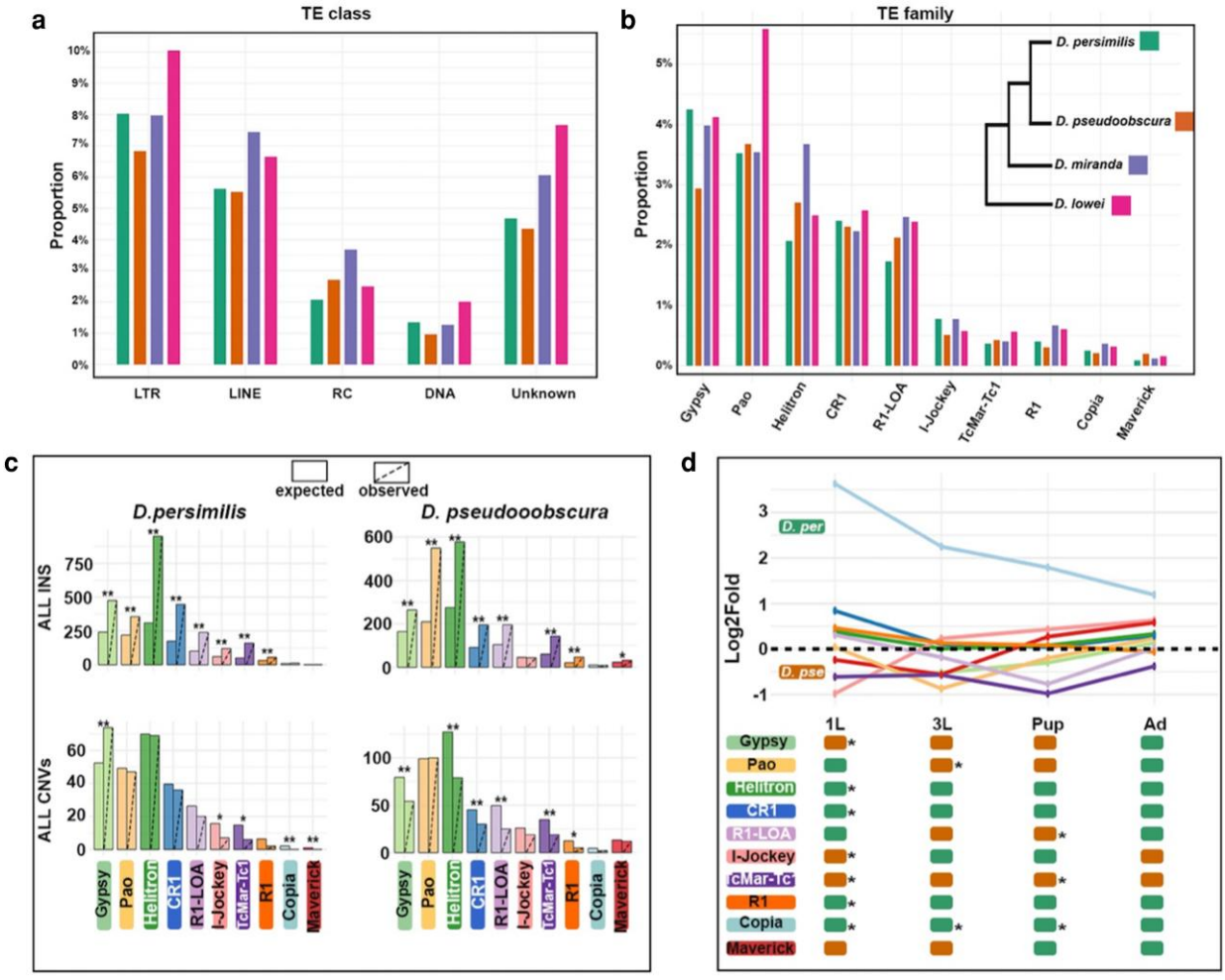

Fig. 4—. TE content, TE-SV associations and TE expression of the pseudoobscura subgroup. a and b) proportion of TE classes and families across the pseudoobscura subgroup. c) Permutation analysis of INS and CNVs overlapping TE annotations; * $P < 0.05$; ** $P < 0.01$ significant difference between observed and expected counts. d) Differential expression analysis during development for each TE family. The scatterplot depicts the log2fold expression change, relative to *D. pseudoobscura*, of each TE family during development. Ovals at the bottom illustrate in which species each TE family shows higher or significantly higher (*) expression levels. 1L: first instar larvae; 3L: third instar larvae; Pup: pupae; Ad: adult.

D. persimilis there is a significant association of the Gypsy family with annotated CNVs (Fig. 4c), whereas most of the TE families are significantly underrepresented in CNVs.

Using RNA-seq data from four developmental stages (supplementary fig. S4, Supplementary Material online), we found significant differences between species in the expression levels of most TE families during development, with most of the significant differences observed in first instar larvae (Fig. 4d). Some of the expression differences are consistent with the observed relative differences in total (supplementary fig. S36, Supplementary Material online) or intact (supplementary fig. S37, Supplementary Material online) TE copy numbers in the genome assemblies. For instance, the Helitron, CR1, R1, and Copia families are expressed at higher levels in *D. persimilis* consistent with their higher abundance in its genome. On the other hand, the Gypsy family shows the opposite pattern: more highly expressed in *D. pseudoobscura* even though the *D. persimilis* genome has a significantly higher content. Overall, TE genome content and gene expression are positively correlated in both species, for both intact and total TE copy numbers, but only significantly correlated in *D. pseudoobscura* across most developmental stages (supplementary figs. S40 to S47, Supplementary Material online). These results are consistent with previous observations in vertebrate lineages showing similar strong linear relationships between germline TE expression and TE genome content (Pasquesi et al 2020), consistent with expectations of a stochastic model of genome-wide transcription (Encode Project Consortium 2012). Given that TE expression patterns suggest potential species differences in TE regulatory mechanisms (Lee et al. 2004), we compared the expression of genes known to be

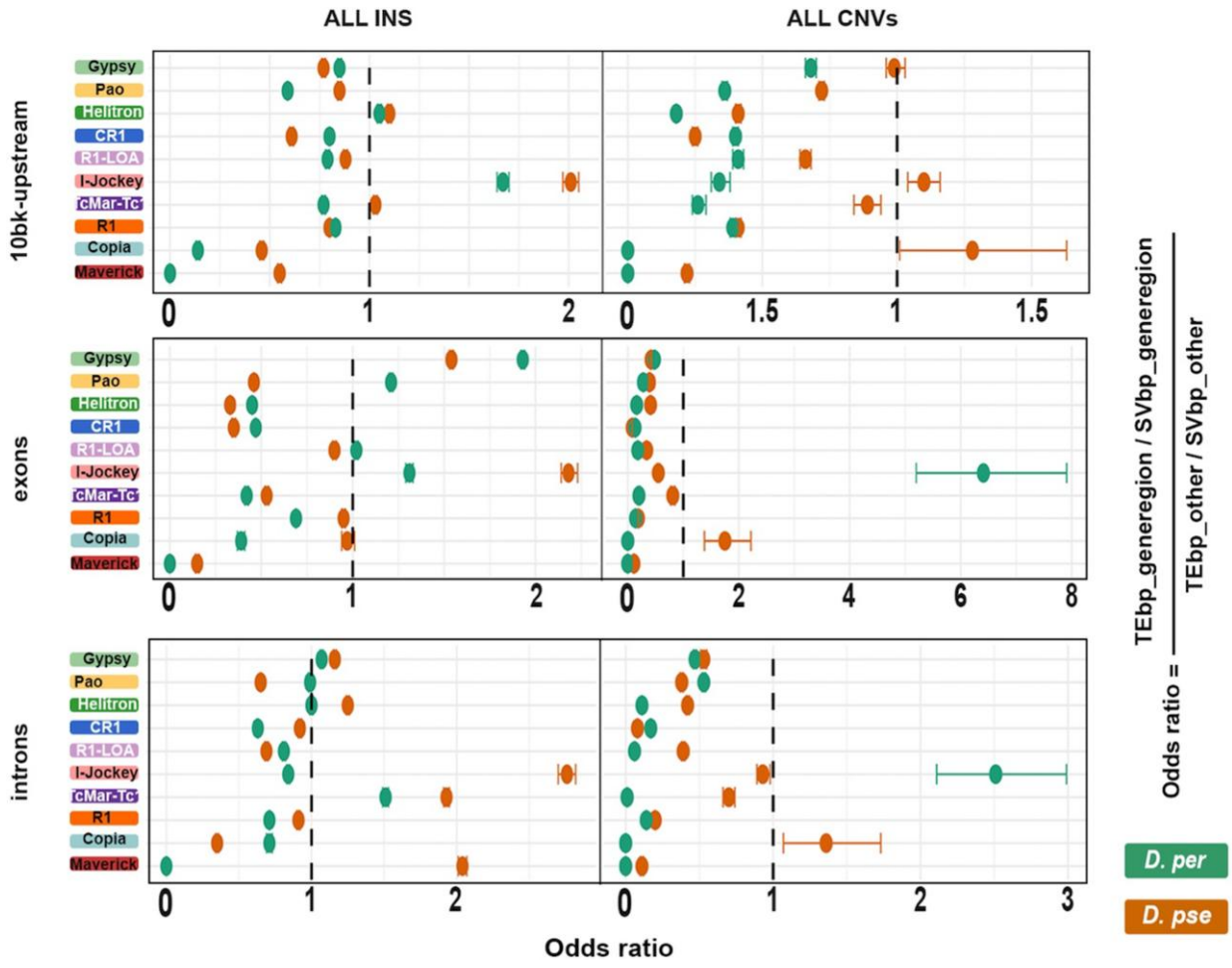

Fig. 5—. Odds ratios of 2 × 2 contingency tables for TE-SV (INS and CNV) associations with different gene regions (10 kb-upstrean, exons, introns).

involved in the regulation of TE expression in Drosophila (Ozata et al. 2019). Two genes known to be at the center of mechanisms of defense against TE proliferation in the germline (*Rhi*) and in somatic tissue (*Dcr-2*) show significant expression differences between species. *Rhi*, a gene that regulates Piwi-interacting RNA (piRNA) expression, is upregulated in *D. pseudoobscura adult females, and in D. pseudoobscura* males across all developmental stages analyzed in this study (supplementary fig. S6, Supplementary Material online). *Dcr-2 a* gene involved in the generation of siRNAs (Galiana-Arnoux et al. 2006; Chung et al. 2008; Naganuma et al. 2021), is expressed at significantly higher levels in the third instar larvae of *D. persimilis in* both sexes (supplementary fig. S6, Supplementary Material online).

Even if there are striking genome-wide differences across TE families, TE associations with gene regions are not always enriched with the most abundant TE families (Figs. 4 and 5). We calculated the odds ratios between each TE family overlapping SVs within gene regions and intergenic regions (Fig. 5) and found that even though TEs are significantly underrepresented in gene regions (supplementary fig. S7, Supplementary Material online) some TE families are significantly associated with INS and CNVs within or close to gene regions. Members of DNA transposon families i-Jockey and Helitron are enriched near INS located in the 10 kb regions upstream of genes in both species, while the DNA transposon family TcMar-Tc1 is enriched only in *D. pseudoobscura*. For CNVs, the Copia and i-Jockey families are enriched only in *D. pseudoobscura*. INS associated with exons appear to be associated with the Gypsy and i-Jockey families in both species, whereas Pao is enriched only in *D. pseudoobscura*. CNVs associated with exons appear to be associated with i-Joc key in *D. persimilis* and with Copia for *D. pseudoobscura*. Further, INS associated with introns are enriched for Gypsy and TcMar-Tc1 in both species but for Helitron, i-Jockey, and Maverick only in *D. pseudoobscura*. CNVs associated with introns are enriched with i-Jockey in *D. persimilis* and with Copia in *D. pseudoobscura* (Fig. 5).

Accumulation of Transposable Elements in Regions of Low Recombination

We estimated population-based fine-scale recombination rates for both species and observed significant negative correlations between TE content and recombination rates in both species ($P < 0.05$) (supplementary fig. S7, Supplementary Material online). Although recombination rates seem to be lower in all *D. persimilis* chromosomes (supplementary figs. S8 and S9, Supplementary Material online), this is likely a function of differences in effective population size between the species. Comparing collinear and INV we observe that recombination rates are significantly higher in the fixed INV from chromosome X in both species (supplementary figs. S10 and S11, Supplementary Material online), but significantly lower in the INV of chromosome 2 in *D. persimilis* (supplementary fig. S11, Supplementary Material online). Consistent with the expected negative correlation between TE content and recombination rate (Dolgin and Charlesworth 2008), TE content is significantly lower inside the INV of chromosome X in both species, while the INV of chromosome 2 in *D. persimilis* shows a slight albeit nonsignificant increase in TE content (supplementary figs. S10 and S11, Supplementary Material online). Further, there is also a significant decrease in TE content in chromosome 3 for both species (supplementary figs. S10 and S11, Supplementary Material online).

We observe significant increases in the proportion of TEs in the inversion breakpoint regions from chromosomes 2 and XL. For chromosome 2, we observe a significant increase in TE content on both sides of the proximal inversion break point in *D. pseudoobscura* (Fig. 6). A similar but more pronounced pattern is observed toward the distal inversion break point in *D. persimilis*, and this region has the highest overall TE content compared to *D. pseudoobscura* (Fig. 6). Interestingly, this increase in TE content around these inversion breakpoints is also accompanied by a reduction on the local recombination rate, where *D. persimilis* shows a block of reduced recombination of ~350 kb that overlaps with the inversion breakpoint (supplementary fig. S12, Supplementary Material online). These results imply that TEs were already abundant around the breakpoint regions in the ancestor of both species, facilitating the generation of inversions in *D. persimilis* and further accumulation of TEs due to strong reduction in recombination rate around the breakpoints.

Among the annotated TE families, we observed that their proportion varies toward the inversion break points, while elements annotated as "Unknown" are highly abundant in the two species. In addition, we observed that the four most abundant TE families in *D. persimilis* are present in similar proportions at the closest window of the distal inversion breakpoint, and that the Helitron family is highly abundant in the COL just outside the proximal inversion breakpoint for both species (Fig. 6). For chromosome XL, we only observe a significant increase in TE content toward the proximal inversion breakpoint (within the inversion) in *D. persimilis* (Fig. 6). Although the negative correlation between TE content and recombination rate is less obvious for the breakpoints from chromosome XL, upstream and downstream regions to the breakpoints in both species show peaks of elevated TE content in low recombining regions (supplementary fig. S13, Supplementary Material online). Even though "Unknown" TEs are highly abundant in both species, we observed that Gypsy is highly abundant in the proximal inversion break point in *D. pseudoobscura*, but CR1 is more abundant in the corresponding distal breakpoint region in *D. persimilis* (Fig. 6).

Genome-Wide Gene Differential Expression is Significantly Associated With SVs in *D. pseudoobscura* and *D. persimilis*

We assessed differences in gene expression between both species for a total of 8,639 one-to-one single-copy orthologous genes using RNA-seq data from four different developmental stages (see methods). We analyzed each developmental stage independently to unveil patterns of gene expression across development. A total of 659, 714, 727, and 740 genes constituted the top 5% of the differentially expressed (DE) genes in first instar larva (1L), 3L, midstage Pupa and Adults, respectively. Overall, a higher proportion of genes are more highly expressed in *D. persimilis*, except in the 3L stage where there is a higher frequency of genes more highly expressed in *D. pseudoobscura* (supplementary fig. S14, Supplementary Material online). Gene ontology (GO) enrichment analyses of DE genes show overrepresentation of genes involved in a wide variety of functions from gene regulation to general developmental processes (supplementary table S5 to S6, Supplementary Material online).

Correspondence analyses, considering all INDELs and CNVs, indicate that DE genes are significantly associated with SVs in both *D. pseudoobscura* and *D. persimilis* (Fig. 7a; supplementary figs. S14 to S17, Supplementary Material online). This correlation signal mostly arises from SVs that overlap the 10 kb upstream region of genes. Our results also indicate that there is a strong association between differential expression and SVs on genes that are located inside the INV. This pattern is stronger in chromosome 2 and is significant across all developmental stages (Fig. 7a and b; supplementary figs. S14 to S17, Supplementary Material online). For chromosome XL, we only observed a strong association between SVs and DE genes in INV during the pupal stage; for chromosome XR the association is significant on genes expressed during

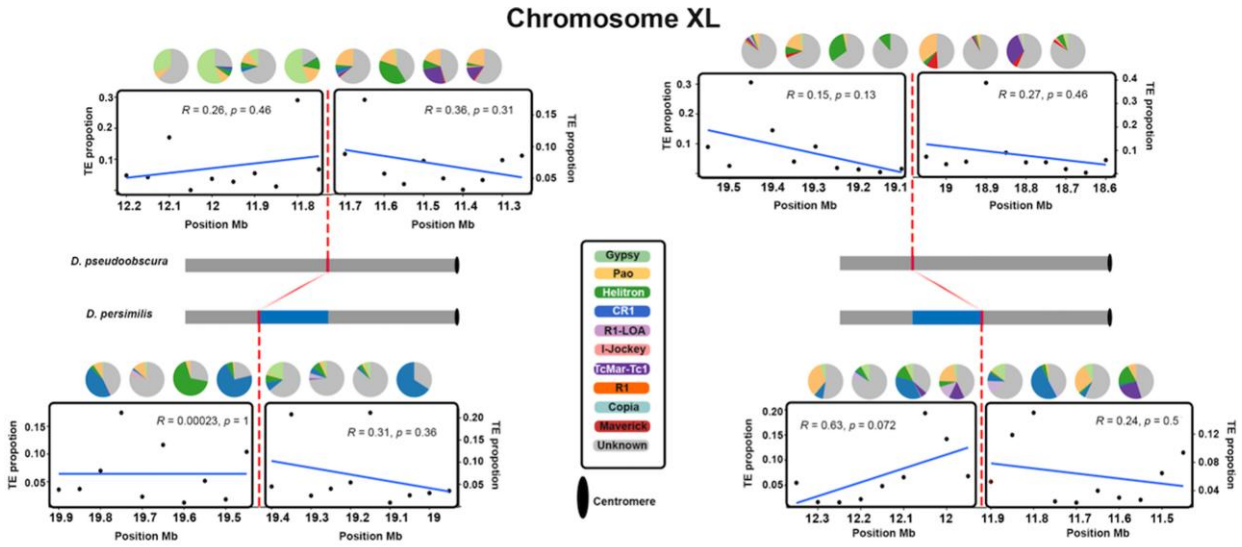

Fig. 6—. TE content at the proximal *D. pseudoobscura* and distal *D. persimilis* (left) and distal *D. pseudoobscura* and proximal *D. persimilis* (right) inversion breakpoints in chromosomes 2 and XL. Each dot from the scatterplots represents a 50 kb sliding window. Solid and dashed red lines depict the inversion break points. The blue section of the chromosome represents the INV. Pie charts show the proportion of the color-coded TE families in the four 100 kb windows closest to the inversion break points.

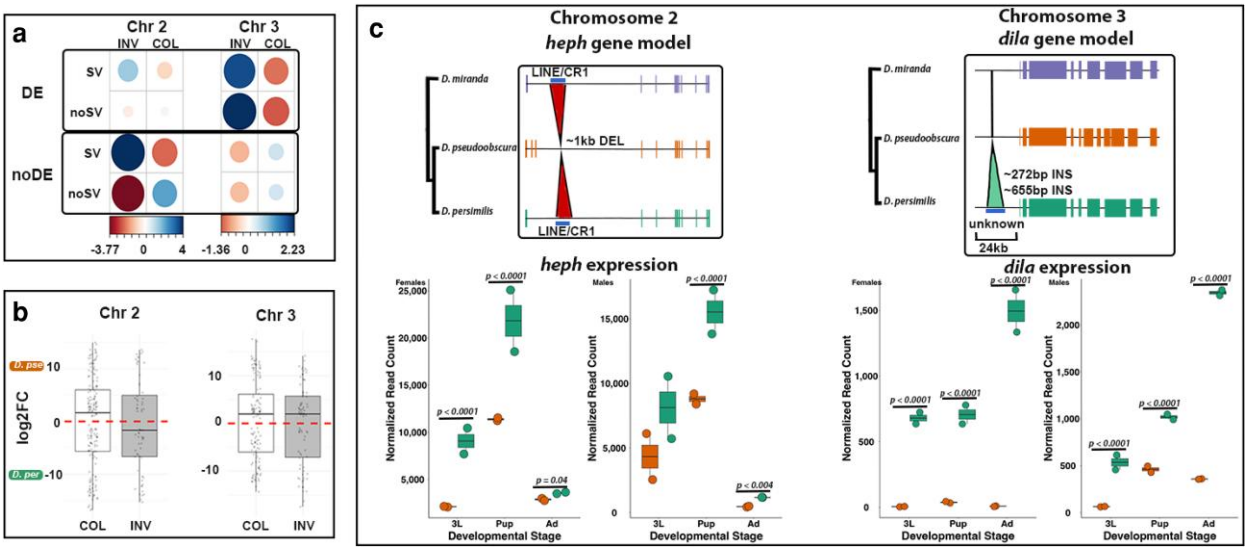

Fig. 7—. Gene expression and its association with SVs for chromosomes 2 and 3 in *D. pseudoobscura* and *D. persimilis* for the 3L developmental stage. a) Correspondence analysis showing the association of genes DE or not (noDE) with the presence or absence of SVs in the 3L stage; circle sizes depict number of genes, and color depicts correlation values (contribution to the overall Chi-square statistic). b) Log2 fold change values for DE genes comparing collinear and inverted regions; >0 higher expression in *D. pseudoobscura*; <0 higher expression in *D. persimilis*. c) heph and dila gene models of *D. miranda, D. pseudoobscura, and D. persimilis* showing a deletion affecting the third intron in *D. pseudoobscura* (heph—left) and an insertion affecting the upstream region in *D. persimilis* (dila—left). Boxplots show the DESeq2 normalized read counts for heph and dila over 4 developmental stages 3L: third instar larvae, Pup: Pupae, Ad: Adult, between *D. pseudoobscura* (left datapoints in each boxplot) and *D. persimilis* (right datapoints in each boxplot). Expression for 1L: first instar larvae can be found in supplementary fig. S34, Supplementary Material online.

the first (1L) and third (3L) instar larvae stages (Fig. 7b, supplementary figs. S14 to S17, Supplementary Material online). For chromosome 3, which harbors a rich suite of nonfixed polymorphic inversions, we observed a significantly high proportion of DE genes inside the INV regardless of their association with SVs.

Lineage Specific SVs are Associated With Genes Involved in Neural System Development and Gametogenesis Inside the Fixed Inversions

We found a strong association between SVs and genes in the INV, specifically for genes that are involved in neural development and spermatogenesis. Using genome-based called variants with *D. miranda*, we identified a total of 852 and 656 lineage-specific deletions and insertions, respectively, in *D. pseudoobscura* and a total of 689 and 793 lineage-specific deletions and insertions, respectively, in *D. persimilis*. In addition, we identified a total of 150 CNVs in *D. pseudoobscura* and a total of 133 CNVs in *D. persimilis*. The Helitron family is the main TE family associated with recent INDELs for both species (supplementary figs. S48 to S51, Supplementary Material online), although for CNVs no family other than TEs classified as "Unknown" show significant abundance (supplementary figs. S52 and S53, Supplementary Material online).

We selected all the DE genes located inside the inversions that were associated with the polarized SVs (see methods) and ran a second GO enrichment analysis focused on all the genes located inside inversions. Our results indicate an overall overrepresentation of DE genes involved in neural system development (GO:0007399) during 1L, 3L, and Pupal stages, and in protein and nutrient transport in adults (supplementary table S5 to S7, Supplementary Material online). While inversions in chromosomes 2 and XR show the highest overrepresentation of genes involved in neural system development in both species, we did not find any overrepresented category in chromosome XL (supplementary table S7, Supplementary Material online). Finally, in the polymorphic inversion from chromosome 3, we found an overrepresentation of genes involved in transport activity (GO:0005215) in *D. persimilis* and in transcriptional silencing (GO:0016458) in *D. pseudoobscura*. Interestingly, none of the DE genes associated with SVs in the third chromosome overlap with the genes reported by Fuller et al. (2016) that are DE between different third chromosome arrangements in *D. pseudoobscura*.

Literature surveys confirmed the GO enrichment analysis for most of the genes associated with neural system development. Interestingly, we found that 3 of those genes (*cnc*, *dila*, *heph*) are also involved in spermatogenesis, while one gene is involved in oocyte-to-embryo transition (*nebu*) in *D. melanogaster* (Sridharan et al. 2016; Vieillard et al. 2016; Avilés-Pagán et al. 2020; Chen et al. 2020). *cnc* and *heph* are genes located in chromosome 2 that show *D. pseudoobscura*-specific indels, an INS in the 10 kb upstream region of *cnc* and a DEL in *heph* inside an intron (Figs. 7c, supplementary figs. S18 to S19, Supplementary Material online). The recent ~130 bp INS in *cnc* overlaps an "unknown" RepeatMasker annotation in *D. pseudoobscura*. We did not observe any TE annotation in the homologous region of *D. persimilis* or *D. miranda* (supplementary fig. S19, Supplementary Material online). For *heph* we observed that the recent ~1 kb DEL in *D. pseudoobscura* corresponds to a region overlapping a LINE/CR1 in both *D. persimilis* and *D. miranda* (Fig. 7). These two recent INDELs occurred in *D. pseudoobscura*, leading to a decrease in the level of expression relative to *D. persimilis* in both males and females for *heph* (Fig. 7c) whereas for *cnc* we observe an increase in the levels of expression in *D. pseudoobscura* (supplementary fig. S20 and S34, Supplementary Material online). In *D. persimilis*, *dila* (chromosome 3) shows two recent ~1 kb INS occurring close to each other (supplementary fig. S18, Supplementary Material online), whereas *nebu* (chromosome XR) has a recent ~130 bp DEL, both in the 10 kb upstream regions (supplementary fig. S19, Supplementary Material online). The recent INS in *dila* overlaps with an "unknown" RepeatMasker annotation, not found in *D. pseudoobscura* or *D. miranda*. In *nebu,* the recent DEL in *D. persimilis* corresponds to a region that overlaps with a RC/Helitron annotation in the three species (supplementary fig. S19, Supplementary Material online). *dila* and *nebu* are more highly expressed in *D. persimilis* for both males and females across development (Fig. 7c, supplementary fig. S20, Supplementary Material online).

Discussion

The use of long reads for genome assembly projects has enhanced our understanding of the origin and evolution of complex genomic variation (Bracewell et al. 2019; Hufford et al. 2021; Rhie et al. 2021). In this study, we generated the first high-quality genome assembly for *D. persimilis* along with a high-quality genome assembly for a new strain of *D. pseudoobscura*. Although Miller et al. (2018) reported the first *D. persimilis* genome assembly built with long (ONT) reads, we present here the first fully de-novo and chromosome-level assembly generated for this species using a mix of high-coverage PacBio and Illumina data. Independent genome-wide alignments against the most recent reference *D. pseudoobscura* genome (Liao et al. 2021) and mapping of long reads discarded any mis-assemblies, and show that there are no other major rearrangements in either of our assemblies (supplementary figs. S21 to S23, Supplementary Material online). We only observed a potential missing section corresponding to the centromeric region of the X chromosome (supplementary figs. S21 to S23, Supplementary Material online), which is not surprising given the difficulty of properly assembling centromeric regions due to their high repetitive element content (Rhie et al. 2021). In addition, we also compared our assemblies with the *D. pseudoobscura* genome assembly from FlyBase (FB) (r3.04) and observed discrepancies that can indicate potential misassembles in the FB genome. Although the correct order of contigs and the identification of potential

misassembles of the FB assembly were previously reported by Schaeffer et al. (2008), we detected two additional potential misassembles in chromosome 2 (supplementary figs. S24 to S25, Supplementary Material online).

Although some annotation discrepancies exist between our assemblies and the publicly available genomes (Liao et al. 2021), we observe consistency in the number of genes based on our ortholog and collinearity analyses. Our analyses revealed a significant difference in the number of annotated protein-coding genes between these closely related species, with the *D. pseudoobscura* assembly containing 615 additional genes. Although part of the difference may be the result of annotation artifacts, it is could also reflect biological differences between the species, as well as significant gene content differences that can happen among individuals of the same or very closely related species. Differences in gene content among individuals of the same species are well known in prokaryotes, leading to the concept of the "pan-genome", the overall gene content of a species (Tettelin et al. 2005). Those differences are being increasingly observed in eukaryotes (Gerdol et al. 2020; Hufford et al. 2021), and although little is known about Drosophila pan-genomes, our findings suggest more work needs to be done to study gene content differences within individuals and species of this genus.

Despite the larger number of predicted protein coding genes in *D. pseudoobscura*, we observed a higher number of predicted noncoding transcripts in *D. persimilis* (supplementary table S2, Supplementary Material online). Although differences in the number of predicted transcripts can be partially explained by annotation artifacts (Drosophila 12 Genomes et al. 2007), it is possible that the higher number of noncoding transcripts in *D. persimilis* is the result of spurious transcription or transcriptional noise (Ponjavic et al. 2007; Darbellay and Necsulea 2020), that could arise from lower selection efficiency in this species due to its smaller effective population size (Machado et al. 2002; Korunes et al. 2021).

We observed an overall conservation in the physical order and transcript length of orthologous genes among species of this species group. Even within inversions, we did not detect rearrangements disrupting overall gene collinearity between *D. pseudoobscura* and *D. persimilis*. Nevertheless, we were able to detect several potential gene translocation events occurring both within and between chromosomes (Fig. 2). For example, we observe that 8 genes originally located just outside the proximal inversion breakpoint in chromosome 2 seem to have moved closer to the centromeric region of the same chromosome in *D. pseudoobscura*. In this case, the source breakpoint and recipient centromeric regions have a high proportion of repetitive elements (Fig. 2), and the movement of genes in these species could be the result of recombination events mediated by TEs (Weckselblatt and Rude 2015).

The Landscape of Structural Variation in Chromosomal Inversions

One important challenge in speciation genomics research is elucidating the role of genome architecture in species divergence (Zhang et al. 2021). The advent of genomic analysis has increasingly shown that hybridization and introgression among closely related species have occurred frequently across the tree of life (Taylor and Larson 2019). One of the most important mechanisms that allow species to persist in the face of gene flow is chromosomal inversions (Hoffmann and Rieseberg 2008), and the two focal species of this study have been classic examples of the importance of chromosomal rearrangements for speciation (Dobzhansky 1944; Orr 1987; Noor et al. 2001a, 2001b; Machado et al. 2002; Fuller et al. 2018; Korunes et al. 2021). Our new assemblies allowed us to confirm not only the presence of the 3 large fixed chromosomal rearrangements that differ between *D. pseudoobscura* and *D. persimilis* (Chr. 2, XL and XR; supplementary figs. S26 to S33, Supplementary Material online) which were first inferred in the 1930s using cytogenetic analyses (Tan 1935), but also the known polymorphic inversion in chromosome 3 that distinguishes some strains of both species (Dobzhansky 1944; Fuller et al. 2019).

More importantly, we also report the presence of 9 additional microinversions (10 to 148 Kb) that differ between the two sequenced genomes. Eight of these microinversions appear to be derived in *D. persimilis* (one had no detectable homology in the outgroup *D. miranda*), consistent with the previous inference that the 3 large inversions in chromosomes 2, XL and XR are also derived in this species (Tan 1935; Machado et al. 2007). This bias toward derived inversions in *D. persimilis* remarkable as the probability of observing this many inversions by chance in the same lineage is low ($(½)^{11} = 0.00048$). This result is consistent across 3 available assemblies available for *D. pseudoobscura* (MV2-25, MV-25-SWS-2005, *Dpse\wild-type 14011-0121.41*), but it is still unclear if these smaller structural differences are fixed because we are only analyzing a single genome of *D. persimilis*.

Inversions can readily arise due to a variety of molecular mechanisms, most of which involve TEs (Huang and Rieseberg 2020). Although no specific TE families are associated with the generation of chromosomal rearrangements, the association between inversion breakpoints and TE content has been found across kingdoms (Zhang and Peterson 2004; Richards et al. 2005; Delprat et al. 2009; Bracewell et al. 2019; Sharma et al. 2021). Here we found that at least one inversion breakpoint overlaps with annotated INDELs in *D. persimilis*, and that the most abundant TE family (Helitron) is observed in 14% (3/22) of the inversion breakpoints (supplementary table S1, Supplementary Material online). Further, we found increases in the proportion of specific TEs right next to identified inversion

breakpoints, consistent with their potential role in the origin of these fixed inversions (Fig. 6).

Although former reports of increased genetic differentiation between the fixed inversions separating this species pair were based on SNP differences (Noor et al. 2007; Korunes et al. 2021), we also show that there is a significant overrepresentation of INDELs inside inversions in chromosomes 2 and XL (but not XR). We speculate that this pattern could have been generated by reduction in the efficiency of selection to remove slightly deleterious indels caused by (i) the initial reduction of recombination in the INV during the time these inversions were still polymorphic within species, (ii) by selection interference due to concurrent selective sweeps along the INV that could have also had an effect on reducing the efficacy of selection in removing slightly deleterious indels.

We show that genes located inside the major fixed INV show an overrepresentation of linked SVs, and that SVs are significantly associated with gene expression differences between species. Even though the lower frequency of SVs inside gene regions implies the effect of purifying selection, we found some SVs affecting not only potential regulatory elements in upstream regions but also overall gene structure (Fig. 7). Previous studies have provided vast evidence of SVs involved in gene expression differences that ultimately promote important phenotypic differences either between or within species (Jones et al. 2012; Chiang et al. 2017; Chakraborty et al. 2018; Alonge et al. 2020), and we show association patterns that suggest a significant relationship between SVs and differential gene expression between this species pair (Fig. 7, supplementary figs. S14 to S17, Supplementary Material online). Moreover, we observed a strong signal of differential expression for genes inside inversions compared to COL of the genome consistent with previous studies that show high levels of sequence divergence between inversions in this species pair (Machado et al. 2007; Noor et al. 2007; Kulathinal et al. 2009; Korunes et al. 2021). Therefore, the evolutionary forces that have influenced the increased sequence divergence between species at the nucleotide and structural levels inside fixed inversions have also led to increased gene expression differences between these species.

The Influence of Transposable Elements on Genomic Divergence

TEs appear to be the main players involved in the generation of structural variation in this group, similar to observations in other *Drosophila* species (Mérel et al. 2020) and in many model systems, including humans (Kofler et al. 2015; O'Neill et al. 2020). The genome coverage of repetitive elements is significantly different for several major TE families across the *pseudoobscura subgroup* (Fig. 4, supplementary table S4, Supplementary Material online), and a significant proportion of SVs overlap with TE annotations (Fig. 4). Consistent with previous findings (Drosophila 12 Genomes et al. 2007; Hill and Betancourt 2018), our annotation pipeline indicates that *D. persimilis* has a higher TE content (supplementary fig. S8 and table S4, Supplementary Material online) genome-wide and inside the INV (supplementary fig. S9, Supplementary Material online), although the observed difference between species is not significant and not as large as previously observed (1% here, 11% in (Hill and Betancourt 2018), 5% in (Drosophila 12 Genomes et al. 2007)) probably due to our significantly better *D. persimilis* assembly. Interestingly, even though TE content is slightly higher in *D. persimilis*, RepeatMasker annotations show a higher proportion of non-TE repetitive elements such as satellites and simple repeats *in D. pseudoobscura* (supplementary table S4, Supplementary Material online). Little is known about the evolution of satellite DNA in these species, but previous work indicates that rapid turnovers of satellite DNA are caused mainly by gains rather than losses (Wei et al. 2018).

Previous work has shown that the frequency of TE insertions often correlates with overall TE activity (Hill and Betancourt 2018; Pasquesi et al 2020; Liu et al. 2021), a pattern also observed in our data in agreement with predictions from a stochastic model of genome-wide transcription (Encode Project Consortium 2012) Both species exhibit differences not only in the proportion of TE families but also in the expression of two key players involved in TE suppression pathways (piRNA and siRNA): *Rhi* and *Dcr*-2. These results are consistent with the idea that because TE family expansions and turnovers can happen very rapidly, efficient silencing of the most abundant TE families in a genome can generate a disconnect between genome abundance and levels of expression (Kofler et al. 2012, 2015; Yang and Xi 2017; Ozata et al. 2019), and can also explain heterogeneous TE family abundances across the Drosophila phylogeny (Hill and Betancourt 2018; Wei et al. 2018).

D. pseudoobscura and *D. persimilis* show a strong negative correlation between recombination rate and TE content (supplementary fig. S7, Supplementary Material online). This result is consistent with the idea that recombination suppression can promote the accumulation of TEs due to a reduction in the efficiency of selection to remove slightly deleterious TEs and SVs generated by TE activity (Dolgin and Charlesworth 2008). Interestingly, we observed significant increases in the proportion of TEs at the inversion breakpoint regions from chromosomes 2 and XL (Fig. 6), as well as a reduction of local recombination rates on those genomic regions (supplementary figs. S12 and S13, Supplementary Material online). Because the chromosomal rearrangements only occurred in *D. persimilis*, these results imply that TEs were already abundant at those genomic

locations in the ancestor of both species. The local increase in TEs at breakpoint regions probably favored the formation of the rearrangements in the ancestor of *D. persimilis* and further reduction of recombination rates may have favored the accumulation of more TEs. The latter scenario is predicted by different models proposed to explain the establishment of inversions across populations that posit the repressing effect of inversions on the local recombination rate (Kirkpatrick and Barton 2006; Feder et al. 2011; Huang and Rieseberg 2020). Once inversions arise, they are usually in heterozygotes and those individuals experience a reduction in recombination that can facilitate the accumulation of deleterious alleles (Dolgin and Charlesworth 2008; Charlesworth and Barton 2018).

Methods

Sequencing

We sequenced one inbred line of *Drosophila pseudoobscura* (*Dpse\wild-type ST, National Drosophila stock center #14011-0121.41, collected in Mather CA*) and of *D. persimilis* (Mather 40, *collected in Mather* CA) (Machado et al. 2002). These strains are different than those used in the original genome projects for those species (Richards et al. 2005; Drosophila 12 Genomes et al. 2007). High molecular weight DNA from a mix of males and females was extracted using the Blood and Cell culture DNA Midi Kit for DNA extraction (Qiagen) following a previously described protocol (Chakraborty et al. 2016). DNA was then sent to Pacific Biosciences to perform SMRT sequencing using the Sequel system. Sequencing coverage for *D. pseudoobscura* ST and *D. persimilis* Mather 40 was 114X and 72X, respectively. PacBio sequences were deposited in NCBI's SRA database: PRJNA753500 (*D. pseudoobscura*) and PRJNA753501 (*D.* persimilis). Short-read sequences (Illumina, 150 PE) were obtained from male DNA and sequenced at the University of Maryland Genome core facility (IBBR). Short-read sequences were deposited in NCBI's SRA database (*Dpse\wild-type ST* SRA Accession PRJNA753500, *D. persimilis* Mather 40 SRA Accession SAMN16555934 (Korunes et al. 2021)).

Genome size for both species was estimated with a k-mer approach using Illumina short reads. The k-mer abundance spectrum ($k = 21$) was generated using jellyfish v2.2.8 (Marcais and Kingsford 2011) and genome size was estimated using GenomeScope v1.0 (Vurture et al. 2017).

Genome Assembly

PacBio long reads were used to generate de novo genome assemblies using HGAP4-Arrow with default parameters (supplementary table S8, Supplementary Material online). Default parameters of PbJelly (PBSuite v15.8.24) (English et al. 2012) and Pilon v1.22 (Walker et al. 2014) were used later to fill assembly gaps and to polish the final gap-filled contigs using both PacBio long reads and Illumina short reads. A hybrid assembly for both species was also generated by combining long and short DNA reads using DBG2OLC (Ye et al. 2016). DBG2OLC combines both De Brujin graphs and Overlap-Layout-Consensus approaches. Briefly, SparseAssembler was used to generate an initial assembly of the short reads into short but accurate contigs using default parameters. Those fragmented but accurate assemblies were used by DBG2OLC to find overlaps with the PacBio long reads (supplementary table S8, Supplementary Material online).

As PacBio-only assemblies can be improved with the incorporation of a hybrid assembly, both the de novo and the hybrid assemblies were merged to perform a final round of scaffolding using quickmerge v0.2 (Chakraborty et al. 2016), which finds highly homologous overlaps between the contigs from the hybrid and PacBio-only assemblies. After the merging step, a final round of gap-filling with PacBio long reads was performed using PbJelly. Redundant contigs were removed from each assembly based on nucmer alignments using custom bash scripts. Chimeric contigs containing mitochondrial and yeast genomes were also removed from the final genome assemblies. A full list of commands used for PbJelly, pilon, DBG2OLC, and quickmerge can be found at https://github.com/javibio-git/SV_analysis_for_Dpse_and_Dper. Final genome assemblies were deposited to NCBI's genome database: JAIUWF000000000 (*D. pseudoobscura*) and JAIUWG000000000 (*D. persimilis*).

Contig Orientation and Assembly Comparisons

Final genome assemblies for both species were aligned to the *D. pseudoobscura* reference genome assembly (Flybase v3.2) using nucmer ver 3.1 (Kurtz et al. 2004) to orientate and assign contigs to chromosomes. Additionally, more recent genome assemblies (Miller et al. 2018; Bracewell et al. 2019; Liao et al. 2021) were used to confirm contig orientation, contiguity, and chromosome assignment of contigs (supplementary fig. S13, Supplementary Material online). Genetic markers from (Schaeffer et al. 2008) and (Bracewell et al. 2019) were used to validate the order of both species assemblies and to confirm centromere regions for all chromosomes.

Repeat Annotation

De novo transposable element identification was performed using RepeatModeler v1.0.11 (Smit and Hubley 2008–2015). Subsequently, a full repeat element annotation was performed using RepeatMasker v4.0.9 (Smit et al. 2013–2015) using the drosophila library from RepBase in 2017 (Bao et al. 2015). Annotations from RepeatModeler and RepeatMasking were merged to generate the final repeat annotation gff3 input file used in the genome annotation.

Genome Annotation

We used newly collected developmental RNA-seq data for *D. persimilis* (supplementary table S9, Supplementary Material online) and previously published data from *D. pseudoobscura* (Paris et al. 2015; Nyberg and Machado 2016). RNA-seq reads were mapped to the new genome assemblies using Hisat2 v2.1.0 (Kim et al. 2019). These mapped reads were used to build transcriptome assemblies for each sample for both species using StringTie v2.1.1 (Pertea et al. 2015) using Drosophila-optimized parameters (Yang et al. 2018). The assembled transcripts for each sample were then merged using "StringTie merge" with the Drosophila-optimized parameters to get the final transcriptome for each species.

Isoseq RNA sequencing data for *D. pseudoobscura* heads was also generated and used as another source of empirical evidence for gene annotations. Best practices for Isoseq data were implemented to get the final nonredundant isoform sequences using the IsoSeq3 tools [https://github.com/Magdoll/cDNA_Cupcake/wiki/Iso-Seq-Single-Cell-Analysis:-Recommended-Analysis-Guidelines]. In brief, circular consensus sequencing reads were generated using the css command (–skip-polish –minPasses 1) and primers were removed using lima (–isoseq –no-pbi). Isoseq3 refine, cluster, and polish were used with default parameters to generate the final subreads (https://github.com/PacificBiosciences/pbbioconda). Subreads were mapped to the new *D. pseudoobscura* genome using minimap2 (Li 2018) and final collapsed transcripts were retrieved using the tama_collapse.py script from https://github.com/GenomeRIK/tama/wiki/Tama-Collapse (Kuo et al. 2017). Final transcriptomes were used as additional EST evidence during the initial genome annotation. Protein sequences for *D. pseudoobscura* ver. 3.2 and *D. melanogaster* (r6.37) were downloaded from FB and used as protein homology evidence.

We used the MAKER pipeline (Cantarel et al. 2008) for the basic genome annotation. The initial MAKER run created gene models based only on empirical evidence coming from de novo assembled ESTs and protein sequences ('est2genome = 1", "protein2genome = 1"). For all subsequent MAKER runs, other parameters were modified as follows: "pred_flank = 2000", "alt_splice = 1", "split_hit = 30,000", "min_intron = 20" (Venturini et al. 2018). SNAP v2006-07-28 (Korf 2004) and Augustus v3.3.3 (Stanke et al. 2006) ab initio gene predictors were trained based on the gene annotations created from the empirical evidence for both species (AED > 0.5, amino acid length > 50). Augustus training was conducted by using BUSCO v3.1.0 (Simao et al. 2015; Seppey et al. 2019). (insectadb, -m genome, -long) for genomic regions that contained mRNA annotations generated from the empirical annotation. A second round of annotation with MAKER was conducted to create a new set of gene models predicted by SNAP and Augustus (est2genome = 0, protein2genome = 0). Finally, one more round of annotation was run to improve previous annotated gene models.

In addition to the MAKER annotations, annotations from *D. pseudoobscura* from FB, from more recent improved annotations (Yang et al 2018) and lncRNA annotations (Nyberg and Machado 2016), were also transferred to the two new genome assemblies using liftOver (Kent et al. 2002) implemented in the flo pipeline (Pracana et al. 2017). Transferred annotations and MAKER annotations were then compared using gffcompare v0.11.6 (Pertea and Pertea 2020). Only transferred annotations that did not overlap with MAKER annotations were considered new relative to the MAKER annotations.

Consensus Genome Annotation

Annotations from different data sources can lever a noisy annotation dataset simply because of subtle differences in the annotation algorithms. To create a final annotation dataset, Mikado v1.2.4 (Venturini et al. 2018) was implemented using three different annotation sources: transcriptome assembly, MAKER annotations, and FB liftovers. As annotations for ncRNAs from (Nyberg and Machado 2016) were not included on the Mikado runs, those were merged later using GffCompare (Pertea and Pertea 2020) and custom bash scripts. Final genome annotations for both species were formatted using the packages AGAT v0.2.3 (https://github.com/NBISweden/AGAT) and GenomeTools v1.6.1 (Gremme et al. 2013). Potentially spurious annotations (genes < 100 bp) were removed from the final consensus annotation. The full postprocessing protocol is available at: https://github.com/javibio-git/SV_analysis_for_Dpse_and_Dper.

Gene Orthology and Collinearity

Gene synteny analysis was conducted to estimate the degree of gene collinearity between *D. pseudoobscura* and *D. persimilis* and two more species of the *Drosophila pseudoobscura* subgroup: *D. loweii* and *D. miranda*. Genome assemblies and annotations for *D. loweii* and *D. miranda* were retrieved from (Mahajan et al. 2018) and (Bracewell et al. 2019) respectively, and gene collinearity was determined using CLfinder-OrthNet (Oh and Dassanayake 2019). Briefly, CLfinder-OrthNet establish collinearity based on the number of genes that exist on the same order across all genomes of interest. Because CLfinder-OrthNet construct groups of local collinear genes, it is suitable to determine how gene collinearity has been maintained inside INV across the *pseudoobscura subgroup*. Parameters and dependencies used to run CLfinder-OrthNet were the same as in (https://github.com/ohdongha/OrthNet#1-obtaining-one-representative-gene-model-per-locus).

Gene orthology between species of the *pseudoobscura subgroup* was established using OrthoFinder (Emms and Kelly 2019) with default parameters. For these analyses, all protein-coding genes were considered, including annotated genes without start codons. A second run of OrthoFinder was performed including the D. melanogaster reference genome from FB to transfer putative functional gene annotation using GO terms. The results from CLfinder-OrthNet and OrthoFinder analyses were used to generate a collinearity figure (Fig. 2) with the package genoPlotR v0.8.11 (Guy et al. 2010).

Structural Variant Calling

Genome assemblies and PacBio long-reads were used to call and quantify the number of INDELs and CNVs between *D. pseudoobscura* and *D. persimilis* using svim v1.4.2 (Heller and Vingron 2019). Reciprocal svim callings were conducted using the 2 species as a reference. Two additional svim callings using reads and genome assemblies of the same species were used as a control to correct for potential false positives produced by assembly errors. Remaining variants from each reciprocal calling were filtered again based on the svim score (>10). Filtered svim variants were then cross-validated using the two reciprocal callings. The error-correction and validation steps were conducted using "bedtools intersect" and custom perl and bash scripts.

Variants were also called with svmu v0.4-alpha (Chakraborty et al. 2019) and paftools.js from minimap2 (Li 2018) using whole-genome alignments (Chakraborty et al. 2018). Svim, svmu, and minimap2 variants were merged to obtain the final set of INDELs and CNVs for downstream analyses.

Final validated variants between *D. pseudoobscura* and *D. persimilis* were polarized using the *D. miranda* reference genome (Mahajan et al. 2018). Variants with D. miranda were called using paftools.js from minimap2 and the polarization step was conducted using bedtools intersect and custom perl scripts (https://github.com/javibio-git/SV_analysis_for_Dpse_and_Dper).

Inversion Breakpoint Validation

Previously identified fixed chromosomal inversions between *D. persimilis* and *D. pseudoobscura* were confirmed in the new genome assemblies using nucmer genome alignments chromosomes 2, 3, XL, and XR (supplementary fig. S13, Supplementary Material online) and validated inversion breakpoint regions using reciprocal mapping of CLR reads for chromosomes 2 and XL (supplementary figs. S13 to S25, Supplementary Material online). In addition, we further validated inversion breakpoints (Machado et al. 2007) for those major rearrangements using SyRI v1.3 (Goel et al. 2019). Full command lines are shown at https://github.com/javibio-git/SV_analysis_for_Dpse_and_Dper.

Recombination Landscapes

DNA-seq data from 37 and 20 populations of *D. pseudoobscura* and *D. persimilis* (supplementary table S10, Supplementary Material online), respectively, were used to estimate the number of recombination events implementing a nonoverlapping 50 kb sliding-windows approach (Chan et al. 2012). Trimmed raw reads of each line were mapped to their corresponding genome assembly using bwa v 0.7.17-r1188 (Li and Durbin 2009) and resulting bam files were sorted using samtools v1.7 (Li et al. 2009). We used GATK v4.2.0.0 to call single nucleotide polymorphisms (SNPs) according to GATK Best Practices recommendations (DePristo et al. 2011; Van der Auwera and O'Connor 2020). Filtered bi-allelic SNPs were used to estimate the mean number of crossover events per generation (*r*/bp) using LDhelmet v1.9 (Chan et al. 2012).

Association Tests

Spatial correlation analyses between SVs and annotated genes were conducted using the GenometriCorr v1.1.24 package in R, using 1,000 permutations. Full mRNA annotation for each annotated gene was taken as the input set for the permutation analysis. RepeatMasker annotations (.align file) were parsed and formatted using the parseRM.pl script (https://github.com/4ureliek/Parsing-RepeatMasker-Outputs) for both *D. pseudoobscura and D. persimilis*. The resulting parsed bed file was the input for the TE-analysis_Shuffle_bed.pl script (https://github.com/4ureliek/TEanalysis), which was used to test for significant associations between the most abundant TE families, SVs and gene regions using 1,000 permutations. The TE-analysis_pipeline.pl v4.6 script (https://github.com/4ureliek/TEanalysis) was used to characterize TE content in gene regions 10 kb upstream of the transcript start site, exons, and introns. Intergenic regions less than 10 kb were also included, but excluding regions with overlapped annotations.

Correspondence analysis was conducted using the corrplot package v0.90 (Wei and Simko 2021) in R, to determine significant associations of SVs versus gene regions and SVs versus DE genes. Odds ratios of the most abundant TEs and INS associated with gene regions were assessed by counting the proportion of bp overlapping with annotated insertions in both species.

Global TE Expression Analyses

RNA-seq data from four developmental stages: first and third instar larvae, pupae, and adults for *D. pseudoobscura* (MV225 line) and *D. persimilis* (M40) were used to measure gene and global TE expression differences between the two species; males and females combined. Alignments were conducted using our *D. pseudoobscura* genome assembly as a reference. Each developmental stage was analyzed independently using the best practices for TETranscripts

v2.2.1 (Jin et al. 2015). Briefly, alignments were made using STAR v2.7.6a (Dobin et al. 2013) and the resulting bam files were the input for TETranscripts to measure gene and TE differential expression. A reciprocal analysis was performed using the *D. persimilis* genome assembly to account for alignment biases. Normalized counts were pooled for each TE family to measure global TE expression across developmental stages in the 2 species. Differential expression of TE families between species was conducted for each developmental stage independently using DESeq2 v1.30.1 (Love et al. 2014).

Differential Gene Expression and SV Variant Associations

The same developmental RNA-seq data was used to assess differential gene expression between both species using salmon v1.5.2 (Patro et al. 2017). Expression quantification was conducted using the corresponding transcript and read sequences for each species independently. Subsequently, only expression data for 1:1 orthologs was used to test for significant differential expression using DESeq2. The 0.05 quantile of the distribution of *P*-values was set up as a hard threshold to establish significant expression for each developmental stage. TE annotations and SVs overlapping either exons, introns, or 10 kb upstream regions were counted using custom perl scripts. Significant association between SVs and DE genes was assessed using custom scripts in R.

GO Enrichment Analyses

GO enrichment analysis was conducted using GOrilla web tool (Eden et al. 2007, 2009). GO terms associated with all the genes considered for differential expression were used as the main background list.

Figures

All figures were generated with ggplot2 (Wickham 2016) under R v4.0.3 (R Core Team 2021) and circos v0.69-9 (Krzywinski et al. 2009).

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Thomas Kocher and Phillip Johnson for helpful discussions, technical advice and comments on the manuscript. We are grateful to Sarah Kingan (PacBio) for support during sequencing and initial assembly of long read data, and to Suwei Zhao (UMD IBBR) for library preparation and Illumina sequencing. Two anonymous reviewers provided useful suggestions that improved the manuscript.

Funding

J.C.-P. was partially supported by CONACYT CVU 563701 and the COMBINE program at the University of Maryland (National Science Foundation award number 1632976). Research supported by National Science Foundation grants MCB-1716532 and DEB-1754572 to C.A.M.

Data Availability

The new genome assemblies reported in this article are available in NCBI's genome database under accessions JAIUWF000000000 (*D. pseudoobscura*) and JAIUWG000000000 (*D. persimilis*). PacBio raw sequences were deposited in NCBI's SRA database: PRJNA753500 (*D. pseudoobscura*) and PRJNA753501(*D.* persimilis). Raw Illumina DNA and RNA sequence data will be available on NCBI's SRA database by the time of publication. A full list of commands and scripts used in the analyses can be found at https://github.com/javibio-git/SV_analysis_for_Dpse_and_Dper.

Literature Cited

Allen SL, Delaney EK, Kopp A, Chenoweth SF. Single-Molecule sequencing of the Drosophila serrata genome. G3 (Bethesda). 2017:7(3):781–788. https://doi.org/10.1534/g3.116.037598.

Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D, et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell. 2020:182(1):145–161 e123. https://doi.org/10.1016/j.cell.2020.05.021.

Anderson WW, Ayala FJ, Michod RE. Chromosomal and allozymic diagnosis of three species of Drosophila. *Drosophila pseudoobscura, D. persimilis, and D. miranda*. J Hered. 1977:68(2):71–74. https://doi.org/10.1093/oxfordjournals.jhered.a108793.

Avilés-Pagán EE, Kang ASW, Orr-Weaver TL. Identification of new regulators of the oocyte-to-embryo transition in Drosophila. G3 (Bethesda). 2020:10(9):2989–2998. https://doi.org/10.1534/g3.120.401415.

Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015:6:11. https://doi.org/10.1186/s13100-015-0041-9.

Bracewell R, Chatla K, Nalley MJ, Bachtrog D. Dynamic turnover of centromeres drives karyotype evolution in Drosophila. Elife. 2019:8:e49002. https://doi.org/10.7554/eLife.49002.

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell. 2007:128(6):1089–1103. https://doi.org/10.1016/j.cell.2007.01.043.

Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, Yandell M. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008:18(1):188–196. https://doi.org/10.1101/gr.6743907.

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

Chakraborty M, Emerson JJ, Macdonald SJ, Long AD. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. Nat Commun. 2019:10(1):4872. https://doi.org/10.1038/s41467-019-12884-1.

Chakraborty M, VanKuren NW, Zhao R, Zhang X, Kalsow S, Emerson JJ. Hidden genetic variation shapes the structure of functional elements in Drosophila. Nat Genet. 2018:50(1):20–25. https://doi.org/10.1038/s41588-017-0010-y.

Chan AH, Jenkins PA, Song YS. Genome-wide fine-scale recombination rate variation in Drosophila melanogaster. PLoS Genet. 2012:8(12):e1003090. https://doi.org/10.1371/journal.pgen.1003090.

Charlesworth B, Barton NH. The spread of an inversion with migration and selection. Genetics. 2018:208(1):377–382. https://doi.org/10.1534/genetics.117.300426.

Chen WY, Luan XJ, Yan YD, Wang M, Zheng QW, Chen X, Yu J, FangJ. CG8005 mediates transit-amplifying spermatogonial divisions via oxidative stress in Drosophila testes. Oxid Med Cell Longev. 2020:2020:2846727. https://doi.org/10.1155/2020/2846727.

Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L, Montgomery SB, et al. The impact of structural variation on human gene expression. Nat Genet. 2017:49(5):692–699. https://doi.org/10.1038/ng.3834.

Choi JY, Lee YCG. Double-edged sword: the evolutionary consequences of the epigenetic silencing of transposable elements. PLoS Genet. 2020:16(7):e1008872. https://doi.org/10.1371/journal.pgen.1008872.

Chung WJ, Okamura K, Martin R, Lai EC. Endogenous RNA interference provides a somatic defense against Drosophila transposons. Curr Biol. 2008:18(11):795–802. https://doi.org/10.1016/j.cub.2008.05.006.

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

Darbellay F, Necsulea A. Comparative transcriptomics analyses across Species, organs, and developmental stages reveal functionally constrained lncRNAs. Mol Biol Evol. 2020:37(1):240–259. https://doi.org/10.1093/molbev/msz212.

Delprat A, Negre B, Puig M, Ruiz A. The transposon galileo generates natural chromosomal inversions in Drosophila by ectopic recombination. PLoS One. 2009:4(11):e7883. https://doi.org/10.1371/journal.pone.0007883.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011:43(5):491–498. https://doi.org/10.1038/ng.806.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-Seq aligner. Bioinformatics. 2013:29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635.

Dobzhansky T. Chromosomal races in *Drosophila pseudoobscura* and *Drosophila persimilis*. In: Dobzhansky T, Epling C, editors. Contributions to the genetics, taxonomy, and ecology of Drosophila pseudoobscura and its relatives. Washington (DC): Carnegie Institute of Washington; 1944. 554:p. 47–144.

Dobzhansky T, Epling T. Contributions to the genetics, taxonomy, and ecology of Drosophila pseudoobscura and its relatives. Washington (DC): Carnegie Institute of Washington; 1944.

Dolgin ES, Charlesworth B. The effects of recombination rate on the distribution and abundance of transposable elements. Genetics. 2008:178(4):2169–2177. https://doi.org/10.1534/genetics.107.082743.

Eden E, Lipson D, Yogev S, Yakhini Z. Discovering motifs in ranked lists of DNA sequences. PLoS Comput Biol. 2007:3(3):e39. https://doi.org/10.1371/journal.pcbi.0030039.

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009:10(1):48. https://doi.org/10.1186/1471-2105-10-48.

Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019:20(1):238. https://doi.org/10.1186/s13059-019-1832-y.

Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012:489(7414):57–74. https://doi.org/10.1038/nature11247.

English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One. 2012:7(11):e47768. https://doi.org/10.1371/journal.pone.0047768.

Feder JL, Gejji R, Powell THQ, Nosil P. Adaptive chromosomal divergence driven by mixed geographic mode of evolution. Evolution. 2011:65(8):2157–2170. https://doi.org/10.1111/j.1558-5646.2011.01321.x.

Fuller ZL, Haynes GD, Richards S, Schaeffer SW. Genomics of natural populations: how differentially expressed genes shape the evolution of chromosomal inversions in Drosophila pseudoobscura. Genetics. 2016:204(1):287–301. https://doi.org/10.1534/genetics.116.191429.

Fuller ZL, Koury SA, Phadnis N, Schaeffer SW. How chromosomal rearrangements shape adaptation and speciation: case studies in *Drosophila pseudoobscura* and its sibling species *Drosophila persimilis*. Mol Ecol. 2019:28(6):1283–1301. https://doi.org/10.1111/mec.14923.

Fuller ZL, Leonard CJ, Young RE, Schaeffer SW, Phadnis N. Ancestral polymorphisms explain the role o chromosomal inversions in speciation. PLoS Genet. 2018:14(7):e1007526. https://doi.org/10.1371/journal.pgen.1007526.

Galiana-Arnoux D, Dostert C, Schneemann A, Hoffmann JA, Imler JL. Essential function in vivo for dicer-2 in host defense against RNA viruses in Drosophila. Nat Immunol. 2006:7(6):590–597. https://doi.org/10.1038/ni1335.

Gebert D, Neubert LK, Lloyd C, Gui J, Lehmann R, Teixeira FK. Large Drosophila germline piRNA clusters are evolutionarily labile and dispensable for transposon regulation. Mol Cell. 2021:81(19):3965–3978 e3965. https://doi.org/10.1016/j.molcel.2021.07.011.

Gerdol M, Moreira R, Cruz F, Gómez-Garrido J, Vlasova A, Rosani U, Venier P, Naranjo-Ortiz MA, Murgarella M, Greco S, et al. Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. Genome Biol. 2020:21(1):275. https://doi.org/10.1186/s13059-020-02180-3.

Goel M, Sun H, Jiao WB, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. 2019:20(1):277. https://doi.org/10.1186/s13059-019-1911-0.

Gremme G, Steinbiss S, Kurtz S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Trans Comput Biol Bioinform. 2013:10:645–656. https://doi.org/10.1109/TCBB.2013.68.

Guy L, Roat Kultima J, Andersson SGE. genoPlotr: comparative gene and genome visualization in R. Bioinformatics. 2010:26(18):2334–2335. https://doi.org/10.1093/bioinformatics/btq413.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nat Rev Genet. 2009:10(8):551–564. https://doi.org/10.1038/nrg2593.

Heller D, Vingron M. SVIM: structural variant identification using mapped long reads. Bioinformatics. 2019:35:2907–2915. https://doi.org/10.1093/bioinformatics/btz041.

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

Hoffmann AA, Rieseberg LH. Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? Annu Rev Ecol Evol Syst. 2008: 39(1):21–42. https://doi.org/10.1146/annurev.ecolsys.39.110707.173532.

Hon T, Mars K, Young G, Tsai YC, Karalius JW, Landolin JM, Maurer N, Kudrna D, Hardigan MA, Steiner CC, et al. Highly accurate long-read HiFi sequencing data for five complex genomes. Sci Data. 2020:7(1):399. https://doi.org/10.1038/s41597-020-00743-4.

Huang KC, Rieseberg LH. Frequency, origins, and evolutionary role of chromosomal inversions in plants. Front Plant Sci. 2020:11:296. https://doi.org/10.3389/fpls.2020.00296.

Huang Y, Shukla H, Lee YCG. Species-specific chromatin landscape determines how transposable elements shape genome evolution. Elife. 2022:11:e81567. https://doi.org/10.7554/eLife.81567.

Hufford MB, Seetharam AS, Woodhouse MR, Chougule KM, Ou S, Liu J, Ricci WA, Guo T, Olson A, Qiu Y, et al. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science. 2021:373(6555):655–662. https://doi.org/10.1126/science.abg5289.

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018:36(4):338–345. https://doi.org/10.1038/nbt.4060.

Jin Y, Tam OH, Paniagua E, Hammell M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-Seq datasets. Bioinformatics. 2015:31(22):3593–3599. https://doi.org/10.1093/bioinformatics/btv422.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 2012:484(7392):55–61. https://doi.org/10.1038/nature10944.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002:12:996–1006.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019:37(8):907–915. https://doi.org/10.1038/s41587-019-0201-4.

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

Kirkpatrick M, Barton N. Chromosome inversions, local adaptation and speciation. Genetics. 2006:173(1):419–434. https://doi.org/10.1534/genetics.105.047985.

Kofler R, Betancourt AJ, Schlötterer C. Sequencing of pooled DNA samples (pool-seq) uncovers Complex dynamics of transposable element insertions in Drosophila melanogaster. PLoS Genet. 2012:8(1):e1002487. https://doi.org/10.1371/journal.pgen.1002487.

Kofler R, Nolte V, Schlotterer C. Tempo and mode of transposable element activity in Drosophila. PLoS Genet. 2015:11(7): e1005406. https://doi.org/10.1371/journal.pgen.1005406.

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

Korunes KL, Machado CA, Noor MAF. Inversions shape the divergence of Drosophila pseudoobscura and Drosophila persimilis on multiple timescales. Evolution. 2021:75:1820–1834. https://doi.org/10.1111/evo.14278.

Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJP, Dougherty ML, et al. High-resolution comparative analysis of great ape genomes. Science. 2018:360(6393):eaar6343. https://doi.org/10.1126/science.aar6343.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome Res. 2009:19(9):1639–1645. https://doi.org/10.1101/gr.092759.109.

Kulathinal RJ, Stevison LS, Noor MAF. The genomics of speciation in Drosophila: diversity, divergence, and introgression estimated using low-coverage genome sequencing. PLoS Genet. 2009:5(7): e1000550. https://doi.org/10.1371/journal.pgen.1000550.

Kuo RI, Tseng E, Eory L, Paton IR, Archibald AL, Burt DW. Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. BMC Genomics. 2017:18(1):323. https://doi.org/10.1186/s12864-017-3691-9.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. Genome Biol. 2004:5(2):R12. https://doi.org/10.1186/gb-2004-5-2-r12.

Lee YS, Nakahara K, Pham JW, Kim K, He Z, Sontheimer EJ, Carthew RW. Distinct roles for Drosophila dicer-1 and dicer-2 in the siRNA/miRNA silencing pathways. Cell. 2004:117(1):69–81. https://doi.org/10.1016/S0092-8674(04)00261-2.

Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018:34(18):3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009:25(14):1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. Bioinformatics. 2009:25(16):2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Liao Y, Zhang XW, Chakraborty M, Emerson JJ. Topologically associating domains and their role in the evolution of genome structure and function in Drosophila. Genome Res. 2021:31(3):397–410. https://doi.org/10.1101/gr.266130.120.

Liu Z, Zhao H, Yan Y, Wei MX, Zheng YC, Yue EK, Alam MS, Smartt KO, Duan MH, Xu JH. Extensively current activity of transposable elements in natural rice accessions revealed by singleton insertions. Front Plant Sci. 2021:12:745526. https://doi.org/10.3389/fpls.2021.745526

Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Nat Rev Genet. 2020:21(10): 597–614. https://doi.org/10.1038/s41576-020-0236-x.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biol. 2014:15(12):550. https://doi.org/10.1186/s13059-014-0550-8.

Machado CA, Haselkorn TS, Noor MAF. Evaluation of the genomic extent of effects of fixed inversion differences on intraspecific variation and interspecific gene flow in Drosophila pseudoobscura and D. persimilis. Genetics. 2007:175(3):1289–1306. https://doi.org/10.1534/genetics.106.064758.

Machado CA, Kliman RM, Markert JA, Hey J. Inferring the history of speciation from multilocus DNA sequence data: the case of Drosophila pseudoobscura and close relatives. Mol Biol Evol. 2002:19(4):472–488. https://doi.org/10.1093/oxfordjournals.molbev.a004103.

Mahajan S, Wei KHC, Nalley MJ, Gibilisco L, Bachtrog D. De novo assembly of a young Drosophila Y chromosome using single-molecule sequencing and chromatin conformation capture. PLoS

Biol. 2018:16(7):e2006348. https://doi.org/10.1371/journal.pbio.2006348.
Marcais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011:27(6):764–770. https://doi.org/10.1093/bioinformatics/btr011.
Mérel V, Boulesteix M, Fablet M, Vieira C. Transposable elements in Drosophila. Mob DNA. 2020:11(1):23. https://doi.org/10.1186/s13100-020-00213-z.
Miller DE, Staber C, Zeitlinger J, Hawley RS. Highly contiguous genome assemblies of 15 Drosophila Species generated using nanopore sequencing. G3 (Bethesda). 2018:8(10):3131–3141. https://doi.org/10.1534/g3.118.200160.
Naganuma M, Tadakuma H, Tomari Y. Single-molecule analysis of processive double-stranded RNA cleavage by Drosophila dicer-2. Nat Commun. 2021:12(1):4268. https://doi.org/10.1038/s41467-021-24555-1.
Noor MAF, Garfield DA, Schaeffer SW, Machado CA. Divergence between the Drosophila pseudoobscura and D-persimilis genome sequences in relation to chromosomal inversions. Genetics. 2007:177(3):1417–1428. https://doi.org/10.1534/genetics.107.070672.
Noor MAF, Grams KL, Bertucci LA, Almendarez Y, Reiland J, Smith KR. The genetics of reproductive isolation and the potential for gene exchange between *Drosophila pseudoobscura* and *D. persimilis* via backcross hybrid males. Evolution. 2001b:55(3):512–521. https://doi.org/10.1554/0014-3820(2001)055[0512:TGORIA]2.0.CO;2.
Noor MA, Grams KL, Bertucci LA, Reiland J. Chromosomal inversions and the reproductive isolation of species. Proc Natl Acad Sci USA. 2001a:98(21):12084–12088. https://doi.org/10.1073/pnas.221274498.
Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. The complete sequence of a human genome. Science. 2022:376(6588):44–53. https://doi.org/10.1126/science.abj6987.
Nyberg KG, Machado CA. Comparative expression dynamics of intergenic long noncoding RNAs in the genus Drosophila. Genome Biol Evol. 2016:8(6):1839–1858. https://doi.org/10.1093/gbe/evw116.
Oh DH, Dassanayake M. Landscape of gene transposition-duplication within the Brassicaceae family. DNA Res. 2019:26(1):21–36. https://doi.org/10.1093/dnares/dsy035.
O'Neill K, Brocks D, Hammell MG. Mobile genomics: tools and techniques for tackling transposons. Philos Trans R Soc Lond B Biol Sci. 2020:375(1795):20190345. https://doi.org/10.1098/rstb.2019.0345.
Orr HA. Genetics of male and female sterility in hybrids of Drosophila pseudoobscura and D. persimilis. Genetics. 1987:116(4):555–563. https://doi.org/10.1093/genetics/116.4.555.
Ozata DM, Gainetdinov I, Zoch A, O'Carroll D, Zamore PD. PIWI-interacting RNAs: small RNAs with big functions. Nat Rev Genet. 2019:20(2):89–108. https://doi.org/10.1038/s41576-018-0073-3.
Paris M, Villalta JE, Eisen MB, Lott SE. Sex bias and maternal contribution to gene expression divergence in Drosophila Blastoderm embryos. PLoS Genet. 2015:11(10):e1005592. https://doi.org/10.1371/journal.pgen.1005592.
Pasquesi GIM, Perry BW, Vandewege MW, Ruggiero RP, Schield DR, Castoe TA. Vertebrate lineages exhibit diverse patterns of transposable element regulation and expression across tissues. Genome Biol Evol. 2020:12(5):506–521. https://doi.org/10.1093/gbe/evaa068.
Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017:14(4):417–419. https://doi.org/10.1038/nmeth.4197.
Pertea G, Pertea M. GFF utilities: GffRead and GffCompare. F1000Res. 2020:9:304. https://doi.org/10.12688/f1000research.23297.1.
Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-Seq reads. Nat Biotechnol. 2015:33(3):290–295. https://doi.org/10.1038/nbt.3122.
Petersen M, Armisén D, Gibbs RA, Hering L, Khila A, Mayer G, Richards S, Niehuis O, Misof B. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. BMC Evol Biol. 2019:19(1):11. https://doi.org/10.1186/s12862-018-1324-9.
Policansky D, Zouros E. Gene differences between sex-ratio and standard gene arrangements of X-chromosome in Drosophila-Persimilis. Genetics. 1977:85(3):507–511. https://doi.org/10.1093/genetics/85.3.507.
Ponjavic J, Ponting CP, Lunter G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res. 2007:17(5):556–565. https://doi.org/10.1101/gr.6036807.
Pracana R, Priyam A, Levantis I, Nichols RA, Wurm Y. The fire ant social chromosome supergene variant Sb shows low diversity but high divergence from SB. Mol Ecol. 2017:26:2864–2879.
R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2021.
Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021:592(7856):737–746. https://doi.org/10.1038/s41586-021-03451-0.
Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, Nielsen R, Thornton K, Hubisz MJ, Chen R, Meisel RP, et al. Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. Genome Res. 2005:15(1):1–18. https://doi.org/10.1101/gr.3059305.
Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O'Grady PM, Rohde C, Valente VL, Aguade M, Anderson WW, et al. Polytene chromosomal maps of 11 Drosophila species: the order of genomic scaffolds inferred from genetic and physical maps. Genetics. 2008:179(3):1601–1655. https://doi.org/10.1534/genetics.107.086074.
Scully R, Panday A, Elango R, Willis NA. DNA double-strand break repair-pathway choice in somatic mammalian cells. Nat Rev Mol Cell Bio. 2019:20(11):698–714. https://doi.org/10.1038/s41580-019-0152-0.
Seppey M, Manni M, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness. Methods Mol Biol. 2019:1962:227–245.
Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, et al. Nanopore sequencing and the shasta toolkit enable efficient de novo assembly of eleven human genomes. Nat Biotechnol. 2020:38(9):1044–1053. https://doi.org/10.1038/s41587-020-0503-6.
Sharma SP, Zuo T, Peterson T. Transposon-induced inversions activate gene expression in the maize pericarp. Genetics. 2021:218(2):iyab062. https://doi.org/10.1093/genetics/iyab062.
Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015:31(9):3210–3212. https://doi.org/10.1093/bioinformatics/btv351.
Smit A, Hubley R. RepeatModeler Open-1.0 A. 2008–2015. http://www.repeatmasker.org.
Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. 2013–2015. http://www.repeatmasker.org.

Sridharan V, Heimiller J, Robida MD, Singh R. High throughput sequencing identifies misregulated genes in the Drosophila polypyrimidine tract-binding protein (hephaestus) mutant defective in spermatogenesis. PLoS One. 2016:11(3):e0150768. https://doi.org/10.1371/journal.pone.0150768.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006:34:W435-9. https://doi.org/10.1093/nar/gkl200.

Tan CC. Salivary gland chromosomes in the two races of Drosophila pseudoobscura. Genetics. 1935:20(4):392–402. https://doi.org/10.1093/genetics/20.4.392.

Taylor SA, Larson EL. Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. Nat Ecol Evol. 2019:3(2):170–177. https://doi.org/10.1038/s41559-018-0777-y.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A. 2005:102(39):13950–13955. https://doi.org/10.1073/pnas.0506758102.

Thurmond J, Goodman JL, Strelets VB, Attrill H, Gramates LS, Marygold SJ, Matthews BB, Millburn G, Antonazzo G, Trovisco V, et al. FlyBase 2.0: the next generation. Nucleic Acids Res. 2019:47(D1):D759–D765. https://doi.org/10.1093/nar/gky1003.

Van der Auwera GA, O'Connor DB. Genomics in the cloud: using docker, GATK, and WDL in terra. Sebastopol (CA): O'Reilly Media, Incorporated; 2020.

Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. Gigascience. 2018:7(8). https://doi.org/10.1093/gigascience/giy093.

Vieillard J, Paschaki M, Duteyrat JL, Augière C, Cortier E, Lapart JA, Thomas J, Durand B. Transition zone assembly and its contribution to axoneme formation in Drosophila male germ cells. J Cell Biol. 2016:214(7):875–889. https://doi.org/10.1083/jcb.201603086.

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017:33(14):2202–2204. https://doi.org/10.1093/bioinformatics/btx153.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014:9(11):e112963. https://doi.org/10.1371/journal.pone.0112963.

Weckselblatt B, Rude MK. Human structural variation: mechanisms of chromosome rearrangements. Trends Genet. 2015:31(10):587–599. https://doi.org/10.1016/j.tig.2015.05.010.

Wei KHC, Lower SE, Caldas IV, Sless TJS, Barbash DA, Clark AG. Variable rates of simple satellite gains across the Drosophila phylogeny. Mol Biol Evol. 2018:35(4):925–941. https://doi.org/10.1093/molbev/msy005.

Wei T, Simko V. 2021. R package 'corrplot': Visualization of a Correlation Matrix (Version 0.90). https://github.com/taiyun/corrplot.

Weissensteiner MH, Bunikis I, Catalán A, Francoijs KJ, Knief U, Heim W, Peona V, Pophaly SD, Sedlazeck FJ, Suh A, et al. Discovery and population genomics of structural variation in a songbird genus. Nat Commun. 2020:11(1):3403. https://doi.org/10.1038/s41467-020-17195-4.

Wellenreuther M, Mérot C, Berdan E, Bernatchez L. Going beyond SNPs: the role of structural genomic variants in adaptive evolution and species diversification. Mol Ecol. 2019:28(6):1203–1209. https://doi.org/10.1111/mec.15066.

Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019:37(10):1155–1162. https://doi.org/10.1038/s41587-019-0217-9.

Wickham H. Ggplot2: elegant graphics for data analysis. New York (NY): Springer-Verlag New York; 2016.

Yang HW, Jaime M, Polihronakis M, Kanegawa K, Markow T, Kaneshiro K, Oliver B. Re-annotation of eight Drosophila genomes. Life Sci Alliance. 2018:1(6):e201800156. https://doi.org/10.26508/lsa.201800156.

Yang F, Xi R. Silencing transposable elements in the Drosophila germline. Cell Mol Life Sci. 2017:74(3):435–448. https://doi.org/10.1007/s00018-016-2353-4.

Ye C, Hill CM, Wu S, Ruan J, Ma Z. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. Sci Rep. 2016:6:31900. https://doi.org/10.1038/srep31900.

Zhang J, Peterson T. Transposition of reversed Ac element ends generates chromosome rearrangements in maize. Genetics. 2004:167(4):1929–1937. https://doi.org/10.1534/genetics.103.026229.

Zhang L, Reifová R, Halenková Z, Gompert Z. How important are structural variants for speciation? Genes (Basel). 2021:12(7):1084. https://doi.org/10.3390/genes12071084.

Zichner T, Garfield DA, Rausch T, Stutz AM, Cannavó E, Braun M, Furlong EEM, Korbel JO. Impact of genomic structural variation in Drosophila melanogaster based on population-scale sequencing. Genome Res. 2013:23(3):568–579. https://doi.org/10.1101/gr.142646.112.

Associate editor: Cristina Vieira