ABSTRACT

Title of Dissertation:	THE GENOMICS OF SPECIES DIVERGENCE IN DROSOPHILA
	Javier Carpinteyro Ponce, Doctor of Philosophy, 2023
Dissertation directed by:	Professor Carlos A. Machado Department of biology

How do new species arise and diverge? Has been a fundamental question in evolutionary biology. The process of species divergence can be studied at many different levels of biological organization. However, it is until the recent advancements of genome sequencing technologies that genome-wide signatures of species divergence have started to unveil the complex genomic landscape of speciation. In this dissertation we investigate the landscape of genomic divergence using a classic pair of Drosophila species. We generated four new high quality genome assemblies for *Drosophila pseudoobscura* and *D. persimilis* to explore the genomic differences at three different levels. We first characterized the structural variation landscape between *D. pseudoobscura* and *D. persimilis* and stablished its association with transposable elements and tested how intrinsic genomic factors, such as recombination, influence the accumulation of

structural variation associated with transposable elements in both species. With a combination of high-quality genome assemblies and a comprehensive population genomics data set, we also explored how the contribution of recombination rate and introgression promote sequence divergence with the potential of forming species barriers. Moreover, we investigated how gene co-expression networks potentially rewiring between species contribute to the divergence landscape between *D. pseudoobscura* and *D. persimilis*. Our work highlights the complex landscape of species divergence occurring at multiple levels of organization. Moreover, the integration of potential species drivers identified at different scales shed lights on the molecular mechanisms involved in speciation.

THE GENOMICS OF SPECIES DIVERGENCE IN DROSOPHILA

by

Javier Carpinteyro Ponce

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor Carlos A. Machado, Chair
Professor Eric S. Haag
Professor Najib El-Sayed
Assistant Professor Philip Johnson
Professor Thomas D. Kocher

Dedication

Esta tesis está dedicada a todas las personas que directa o indirectamente dejaron una parte de sí mismas en mi viaje por esta etapa de mi vida. Y aunque me resulta imposible nombrarlas a todas, esta lista incluye a todas mis amistades que, a lo largo de este viaje, fueron parte esencial para la culminación de este doctorado.

Ante todo, quiero dedicar este logro a mis padres: Yadira y Javier. Su apoyo incondicional fue sin duda alguna uno de los motores más grandes que me impulso a la culminación de esta parte de mi camino profesional.

A mi esposa: Karina. Por acompañarme en esta aventura y por estar siempre ahí a pesar de todos los tropiezos.

A mis hermanos: Selene y Ulises. De igual manera su apoyo incondicional fue motor esencial.

A los que se me fueron en vida pero que de igual manera sus enseñanzas son parte de lo que hoy soy como profesional y como persona. A mi abuelita: Leonila, a mi abuelito: Ernesto y a mi padrino: Juan.

Acknowledgements

To Carlos, for giving me the opportunity to work under his wing. I am grateful for his dedicated mentoring and support in all aspects of my academic life. Also grateful for the freedom to explore my own research interests but also for the feedback that helped me to stay focus on my short and long-term goals. Carlos is a remarkable scientist and, as a person, he is honorable and kind. I am grateful for all his life lessons that motivated me during this PhD journey.

I also would like to thank each members of my committee: Dr. Thomas Kocher, Dr. Eric Haag, Dr. Philip Johnson, and Dr. Najib El-Sayed. Thank you for all the helpful discussions and guidance that contributed to the development of this dissertation.

To the many members of the Machado lab for the feedback and helpful discussions throughout the years. And to all the people I met during this journey at the Department of Biology and to the BEESst for all the support. Special mention to Ron Clark for his friendship, for the science discussions and for all the emotional support.

Lastly, I would like to thank the Consejo Nacional de Ciencia y Tecnologia (CVU: 36609) for the financial support given through the foreign fellowships for Mexicans citizens program.

Table of Contents

CHAPTER 1: The complex landscape of structural divergence between the *Drosophila pseudoobscura* and *Drosophila persimilis* genomes

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. Here we present the first high-quality genome assembly of *Drosophila persimilis* and a new high-quality genome for *D. pseudoobscura*. We report a complex and previously hidden landscape of structural divergence and study the relationships among structural variants (SVs), transposable elements (TEs), and gene expression divergence between these two species. 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 TEs accumulate on regions with low levels of recombination, and spatial correlation analyses reveal a strong association between TEs and SVs. We also report a strong association between differentially expressed genes and SVs, and an overrepresentation of differentially expressed genes inside the fixed chromosomal inversions that separate this species pair. Interestingly, species specific SVs are overrepresented in differentially expressed genes involved in neural development, spermatogenesis, and oocyte-to-embryo

transition. Overall, our results highlight the association of TEs with SVs and their importance in driving evolutionary change across species.

INTRODUCTION

The rapid development of sequencing technologies has revolutionized the field of comparative 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 (M. Chakraborty, Baldwin-Brown, Long, & Emerson, 2016; Hon et al., 2020; Jain et al., 2018; Logsdon, Vollger, & Eichler, 2020; Nurk et al., 2022; Shafin et al., 2020; Wenger et al., 2019). 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, Brocks, & Hammell, 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 on how structural variation contributes to phenotypic differences between species (M. Chakraborty et al., 2018; Kronenberg et al., 2018; Logsdon et al., 2020; Nurk et al., 2022; O'Neill et al., 2020; Weissensteiner et al., 2020; Wellenreuther, Merot, Berdan, & Bernatchez, 2019)

SVs can originate through a variety of DNA repair mechanisms, errors during meiotic recombination, and the transposition activity of mobile elements (Hastings, Lupski, Rosenberg, & Ira, 2009; Scully, Panday, Elango, & Willis, 2019; Weckselblatt & Rude, 2015). 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 (Bracewell, Chatla, Nalley, & Bachtrog, 2019; Richards et al., 2005). 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 (Chiang et al., 2017; Choi & Lee, 2020; Y. Huang, Shukla, & Lee, 2022; Kronenberg et al., 2018; Weissensteiner et al., 2020; Zichner et al., 2013). 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; Gebert et al., 2021; Ozata, Gainetdinov, Zoch, O'Carroll, & Zamore, 2019; F. Yang & Xi, 2017). However, there is little agreement on the nature of the evolutionary forces shaping TE abundance levels (Dolgin & Charlesworth, 2008). Ultimately, better genome assemblies will increase our understanding on how different evolutionary forces shape genome structure and TE content.

The genus Drosophila has been a model for studying eukaryotic genome evolution (Bracewell et al., 2019; Drosophila 12 Genomes et al., 2007; Richards et al., 2005). Although new genome assemblies based on long-read sequencing have emerged for several species in this genus (Allen, Delaney, Kopp, & Chenoweth, 2017; Liao, Zhang, Chakraborty, & Emerson, 2021; Mahajan, Wei, Nalley, Gibilisco, & Bachtrog, 2018; Miller, Staber, Zeitlinger, & Hawley, 2018), 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 (T. Dobzhansky, 1944; Fuller, Leonard, Young, Schaeffer, & Phadnis, 2018; Korunes, Machado, & Noor, 2021; Kulathinal, Stevison, & Noor, 2009; Machado, Kliman, Markert, & Hey, 2002; M. A. Noor, Grams, Bertucci, & Reiland, 2001; M. A. F. Noor et al., 2001; Orr, 1987). *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 (T. Dobzhansky & Epling, 1944). The genome of these species is organized in four telocentric chromosomes (2^{nd}, 3^{rd}, 4^{th}, 5^{th}), 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) (Anderson, Ayala, & Michod, 1977; Schaeffer et al., 2008; Tan, 1935). 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 & 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") (T. Dobzhansky, 1944).

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 (Bracewell et al., 2019; Liao et al., 2021; Miller et al., 2018), 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 (B. Y. Kim et al., 2021; Miller et al., 2018), but their utility for studying SVs and their divergence are 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 of genes located inside the fixed chromosomal inversions that separate these two species.

MATERIALS AND 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 (Drosophila 12 Genomes et al., 2007; Richards et al., 2005). 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 (M. 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. Default parameters of PbJelly (PBSuite v15.8.24) (English, Richards et

al. 2012) and Pilon v1.22 (Walker, Abeel 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, Hill 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.

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, Baldwin-Brown 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 (Bracewell et al., 2019;

Liao et al., 2021; Miller et al., 2018) were used to confirm contig orientation, contiguity and chromosome assignment of contigs (Figure S13 Appendix A). Genetic markers from (Schaeffer, Bhutkar et al. 2008) and (Bracewell, Chatla 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, Hubley et al. 2013-2015) using the drosophila library from RepBase in 2017 (Boa, Kojima 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* and previously published data from *D. pseudoobscura* (Paris, Villalta et al. 2015; Nyberg and Machado, 2016). RNA-seq reads were mapped to the new genome assemblies using Hisat2 v2.1.0 (D. Kim, Paggi, Park, Bennett, & Salzberg, 2019). These mapped reads were used to build transcriptome assemblies for each sample for both species using StringTie v2.1.1 (M. Pertea et al., 2015) using Drosophila-optimized parameters (H. W. 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 non-redundant 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 FlyBase 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=30000', 'min_intron=20' (Venturini, Caim et al 2018). SNAP v2006-07-28 (Korf 2004) and Augustus v3.3.3 (Stankle, Keller 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, Waterhouse et al. 2015; Seppey, Manni 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). Lastly, one more round of annotation was run to improve previous annotated gene models.

In addition to the MAKER annotations, annotations from *D. pseudoobscura* from FlyBase, from more recent improved annotations (Yang, Jaime et al 2018) and lncRNA annotations (Nyberg and Machado 2016), were also transferred to the two new genome assemblies using liftOver (Kent, Sugnet et al 2002) implemented in the flo pipeline (Pracana, Priyam 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 on the annotation algorithms. To create a final annotation dataset, Mikado v1.2.4 (Venturine, Caim, et al. 2018) was implemented using three different annotation sources: transcriptome assembly, MAKER annotations and FlyBase liftovers. As annotations for ncRNAs from (Nyberg and Machado 2016) were not included on the Mikado runs, those were merged later using GffCompare (G. Pertea & 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, Steinbiss et al. 2013). Potentially spurious annotations (genes < 100 bp) were removed from the final consensus annotation. The full post-procesing 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 stablish 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 inverted regions 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 FlyBase to transfer putative functional gene annotation using gene ontology (GO) terms. The results from CLfinder-OrthNet and OrthoFinder analyses were used to generate a collinearity figure (Figure 2) with the package genoPlotR v0.8.11 (Guy, Roat Kultima, & Andersson, 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 two 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 (M. Chakraborty, Emerson, Macdonald, & Long, 2019) and paftools.js from minimap2 (Li, 2018) using whole-genome alignments (Chakraborty, VanKuren 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 (Figure S13 Appendix A) and validated inversion breakpoint regions using reciprocal mapping of CLR reads for chromosomes 2 and XL (Figures S13-25 Appendix A). In addition, we further validated inversion breakpoints (Machado, Haselkorn, & Noor, 2007) for those major rearrangements using SyRI v1.3 (Goel, Sun, Jiao, & Schneeberger, 2019). Full command lines are shown in https://github.com/javibio-git/SV_analysis_for_Dpse_and_Dper.

Recombination landscapes

DNA-seq data from 35 and 20 populations of *D. pseudoobscura* and *D. persimilis*, respectively, were used to estimate the number of recombination events implementing a non-overlapping 50kb sliding-windows approach (Chan, Jenkins, & Song, 2012). Trimmed raw reads of each line were mapped to their corresponding genome assembly using bwa v 0.7.17-r1188 (Li & 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 & O'Connor, 2020). Filtered bi-allelic SNPs were used to estimate the mean number of crossover events per generation (*p*/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 1000 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 thank 10kb were also included, but excluding regions with overlapped annotations.

Correspondence analysis was conducted using the corrplot package v0.90 (T. Wei & Simko, 2021) in R, to determine significant associations of SVs versus gene regions and SVs versus differentially expressed 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, Tam, Paniagua, & Hammell, 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 two species. Differential expression of TE families between species was conducted for each developmental stage independently using DESeq2 v1.30.1 (Love, Huber, & Anders, 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, Duggal, Love, Irizarry, & Kingsford, 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 10kb upstream regions were counted using custom perl scripts. Significant association between SVs and differentially expressed genes was assessed using custom scripts in R.

GO enrichment analyses

GO enrichment analysis was conducted using GOrilla web tool (Eden, Lipson, Yogev, & Yakhini, 2007; Eden, Navon, Steinfeld, Lipson, & Yakhini, 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)

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 ST, 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* (Figure 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* (Figure 1A).

The genome assembly for *D. pseudoobscura* appears to be less fragmented (lower number of contigs) but the N50 value is higher in *D. persimilis* (Figure 1B). A summary of genome assembly statistics for both species can be found on 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 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 (Figure 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, (Figures 1A, 2). We also confirm a pattern where all derived inversions between the two species appear to have arisen in *D. persimilis* (Machado et al., 2007; Tan, 1935).

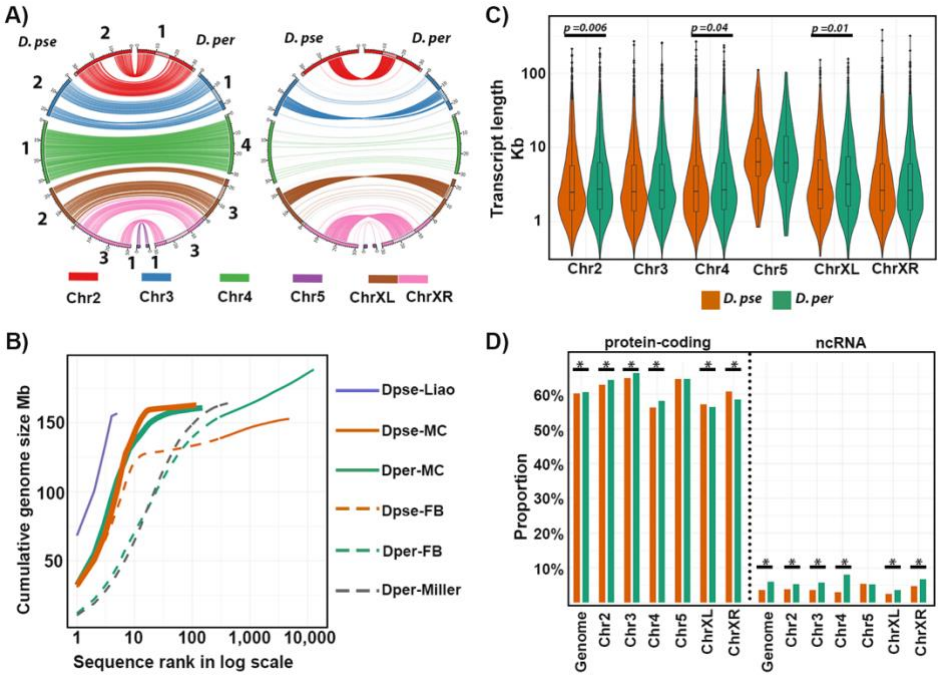


Figure 1. (A) Circos plots showing collinear blocks (left) and inverted regions between *D. pseudoobscura* and *D. persimilis* (right). The number of contigs for each chromosome are indicated on the left plot. (B) Comparison of assembly contiguity with previously published assemblies. MC: our study using 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%)

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 number of annotated ncRNAs and transcript lengths are significantly higher in *D. persimilis* (Figure 1). Because the overall transcript length is longer in *D. persimilis* (Figure 1C), a significantly higher proportion of bp are 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, Figure 1D). However, the difference in transcript length is due to UTR length and not to intron size (Figure S1 Appendix A). When 3' and 5' UTRs are included, whole gene span is significantly longer in *D. persimilis* for chromosomes 2, 4 and XL (Figure 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 (Figure S1

Appendix A). 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 (Figure S1A Appendix A). 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 (Figure S1B Appendix A).

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 300 and 15,000 bp for both species, and the remaining 10% includes genes having a length between 15 kb and 400 kb. 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*. Amino acid length can differ between species up to 20%.

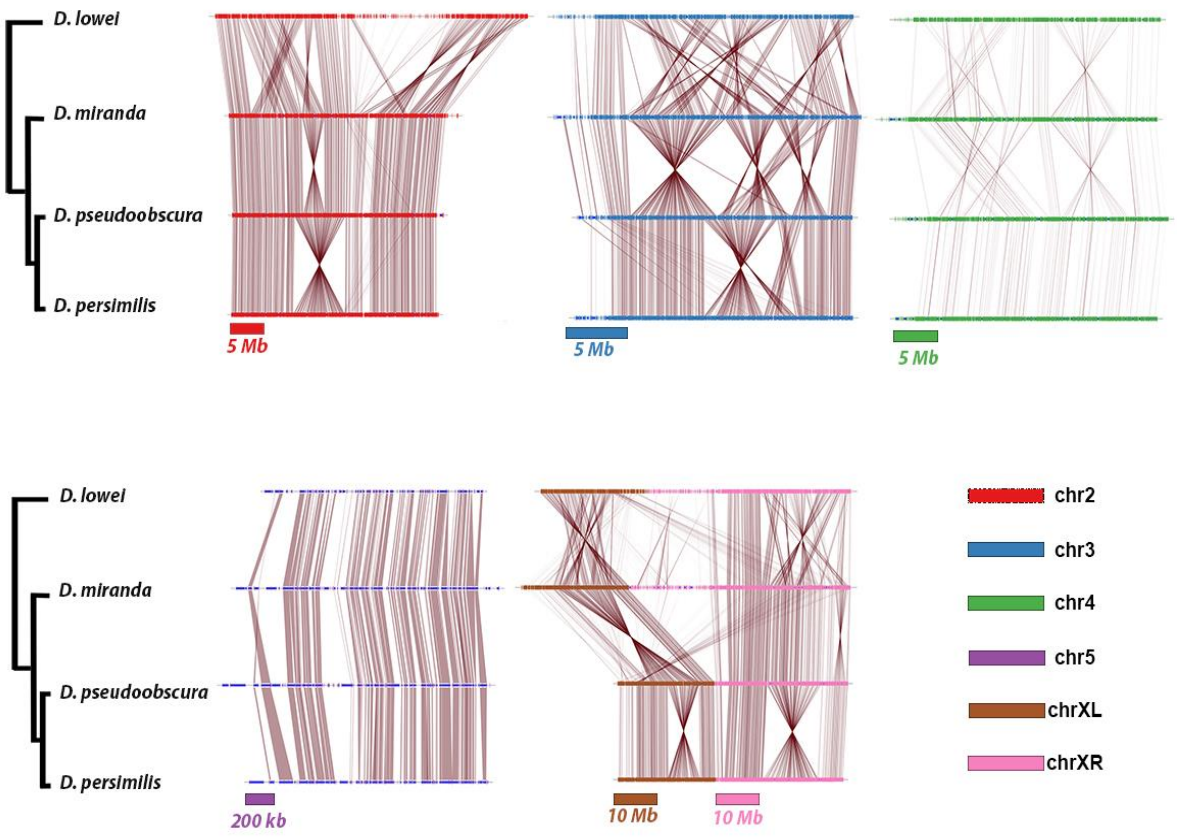


Figure 2. Gene collinearity plots for 1-to-1 single-copy orthologous genes across the *pseudoobscura subgroup*. Chromosomes are color-coded as in Figure 1A. Vertical brown lines represent single-copy orthologs identified by OrthoFinder.

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 range from 11,627 (*D. persimilis*) to 12,361 (*D. miranda*). This range in the number of genes reflects the existence of potential gene 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 inter-chromosomal 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* (Figure 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 (Figure 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) (Figure 4A). We also called a total of 551 and 322 CNVs for *D. pseudoobscura* and *D. persimilis*, respectively (Figure 3A). Our analyses reveal a greater accumulation of INS in *D. persimilis* (Figure 3A). Nevertheless, the size distribution of INDELs suggest that INS in *D. pseudoobscura* are larger than in *D. persimilis* (Figure 3A; Mann-Whitney U Test, p = 2.69e-10). Further, the number of identified CNVs is greater in *D. pseudoobscura* but they have a similar size distribution in *D. persimilis* (Figure 3A; Mann-Whitney U Test, p = 0.388).

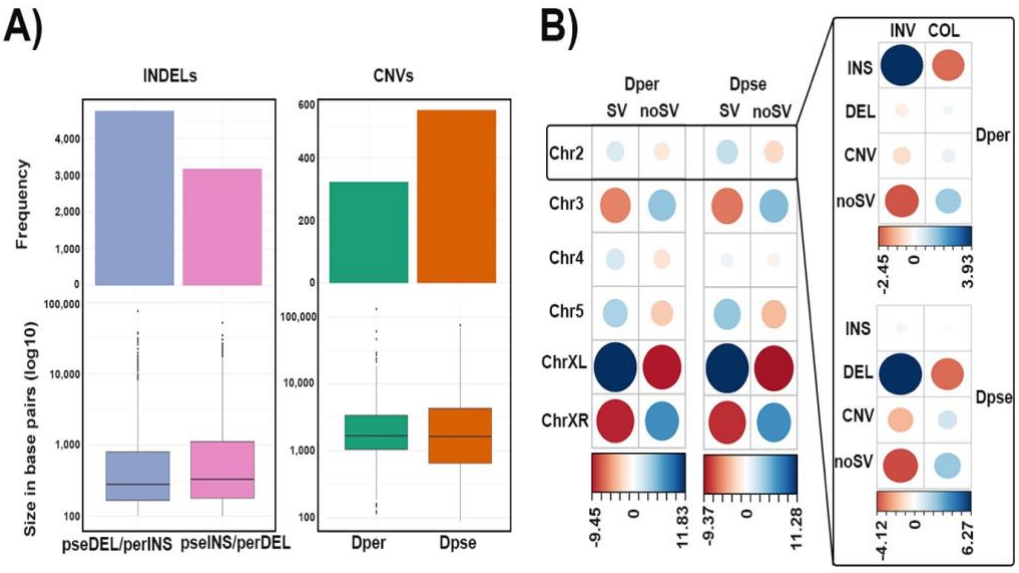


Figure 3. (A) Number of INDELs and CNVs (barplots) and size distribution for each SV type (boxplots). (B) Correspondence analysis showing the association between genes (including the 10kb upstream region) and SVs, for each chromosome. Circle sizes depict the number of genes, and color depicts correlation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

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, 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 (Figure 3B).

Interestingly, we found that genes located inside the major fixed inverted regions are more likely to be associated with SVs than genes in collinear regions. For both species we found a

significantly higher proportion of INDELs associated with genes within inversions than in collinear regions for chromosomes 2 and XL, but not in chromosome XR (Figure S2 Appendix A). The proportion of CNVs is higher in the inverted region of chromosome XR only for *D. pseudoobscura* and for chromosome XL in the two species (Figure 3B, Figure S2 Appendix A).

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 (Merel, Boulesteix, Fablet, & Vieira, 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, 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, and that the four species share the most abundant TE classes and families (Figure 4A, B). Although a considerable proportion of TEs was annotated as ‘unknown’ (Figure 4A), most TE annotations fall in four TE classes and 10 TE families (Figure 4B).

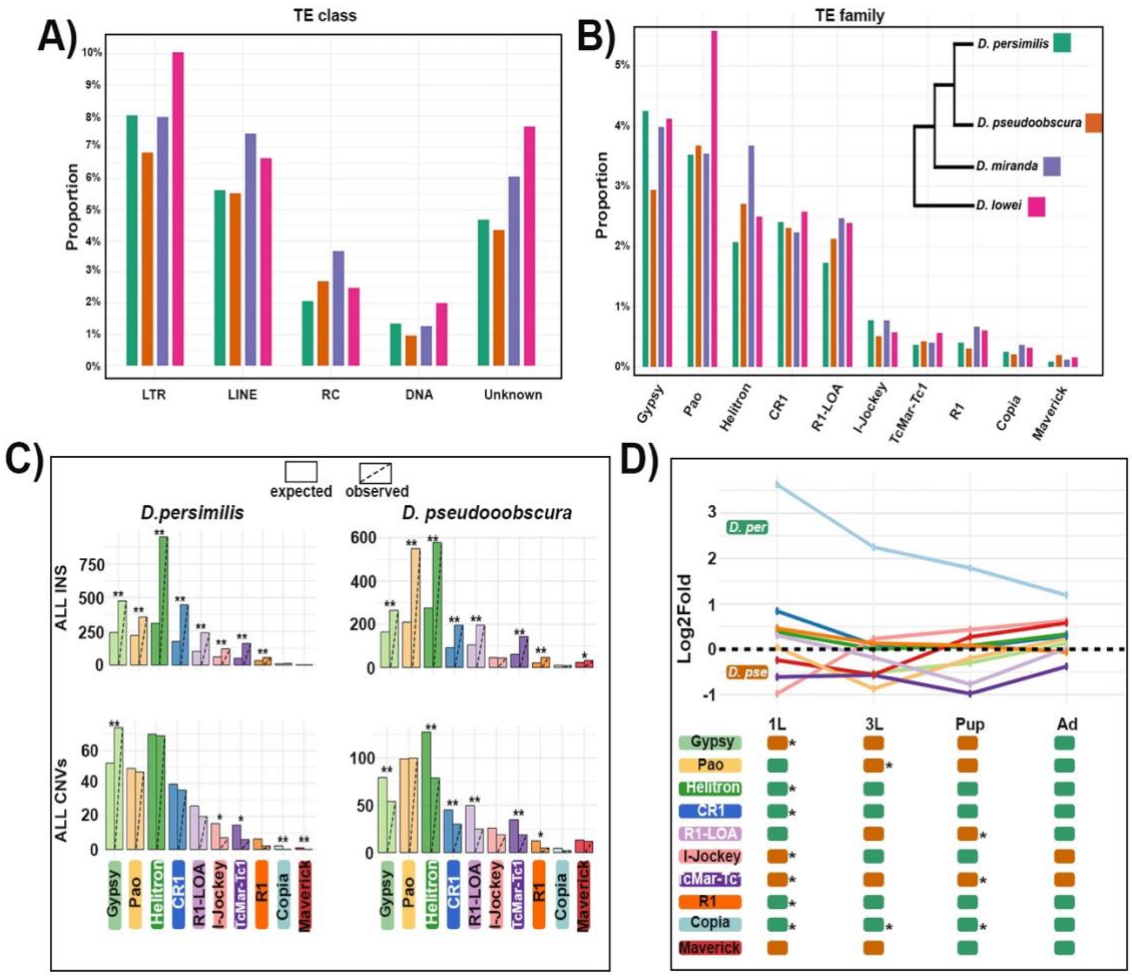


Figure 4. TE content, TE-SV associations and TE expression of the *pseudoobscura subgroup*. (A & B) proportion of TE classes and families across the pseudoobscura subgroup. (C) permutation analysis of INS and CNVs overlapping TE annotations; * $p < 0.05$; ** $p < 0.01$ significant difference between observed and expected counts. (D) Differential expression analysis during development for each TE family. 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.

Almost every TE family is significantly associated with INDELs in both species suggesting that TEs are a primary source of INDEL generation (Figure 4B; Figure S3 Appendix A).

Interestingly, only in *D. persimilis* there is a significant association of the Gypsy family with annotated CNVs (Figure 4C), whereas most of the TE families are significantly underrepresented in CNVs.

Using RNA-seq data from multiple developmental stages, we found that global TE expression is not proportional to the TE content present in the genome assemblies (Figure 4, Figure S4 Appendix A). For instance, the Helitron family is expressed at higher levels in *D. persimilis* even though the *D. pseudoobscura* genome has a higher content of this family. Conversely, the Gypsy family is more highly expressed in *D. pseudoobscura* even though *D. persimilis* has a higher content of this family. Furthermore, even TE families with low representation in the assemblies, such as Copia and TcMar-Tc1, can show expression levels similar to those of the most abundant TE families in the assemblies (Figure S4 Appendix A). In addition, our global analysis of differential expression indicates that TEs are more highly expressed in *D. persimilis* in each of the developmental stages analyzed here (Figure 4D). We found that the most significant differences in expression are observed in the first instar larvae stage, where we observed significant expression differences for four TE families in *D. persimilis* and for four TE families in *D. pseudoobscura*. These overall results suggest that the host genome might have evolved mechanisms to reduce the expression of the most abundant TE families.

Given that TE expression patterns suggest potential species differences in TE regulatory mechanisms, we compared the expression of genes known to be involved in the regulation of TE expression in Drosophila (Ozata et al., 2019). A gene known to be at the center of mechanisms of defense against TE proliferation in the germline and in somatic tissue show significant expression differences between species. *Dcr-2*, a gene involved in the generation of siRNAs (Galiana-Arnoux, Dostert, Schneemann, Hoffmann, & Imler, 2006), is also expressed at

significantly higher levels in *D. pseudoobscura* in all developmental stages (Figure S6 Appendix A).

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 (Figures 4, 5). We calculated the odds ratios between each TE family overlapping SVs within gene regions and intergenic regions (Figure 5) and found that even though TEs are significantly underrepresented in gene regions (Figure S7 Appendix A) 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* (Figure 5).

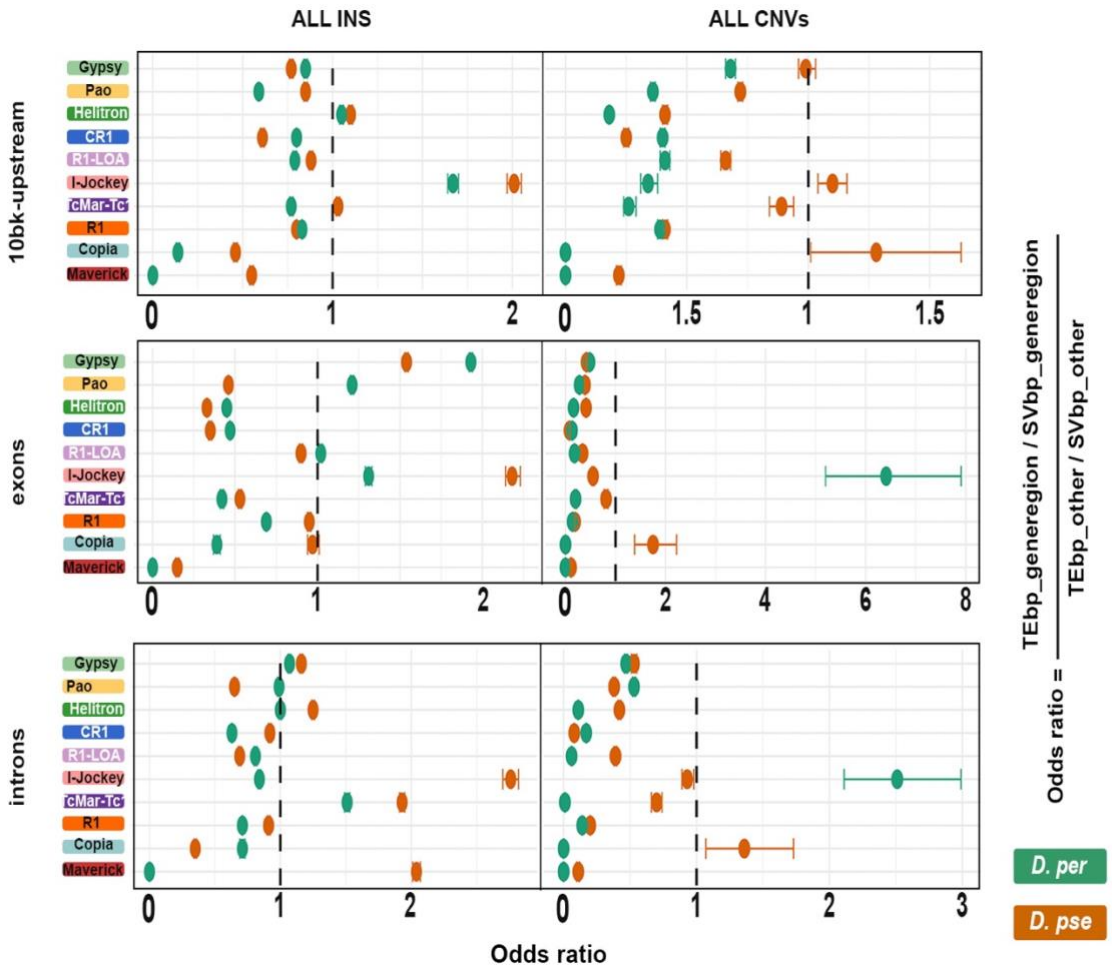


Figure 5. Odds ratios of 2 x 2 contingency tables for TE-SV (INS and CNV) associations with different gene regions (10kb-upstrean, exons, introns).

Accumulation of Transposable Elements in regions of low recombination

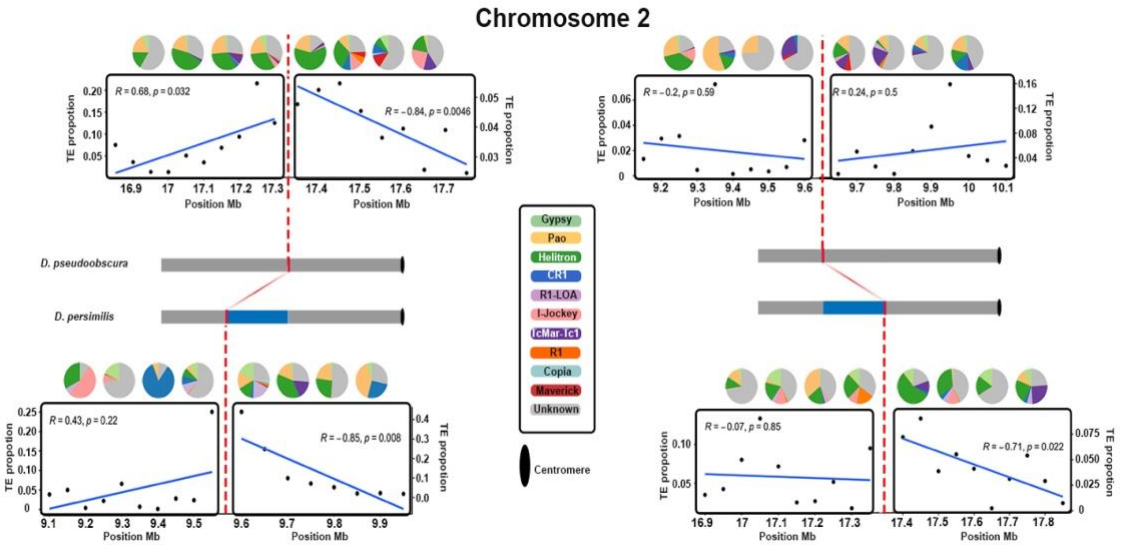
We estimated population-based fine-scale recombination rates for both species and observed significant negative correlations between TE content and recombination rates in both species ($p < 0.05$) (Figure S7 Appendix A). Recombination rates are close to an order of magnitude lower in all *D. persimilis* chromosomes (Figure S8 Appendix A), including the inverted regions of this species (Figure S9 Appendix A). Comparing collinear and inverted regions we observe that recombination rates are significant higher in the fixed inverted regions from chromosome X in

both species (Figures S10, S11 Appendix A), but significantly lower in the inverted region of chromosome 2 in *D. persimilis* (Figure S11 Appendix A). Consistent with the expected negative correlation between TE content and recombination rate (Dolgin & Charlesworth, 2008), TE content is significantly lower inside the inverted regions of chromosome X in both species, while the inverted region of chromosome 2 in *D. persimilis* shows a slight non-significant increase in TE content (Figures S10, S11 Appendix A). Further, there is also a significant decrease in TE content in chromosome 3 for both species (Figures S10, S11 Appendix A).

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* (Figure 6). A similar but more pronounced pattern is observed towards the distal inversion break point in *D. persimilis*, and this region has the highest overall TE content compared to *D. pseudoobscura* (Figure 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 ~350kb that overlaps with the inversion breakpoint (Figure S12 Appendix A). 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 towards 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 collinear region just outside the proximal inversion breakpoint for both species (Figure 6). For chromosome XL, we only observe a significant increase in TE content towards the proximal inversion breakpoint (within the inversion) in *D. persimilis* (Figure 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 (Figure S13 Appendix A). 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* (Figure 6).



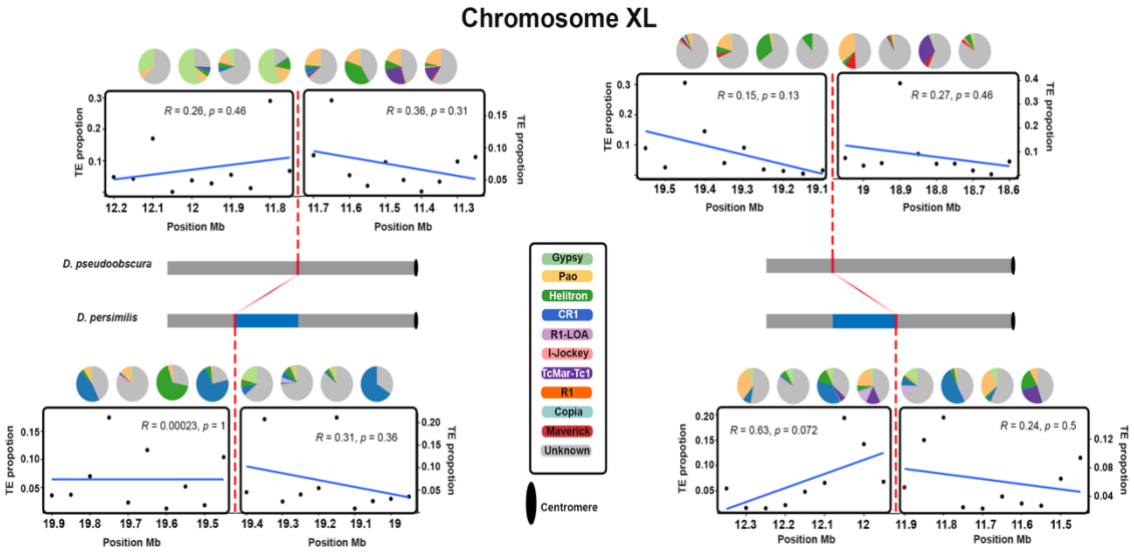


Figure 6. TE content at the proximal *D. pseudoobscura* and distal *D. persimilis* (left) and distal *D. pseudoobscura* and proximal *D. persimilis* (right) inversion breakpoints in chromosomes 2 and XL. Each dot from the scatterplots represents a 50 kb sliding window. Solid and dashed red lines depict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

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 genes in first instar larva (1L), 3L, mid-stage 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* (Figure S14 Appendix A). Gene ontology (GO) enrichment analyses of differentially expressed genes show overrepresentation of genes involved in a wide variety of functions from gene regulation to general developmental processes.

Correspondence analyses, considering all INDELs and CNVs, indicate that differentially expressed genes are significantly associated with SVs in both *D. pseudoobscura* and *D. persimilis* (Figure 7A; Figures S14-17 Appendix A). This correlation signal mostly arises from SVs that overlap the 10kb 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 inverted regions. This pattern is stronger in chromosome 2 and is significant across all developmental stages (Figures 7A,B; Figures S14-17 Appendix A). For chromosome XL, we only observed a strong association between SVs and differentially expressed genes in inverted regions during the pupal stage; for chromosome XR the association is significant on genes expressed during the first (1L) and third (3L) instar larvae stages (Figure 7B, Figures S14-17 Appendix A). For chromosome 3, which harbors a rich suite of non-fixed polymorphic inversions, we observed a significantly high proportion of differentially expressed genes inside the inverted region 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 inverted regions, 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*. We then selected all the differentially expressed 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 differentially expressed genes involved in neural system development (GO:0007399) during 1L, 3L, and Pupal stages, and in protein and nutrient transport in adults. 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 (Table S1; Appendix A). 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*.

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* (Aviles-Pagan, Kang, & Orr-Weaver, 2020; W. Y. Chen et al., 2020; Sridharan, Heimiller, Robida, & Singh, 2016; Vieillard et al., 2016). *cnc* and *heph* are genes located in chromosome 2 that show *D. pseudoobscura*-specific indels, an INS in the 10kb upstream region of *cnc* and a DEL in *heph* inside an intron (Figure 7C; Figures S18-19 Appendix A). 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* (Figure S19 Appendix A). For *heph* we observed that the recent ~1kb DEL in *D. pseudoobscura* corresponds to a region overlapping a LINE/CR1 in both *D. persimilis* and *D.*

miranda (Figure 7). These two recent INDELs occurred in *D. pseudoobscura*, leading to a decrease in the level of expression of the two genes relative to *D. persimilis* (Figure 7C; Figure S20 Appendix A).

In *D. persimilis*, *dila* (chromosome 3) shows two recent ~1kb INS occurring close to each other (Figure 8C), whereas *nebu* (chromosome XR) has a recent ~130 bp DEL, both in the 10kb upstream regions (Figure S19 Appendix A). 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 (Figure S19 Appendix A). Genes involved in gametogenesis are often good candidates for explaining genetic incompatibilities between species. Thus, we also investigated if these genes are differentially expressed between species across development. *dila* is more highly expressed in *D. persimilis* in larval stages (Figure 7C), whereas *nebu* is more highly expressed in *D. pseudoobscura* (Figure S20 Appendix A).

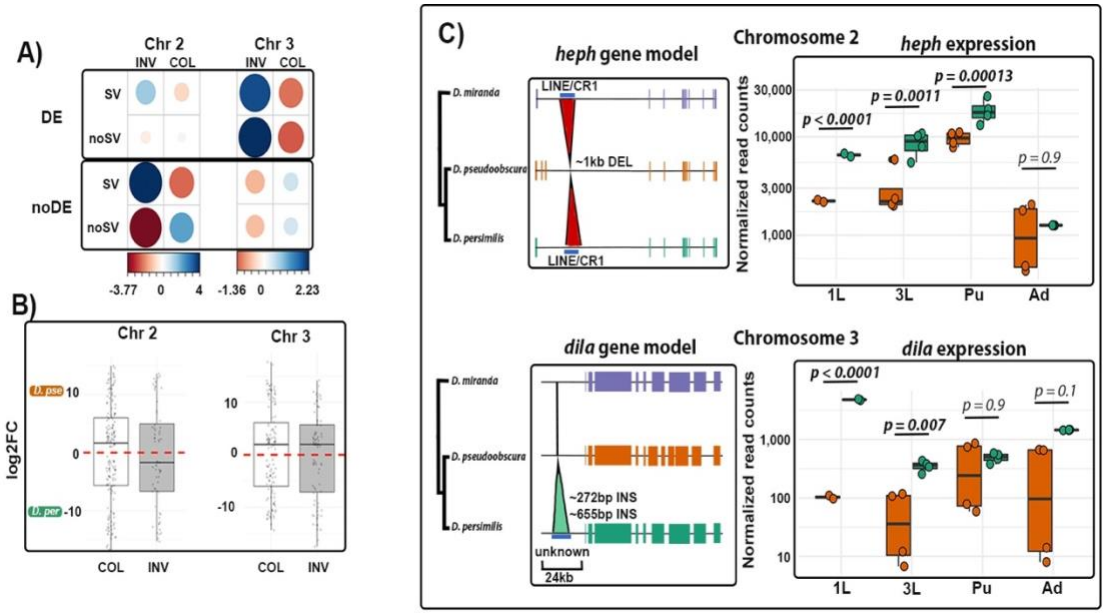


Figure 7. Gene expression and its association with SVs for chromosomes 2 and 3 in *D. pseudoobscura* and *D. persimilis* for the 3L developmental stage. (A) Correspondence analysis showing the association of genes differentially expressed (DE) or not (noDE) with the presence or absence of SVs in the 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 differentially expressed 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 four developmental stages 1L: first instar larvae, 3L: third instar larvae, Pup: Pupae, Ad: Adult, between *D. pseudoobscura* (orange) and *D. persimilis* (green).

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 reported the first *D. persimilis* genome assembly built with long (ONT) reads (Miller et al., 2018), 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. Our hybrid approach resulted in a single contig for chromosomes 2, 3 and 5 and less than five contigs

for each of the remaining chromosomes (Figure 1AB, Table 1). For *D. pseudoobscura* we provide a de-novo high-quality assembly for a new strain that resulted in a single contig for chromosomes 2 and 5, but less than five contigs for each of the remaining chromosomes. 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 shows that there are no other major rearrangements in either of our assemblies (Figures S21-23 Appendix A). We only observed a potential missing section corresponding to the centromeric region of the X chromosome (Figures S21-23 Appendix A), 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 (r3.04) and observed discrepancies that can indicate potential misassembles in the FlyBase genome. Although the correct order of contigs and the identification of potential misassembles of the FlyBase assembly were previously reported by Schaeffer et al (Schaeffer et al., 2008), we detected two additional potential misassembles in chromosome 2 (Figures S24-25 Appendix A).

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 errors, it is possible that it reflects 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 that 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 non-coding transcripts in *D. persimilis*. 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 non-coding transcripts in *D. persimilis* is the result of spurious transcription or transcriptional noise (Darbellay & Necsulea, 2020; Ponjavic, Ponting, & Lunter, 2007), that could arise from lower selection efficiency in this species due to its smaller effective population size (Korunes et al., 2021; Machado et al., 2002).

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 (Figure 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 (Figure 2), and the movement of genes in these species could be the result of recombination events mediated by TEs (Weckselblatt & 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 (L. Zhang, Reifová, Halenková, & Gompert, 2021). The advent of genomic analysis has increasingly shown that hybridization and introgression among closely related species has occurred frequently across the tree of life (Taylor & Larson, 2019). One of the most important mechanisms that allow species to persist in the face of gene flow are chromosomal inversions (Hoffmann & Rieseberg, 2008), and the two focal species of this study have been classic examples of the importance of chromosomal rearrangements for speciation (T. Dobzhansky, 1944; Fuller et al., 2018; Korunes et al., 2021; Machado et al., 2002; M. A. Noor et al., 2001; M. A. F. Noor et al., 2001; Orr, 1987). 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; Figures S26-33 Appendix A) 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 (T. Dobzhansky, 1944).

Inversions can readily arise due to a variety of molecular mechanisms, most of which involve TEs (K. C. Huang & Rieseberg, 2020). Although no specific TE families are associated with the generation of chromosomal rearrangements, association between inversion breakpoints and TE content has been found across kingdoms (Bracewell et al., 2019; Delprat, Negre, Puig, & Ruiz, 2009; Richards et al., 2005; Sharma, Zuo, & Peterson, 2021; J. Zhang & Peterson, 2004). Our results suggest little TE conservation between species, but a slight increase of total TE content near inversion breakpoints in both species for the younger inversion in the 2nd chromosome. Moreover, we found an increase in the proportion of specific TEs next to

breakpoints from the three major inversions, consistent with their potential role in the origin of these fixed rearrangements in *D. persimilis* (Figure 6).

Although former reports of increased genetic differentiation between the fixed inversions separating this species pair were based on SNP differences (Korunes et al., 2021; M. A. F. Noor, Garfield, Schaeffer, & Machado, 2007), we also show that there is a significant overrepresentation of INDELs inside inversions in chromosomes 2 and XL (but not XR). Furthermore, we show that genes located inside the major fixed inverted regions 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 (Figure 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 (M. Alonge et al., 2020; M. Chakraborty et al., 2018; Chiang et al., 2017; Jones et al., 2012), and we show association patterns that suggest a significant relationship between SVs and differential gene expression between this species pair (Figure 7; Figures S14-17 Appendix A). Moreover, we observed a strong signal of differential expression for genes inside inversions compared to collinear regions of the genome consistent with previous studies that show high levels of sequence divergence between inversions in this species pair (Korunes et al., 2021; Kulathinal et al., 2009; Machado et al., 2007; M. A. F. Noor et al., 2007).

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 (Merel et al., 2020) and in many

model systems, including humans (Kofler, Nolte, & Schlotterer, 2015; O'Neill et al., 2020). *D. persimilis* has higher (but no-significant) TE content than *D. pseudoobscura* (Figure S8 Appendix A) genome-wide and inside the inverted regions (Figure S9 Appendix A), consistent with its smaller effective population size (Korunes et al., 2021; Machado et al., 2002). The genome coverage of repetitive elements is significantly different for several major TE families across the *pseudoobscura subgroup* (Figure 4), and a significant proportion of SVs overlap with TE annotations (Figure 4). Consistent with previous findings (Drosophila 12 Genomes et al., 2007; Hill & Betancourt, 2018), our annotation pipeline indicates that *D. persimilis* has a higher TE content, although the observed difference between species is not significant and not as large as previously observed (1% here, 11% in (Hill & 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*. Little is known about the evolution of satellite DNA in these species, but previous work indicate that rapid turnovers of satellite DNA are caused mainly by gains rather than losses (K. H. C. Wei et al., 2018).

Previous work has shown that the frequency of TE insertions often correlates with overall TE activity (Hill & Betancourt, 2018; Z. Liu et al., 2021). However, our data indicate that the most abundant TE families in both genomes do not correspond to the most highly expressed TE families (Figure 4D). It is likely that TE repression mechanisms have evolved differently after the separation of these species leading to the emergence of either more efficient ways to silence TEs or relaxation of TE suppression. Both species exhibit differences not only in the proportion of TE families but also in the expression of a key player involved in TE suppression pathways

(piRNA and siRNA): *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, Betancourt, & Schlotterer, 2012; Kofler et al., 2015; Ozata et al., 2019; F. Yang & Xi, 2017), and can also explain heterogeneous TE family abundances across the Drosophila phylogeny (Hill & Betancourt, 2018; K. H. C. Wei et al., 2018).

D. pseudoobscura and *D. persimilis* show a strong negative correlation between recombination rate and TE content (Figure S7 Appendix A). 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 & Charlesworth, 2008). Interestingly, we observed significant increases in the proportion of TEs at the inversion breakpoint regions from chromosomes 2 and XL (Figure 6), as well as a reduction of local recombination rates on those genomic regions (Figures S12, S13 Appendix A). 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 (Feder, Gejji, Powell, & Nosil, 2011; K. C. Huang & Rieseberg, 2020; Kirkpatrick & Barton, 2006). Once inversions arise, they are usually in heterozygotes and those individuals experience a reduction in recombination that can facilitate

the accumulation of deleterious alleles (Charlesworth & Barton, 2018; Dolgin & Charlesworth, 2008).

ACKNOWLEDGEMENTS

I thank Dr. Therese A. Markow, Dr. Thomas Kocher, Dr Philip Johnson for helpful discussions and technical advice. I also thank Dr. 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. Lastly, I also thank to the members of the Machado Lab for helpful discussions. Research supported by National Science Foundation grants MCB-1716532 and DEB-1754572 to Carlos Machado.

CHAPTER 2: The role of natural selection, recombination, and introgression in the divergence of a classic Drosophila species group

ABSTRACT

Recombination and introgression play critical roles in speciation. In the context of interspecific gene flow, low recombination at regions carrying adaptive alleles can favor adaptive divergence by increasing the efficacy of selection on removing migrant deleterious alleles. These effects can be potentiated by the generation of chromosomal inversions that reduce recombination and generate barriers to introgression between species but can also be enhanced by the evolution of local modifiers of recombination. Here, we study the direct link between gene flow, recombination, and selection in driving patterns of genomic divergence of a classic group of *Drosophila* species (*D. pseudoobscura* and *D. persimilis*). We present two new genome assemblies for both species using high-fidelity long-read sequencing technology. We report that, in colinear regions of the genome, significantly divergent genomic regions ("islands of genomic divergence") between these species tend to be located on regions of low recombination and low introgression. Although inversion differences have been the main contributor to the divergence between these two species, our results show that co-linear regions also harbor genomic outliers in regions where recombination rate and introgression are low. These results suggest that the genetic architecture of species divergence between these species also includes loci in collinear regions of the genome that have diverged thanks to the interplay between recombination and introgression.

INTRODUCTION

One of the main goals of evolutionary genomics is to understand how the interplay of different evolutionary forces can drive processes of adaptive divergence that can eventually lead to the formation of new species. However, it is still unclear how interactions among evolutionary forces (e.g. natural selection, gene flow) and intrinsic genomic factors (e.g. recombination rate, mutation rate) promote adaptive genetic differentiation between populations. Although the study of these interactions has drawn increasing attention in recent years (Aeschbacher, Selby, Willis, & Coop, 2017; Burri et al., 2015; Chase, Ellegren, & Mugal, 2021; Martin, Davey, Salazar, & Jiggins, 2019; Nachman & Payseur, 2012; Renaut et al., 2013; Samuk et al., 2017), testing the specific contribution of different factors is still limited by the availability of data from suitable species models. Genomic data from species complexes that have evolved under different scenarios (e.g. allopatry vs sympatry/parapatry, introgression vs no introgression) can provide important material for testing the combined contributions of natural selection, gene flow and recombination in the processes of adaptive divergence and speciation.

Models that describe population or species divergence in the presence of gene flow (under sympatry or parapatry) predict incipient genomic divergence occurring at small regions presumably under strong divergent selection due to the presence of loci involved in genetic incompatibilities or adaptive divergence (C. I. Wu & Ting, 2004). Those small regions will eventually increase in size through divergence hitchhiking (Feder, Egan, & Nosil, 2012; Nosil & Feder, 2012; Via, 2012), and the maintenance and increase in the size of this so-called "genomic islands of divergence" under divergent selection will be favored by decreased levels of recombination and selection against gene flow at those regions. Reduced recombination

increases the efficiency of natural selection in removing deleterious alleles entering a population via gene flow (Aeschbacher et al., 2017; Barton & Bengtsson, 1986), and regions with reduced recombination can be generated by chromosomal rearrangements (Navarro & Barton, 2003; Rieseberg, 2001) or by the evolution of recombination modifiers (reviewed in (Ortiz-Barrientos, Engelstadter, & Rieseberg, 2016)). Under this general scenario, and when the genetic architecture of species differences is the result of many loci spread through the genome, theory predicts a positive correlation between recombination rate and gene flow across the genome (Aeschbacher et al., 2017; Yeaman & Whitlock, 2011). Such pattern has been observed in several study systems (Aeschbacher et al., 2017; Chase et al., 2021; Marques et al., 2016; Martin et al., 2019; Samuk et al., 2017) but not in others (Burri et al., 2015; Renaut et al., 2013), and there is now a theoretical framework to estimate the strength of selection against migrant alleles (Aeschbacher et al., 2017).

The genus Drosophila has been widely used for understanding the relationship between genome evolution and species diversification (Bracewell et al., 2019; Drosophila 12 Genomes et al., 2007; Sanchez-Flores et al., 2016). Although extensive work has been conducted to try to elucidate the genetic basis of speciation and patterns of genomic divergence, evolutionary inferences on the combined effects of recombination rate and introgression on species divergence are lacking. For instance, the *Drosophila pseudoobscura* subgroup has been a classic group widely used to study the genetics of species divergence (T. Dobzhansky & Epling, 1944; M. A. Noor et al., 2001; Orr, 1987; Nitin Phadnis & Orr, 2009), and to study patterns of genomic divergence in the context of recombination suppression caused by fixed chromosomal inversions between species (Korunes et al., 2021; Machado et al., 2007; M. A. F. Noor et al., 2007). The two most studied members of this species group are *D. pseudoobscura* and *D. persimilis* which

diverged in the last 1 Mya in parapatry with evidence of gene flow between lineages (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 (T. Dobzhansky & Epling, 1944). The genome of these species is organized in four telocentric chromosomes (2^{nd}, 3^{rd}, 4^{th}, 5^{th}), 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) (Anderson et al., 1977). Furthermore, a large inversion in the right arm of X (XR) is fixed among *D. pseudoobscura* and non Sex-Ratio (SR) XR *D. persimilis* strains (Policansky & 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") (T. Dobzhansky, 1944). The other key member of this species group is *D. p. bogotana* (T. Dobzhansky & Epling, 1944), which is an allopatric subspecies of *D. pseudoobscura* found only at high elevations in the Andes mountains in Colombia. These two taxa are considered subspecies due to their more recent divergence time (0.15 Mya), incomplete reproductive isolation (male sterility only occurs on one mating direction), and lack of fixed genomic rearrangements.

This species trio has become a great model for studying the genomics of species divergence under different geographic modes of speciation: one in which divergence has happened with gene flow (*D. pseudoobscura vs D. persimilis*), and one in which recent introgression has played no role (*D. pseudoobscura vs D. p. bogotana,* and *D. persimilis vs D. p. bogotana*). Although patterns of sequence divergence (Dxy and Fst) among these three species have been surveyed in previous studies (Korunes et al., 2021; Kulathinal et al., 2009; Machado et

al., 2007; Machado et al., 2002; McGaugh & Noor, 2012; Stevison, Hoehn, & Noor, 2011), the only inferences made about the genomic effects of fixed inversions on genomic patterns of divergence have been qualitative: inverted regions of the genome are more diverged, and there is evidence of introgression in collinear regions of the genome. So far, no studies assessing the combined effects of natural selection, introgression and gene flow have been published in this system.

In the present study we explore the relationship between recombination rate and introgression to test if the signal of adaptive divergence between species is associated with such intrinsic genomic factors. First, we present new highly contiguous genome assemblies for *D. pseudoobscura* and *D. persimilis*. Using these high-quality genomes, we revisit patterns of diversity and divergence for the whole subgroup including *D. miranda* and *D. lowei* as outgroups. We then measure population recombination rates and admixture proportions to test the hypothesis that adaptive divergence happens more often in regions of low recombination and introgression focusing only on collinear genomic regions.

MATERIALS AND METHODS

Whole genome sequencing, genome assembly and genome annotation.

40 female individuals of one strain of *D. pseudoobscura* (strain: 14011-0121.12) and one strain of *D. persimilis* (strain: 14011-0111.35) were collected and allowed to starve for one day. High molecular weight DNA was isolated by Dr. Michelle Kim from Circulomics (Baltimore, MD) using their Nanobind Tissue Big DNA kit for Animal tissues. DNA samples were sent to the Institute for Genome Sciences at the University of Maryland where Pacific Biosciences HiFi

sequencing was conducted using the Sequel II system. Sequence coverage was 98.2X and 73.6X for *D. pseudoobscura* and *D. persimilis*, respectively.

Genome assembly was conducted using hifiasm v0.15.4 (Cheng, Concepcion, Feng, Zhang, & Li, 2021) for both species. Purge_dups v1.2.5 was implemented to remove overlapping haplotigs and contig overlaps. The final *D. pseudoobscura* genome assembly was based on HiFi reads only. For *D. persimilis*, we implemented a hybrid assembly approach incorporating short read sequencing using MaSuRCa v4.0.3 (Zimin et al., 2017) and quickmerge v0.3 (M. Chakraborty et al., 2016) for merging final HiFi-only and hybrid contigs. Final contigs of both assemblies were oriented and scaffolded using RagTag v2.0.1 (Michael Alonge et al., 2021).

SNP and SV calling

DNA short-read sequencing data for 36 lines of *D. pseudoobscura*, 20 lines of *D. persimilis*, 8 of *D. p. bogotana*, 11 lines of *D. miranda* and 1 line *D. lowei* (Table 1), were mapped with bwa mem v0.7.17 (Li & Durbin, 2009) using the new *D. persimilis* genome as reference, given that the genome of this species resulted in a less fragmented assembly. GATK v4.2.0.0 (McKenna et al., 2010; Van der Auwera & O'Connor, 2020) best practices workflow was implemented to call SNPs for each sample. Briefly, joint genotyping was performed using HaplotypeCaller (-ploidy 1), GenomicsDBImport and GenotypeGVCFs. SelectVariants was then used to collect SNPs from the final gvcf files. Final SNPs were obtained implementing a hard-filtering approach using VariantFiltration (QD < 2.0 , FS > 60.0 , SOR > 3.0, MQ < 40.0, MQRankSum < -12.5, ReadPosRankSum < -8.0) and SelectVariants was used again to keep only bi-allelic SNPs.

To confirm major inversion differences and identify new inversions between *D. pseudoobscura* and *D. persimilis*, a series of pairwise comparisons were conducted using syri

v1.2 (Goel et al., 2019). To polarize new identified inversion differences between *D. pseudoobscura* and *D. persimilis*, the genome of *D. miranda* (Mahajan et al., 2018) was included in the analysis.

Diversity and Divergence analyses

VCF files containing bi-allelic SNPs were used to calculate nucleotide diversity (Pi), absolute divergence (Dxy) and relative divergence (Fst). These statistics were calculated using PopGenome v2.7.5 (Pfeifer, Wittelsburger, Ramos-Onsins, & Lercher, 2014) under R v4.0.5 (R Core Team, 2021). Calculations were performed for both non-overlapping 10 kb sliding windows and genes independently. Sites with missing data were excluded and treated as monomorphic sites. Centromeric regions were identified based on comparisons with a previously published *D. pseudoobscura* genome (Liao et al., 2021) and excluded for downstream analyses given its high repetitive element content. The alignment file from syri (*D. pseudoobscura* vs *D.* persimilis) was used as a reference to consider only 10 kb windows that were syntenic among species to avoid potential effects of alignment biases due to the presence of major repetitive elements differences between species.

Phylogenetic analysis

VCF files including all sequenced lines for each species were converted to a genotype file using the parseVCF.py (-ploidy 1). Genotype files were the input of the phyml_sliding_windows.py script to create maximum likelihood phylogenies for windows of 50 bi-allelic SNPs using PhyML v3.3.3 (Guindon & Gascuel, 2003). A final list of phylogenies for each window was used as input for Twisst v0.2 (Martin & Van Belleghem, 2017), which quantifies the frequency of alternative genealogical relationships across the genome. Both

phyml_sliding_windows.py and parseVCF.py were retrieved and adapted from https://github.com/simonhmartin/genomics_general#processing-vcf-files.

Topology weighting was also implemented for a set of phylogenies created for each gene of the reference genome (*D. persimilis*).

Genome wide tests of selection

Genome scans for signatures of selection were conducted using three different approaches. iMKT was used to detect positive selection on protein-coding genes under the framework of the FWW McDonald-Kreitman test controlling for low-frequency polymorphisms with a cutoff of 0.15. iMKT runs were performed for *D. pseudoobscura* (with *D. persimilis* as outgroup), *D. persimilis* (with *D. pseudoobscura* as outgroup) and *D. p. bogotana* (with *D. pseudoobscura* as outgroup). iMKT implementation was conducted using the adapted python and R scripts retrieved from https://biovcnet.github.io/_pages/lesson-mktest. To include iMKT results in the 10 kb windows approach, bedtools intersect was used to assign MKT results to the 10 kb windows overlapping genes.

Signatures of selective sweeps were detected using XP-CLR v1.1.2 (H. Chen, Patterson, & Reich, 2010) and RAiSD v2.9 (Alachiotis & Pavlidis, 2018). XP-CLR scans were run for non-overalapping 10 kb sliding windows with a maximum of 200 snps (--size 10000, --maxsnps 200) using the same pairwise comparison scheme as for the MKT tests. RAiSD scans were conducted for each species separately for 200 SNPs sliding windows (-w 200). For both selective sweep scans, VCF files containing the filtered bi-allelic SNPs data were used as inputs and sites with missing data were excluded and treated as monomorphic sites. Values of both XP-CLR and Mu statistic of RAiSD above the 99% quantile, for each independent chromosome, were considered significant.

Recombination and introgression rate estimates

Population recombination rates (*p*) were computed for *D. pseudoobscura*, *D. persimilis* and *D. p. bogotana* using the maximum likelihood method implemented in LDHelmet v1.10 (Chan et al., 2012). *p* is expressed in terms of the average number of crossover events in the whole population at a specific site (*p*/bp). VCF files for each species were converted to fasta files using bcftools v1.7 (Danecek et al., 2021) and the required Watterson's estimator value was calculated using PopGenome. Final population recombination rates were transformed to obtain values for non-overlapping 10 kb windows.

Dsuite v0.4 (Malinsky, Matschiner, & Svardal, 2021) was used to estimate admixture and introgression statistics using *D. miranda* as outgroup. First, Patterson's D statistic (Patterson et al., 2012) was calculated using the Dtrios command for each chromosome. Trios with significant D statistic were used to calculate the admixture proportion *fd* (Martin, Davey, & Jiggins, 2015) for each chromosome in sliding windows of 50 SNPs using the Dinvestigate command. Final f_d values were averaged over non-overlapping 10 kb windows.

Fst and Dxy outlier detection

Outliers were identified using two approaches. In the first approach, regions with Fst and Dxy values in the 95th quantile of 10 kb non-overlapping windows were classified as outliers. In the second approach, we implement outlier detection on the same 10 kb genomic windows using k-nearest neighbor techniques (kNN) (Pfeifer, Alachiotis, Pavlidis, & Schimek, 2020). A final set of outliers was defined based on the overlap between results from both approaches.

We used logistic regression to quantify the tendency for outliers to occur in regions of low recombination and low admixture in the collinear regions, using an approach like the one described in (Samuk et al., 2017). For the *D. pseudoobscura* vs *D. persimilis* and *D. p. bogotana*

vs *D. persimilis* (null model) comparisons, both *fd* and *p* were included as part of the model. For all model fits, the effects of gene density (GD) and GC content were also considered. Using 10 kb windows as data points, we fitted a logistic regression model as follows: outlier status ~ [proportion of introgression] * recombination rate + GD + GC, where outlier status is 1 for windows >95th percentile, and 0 otherwise. The model was fitted for the genomic outliers from each species comparison using R v4.2.1 with the generalized model function 'glm' with 'distribution=binomial'.

To test the hypothesis that species divergence is driven by the recombinational landscape and introgression, permutation tests were implemented to assess if *p* and *fd* coefficients from the model above were significantly lower for the *D. pseudoobscura* vs *D. persimilis* comparison. For each species comparison, logistic regression fits using glm were implemented for chromosomes 2, 4, XL and XR. Chromosome 3 was excluded given its elevated inversion polymorphisms within species. In addition, chromosome 5 was also excluded since there wasn't enough data for the model to converge. *p* and *fd* coefficients for each logistic regression fits were used to implement a permutation test for each of the pairwise species comparisons following the approach from (Samuk et al., 2017).

Test of independence

Test of independence were conducted in R using the chisq.test function of the stats v4.0.3 package. In addition, the corrplot package was used to create the contingency plots showing the contribution of each cell to the chi squared statistic.

Figures and data manipulation

All the data generated, and the plots created in the present study were manipulated and created under R with custom scripts and by using the tidyverse metapackage (Wickham et al., 2019).

RESULTS

High-quality genome assemblies using HiFi sequencing.

We present new high-quality genome assemblies for two new lines of *D. pseudoobscura* (*National Drosophila stock center* #14011-0121.12, *Dpse\w[1]*) and *D. persimilis* (*National Drosophila stock center* #14011-0111.35, Dper\wild-type). Our assembly approach resulted in an assembly size of ~174 Mb with 29 contigs for *D. persimilis* and, ~167 Mb with 73 contigs for *D. pseudoobscura*. After scaffolding using the most recent *D. pseudoobscura* reference assembly (Liao et al., 2021), 13 contigs (~173 Mb) were placed into chromosomes with a total gap size of 800 bp in 8 gap sequences for *D. persimilis*. For *D. pseudoobscura*, RagTag placed 37 contigs (~164 Mb) into chromosomes with a total gap size of 3,200 bp in 32 gap sequences. Although both genome assemblies did not result in chromosome level assemblies, assembly and scaffolding statistics reflect highly contiguous assemblies for both species. It is noteworthy that all inversion differences between species were captured in single contigs from the primary assembly (see below).

Our genome annotation pipeline using Maker2, resulted in the annotation of 14,435 and 13,667 protein-coding genes for *D. persimilis* and *D. pseudoobscura*, respectively. In addition, FEELnc annotated a total of 1,242 and 1,162 lncRNAs for *D. persimilis* and *D. pseudoobscura*, respectively. In addition, we complemented our annotations by comparing and transferring

missing protein-coding genes not annotated by our pipeline using the annotation of (Liao et al., 2021) from *D. pseudoobscura*. Our final annotation contains a total of 15,091 and 16,119 protein-coding genes for *D. pseudoobscura* and *D. persimilis*, respectively.

Full-chromosome alignments captured the main inversion differences from chromosomes 2, 3 and X. In addition, we report 21 additional small inversions (10-600 kb), 18 of which are likely derived in *D. persimilis* based on comparisons including *D. miranda* (Mahajan et al., 2018). Those small inversions contain a total of 15 genes altogether in both species, 7 of which have orthologs with *D. melanogaster*. It is important to note that only two inversions of ~10kb are outside centromeric regions located in chromosomes 2 and 4 (Tables S1-4 Appendix B).

Genome-wide patterns of divergence.

A total of 12,211,551 bi-allelic SNPs were called across the five chromosomes. A multidimensional scaling analysis, using identity-by-state (IBS) pairwise relationships, shows a clear species separation for all chromosomes (Figure 1; Figure S1 Appendix B). However, this pattern is slightly different for chromosome 3 where we observed more genetic similarity between the line MSH1993 of *D. persimilis* with other 28 lines of *D. pseudoobscura*, which are separated from the remaining samples clustering closer with *D. p. bogotana* (Figure 1). This separation of *D. pseudoobscura* lines can reflect previous findings on observed population structure for chromosome 3 (Fuller, Koury, Phadnis, & Schaeffer, 2019). Chromosomes 2 and X show the greatest separation between *D. pseudoobscura* and *D. persimilis*, likely reflecting the effects of fixed inversion differences occurring in these two chromosomes (Figure 1; Figure S1 Appendix B).

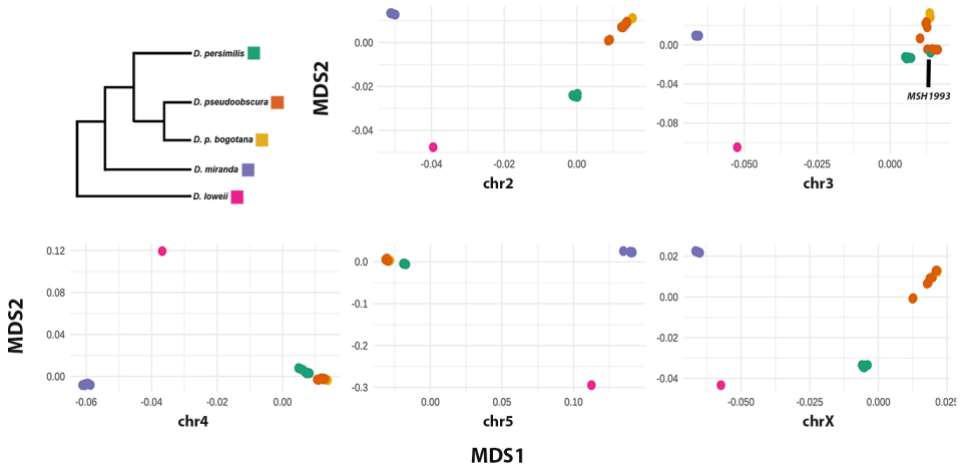


Figure 1. Multidimensional scaling plots (by chromosome) for IBS relationships between members of the *D. pseudoobscura* group. Independent plots for each arm of the X chromosome are shown in Figure S2 Appendix B.

Although patterns of sequence divergence (Dxy and Fst) among species of the *D. pseudoobscura* subgroup have been extensively surveyed in previous studies (Korunes et al., 2021; Kulathinal et al., 2009; Machado et al., 2007; Machado et al., 2002; McGaugh & Noor, 2012; Stevison et al., 2011), we revisited those estimates by including a larger number of samples, and including all autosomes. Divergence levels were lower between *D. pseudoobscura* and *D. p. bogotana*, which only split 0.15 Mya (Figure 2A,B). Although divergence levels reflect the phylogeny of the species (Figures 1, 2), the divergence between *D. persimilis* and *D. p. bogotana* is greater than between *D. persimilis* and *D. pseudoobscura* (Figure 2 A,B) due the effects of recent introgression between *D. pseudoobscura* and *D. persimilis* and a higher rate of divergence in the *D. p. bogotana* lineage since its split from *D. pseudoobscura* (Korunes et al., 2021).

Divergence levels between *D. pseudoobscura* and *D. persimilis* are known to be higher in regions with fixed inversion differences (Korunes et al., 2021; Kulathinal et al., 2009; Machado

et al., 2007; M. A. F. Noor et al., 2007). To test if those patterns only hold for those two species, genomic regions corresponding to their inverted and co-linear regions were also compared for other species pairs. The comparisons between *D. pseudoobscura/D. persimilis* and *D. persimilis/D. p. bogotana* were the only ones showing higher divergence levels in inverted regions (Figure 2 C, D), reflecting the fact that fixed inversions arose before the split of *D. p. bogotana* from *D. pseudoobscura*. Interestingly, in the comparisons including the outgroup *D. miranda* divergence were lower in the inverted regions (Figure 2 C, D). Fixed inversions from chromosomes 2 and XL show the highest levels of divergence between *D. pseudoobscura* and *D. persimilis* (Figure 2 E,F; Figures S2,3,10,11 Appendix B), reflecting the fact that those two inverted regions are the oldest fixed inversions between these species (M. A. F. Noor et al., 2007) and appear to harbor most loci involved in reproductive isolation (M. A. Noor et al., 2001; M. A. F. Noor et al., 2001). Inverted regions show higher absolute divergence (Dxy) than collinear regions in each chromosome that harbors inversions (Figure 2 E; Figures S2,3,10,11 Appendix B), but Fst in the inverted regions from chromosomes 3 and XR shows the opposite pattern (Figure 2 F; Figures S4,5,12,13 Appendix B). This discrepancy is likely the result of higher nucleotide polymorphism within species (Cruickshank & Hahn, 2014), which in the case of chromosome 3 is due to its rich inversion polymorphism (Schaeffer et al., 2003). Chromosome 4, which is fully collinear between *D. pseudoobscura* and *D. persimilis*, has levels of divergence comparable to collinear regions of the other chromosomes (Figure 2 G, H; Figures S6,7 Appendix B). Chromosome 5 has the lowest levels of Dxy but the highest average Fst among all chromosomes, consistent with its reduced diversity (Figures S8,9,14,15 Appendix B) (Larracuente & Clark, 2014; Machado & Hey, 2003).

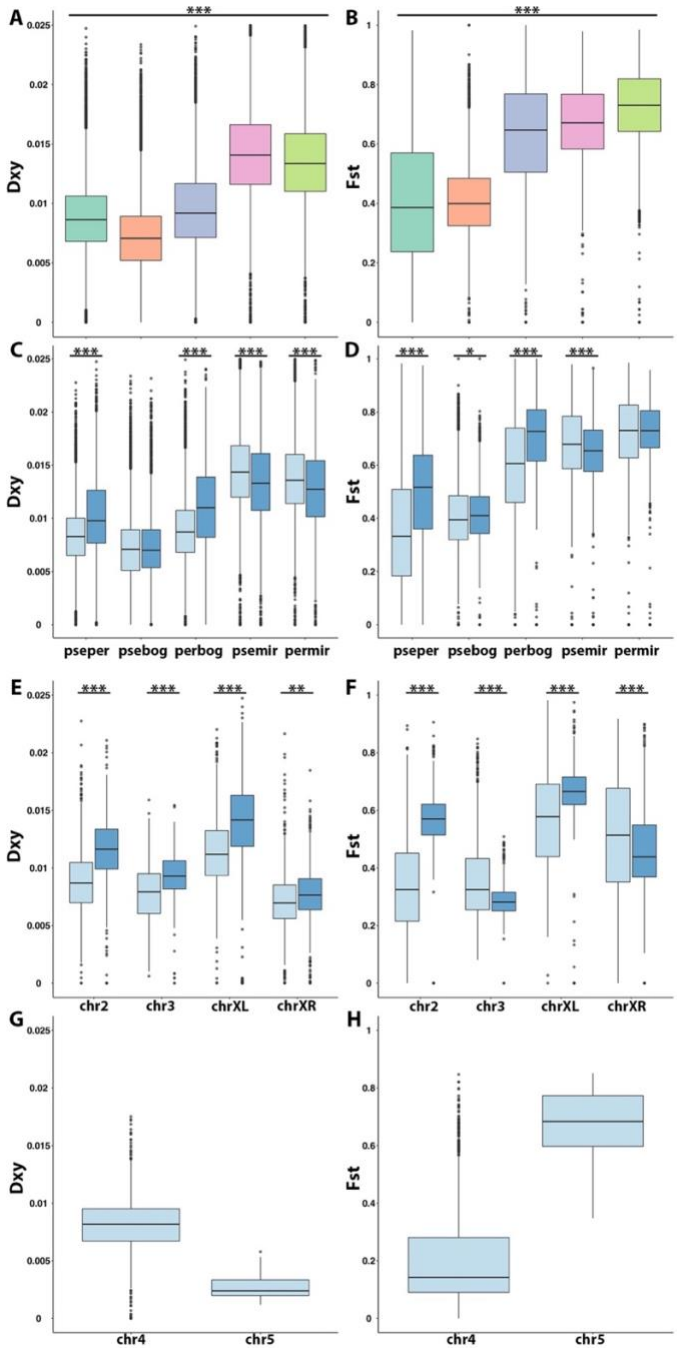


Figure 2. Absolute (Dxy) and relative (Fst) divergence among *Drosophila pseudoobscura* subgroup species. Levels of divergence are shown for pairwise species comparisons (A,B) and for comparisons between inverted (dark blue) and co-linear (light blue) regions (C,D). Inverted regions are based on the fixed inversion differences occurring between *D. pseudoobscura* and *D. persimilis* in

chromosomes 2, XL and XR. Divergence levels are also shown for individual chromosomes only for the *D. pseudoobscura* vs *D. persimilis* (E-H). Abbreviations: *D. pseudoobscura* (pse), *D. persimilis* (per), *D. p. bogotana* (bog), *D. miranda* (mir). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; Kruskal-Wallis test (A, B) ; Mann-Whitney test (C, D-H).

Variation in species relationships across the genome is consistent with patterns of introgression.

We first re-assessed the contribution of introgression to patterns of divergence between *D. pseudoobscura* and *D. persimilis* using topology weighting. This approach quantifies the frequency of alternative genealogical relationships between species across the genome using the SNP data from all sampled individuals. Consistent with previous results about patterns of introgression in this species group (Korunes et al., 2021; Kulathinal et al., 2009; Machado et al., 2007; Machado & Hey, 2003), the observed patterns of genealogical relationships suggest that introgression has contributed to patterns of genomic divergence between *D. pseudoobscura* and *D. persimilis*. We observed that topology 3, which represent the species tree, is the most represented genome-wide (Figure 3; Figures S16-18 Appendix B), but it is also the topology with the highest average weighting (average weighting = 0.35) genome-wide, and with the highest frequency of complete monophyly (weighting = 1; Figure S18 Appendix B). Furthermore, topology 1, which indicates introgression between *D. pseudoobscura* and *D. persimilis*, has the second highest genome wide weighting value (Figures S16-17 Appendix B). Topology 1 has higher weighting in co-linear than inverted regions in chromosomes 2 and 3, and in co-linear chromosomes 4 and 5 there is only a small difference in average weighting between topologies 1 and 3 (Figure 3). These observations are consistent with patterns of introgression described by (Korunes et al., 2021), which show higher signal of introgression occurring in co-linear regions compared to inverted regions in this species pair.

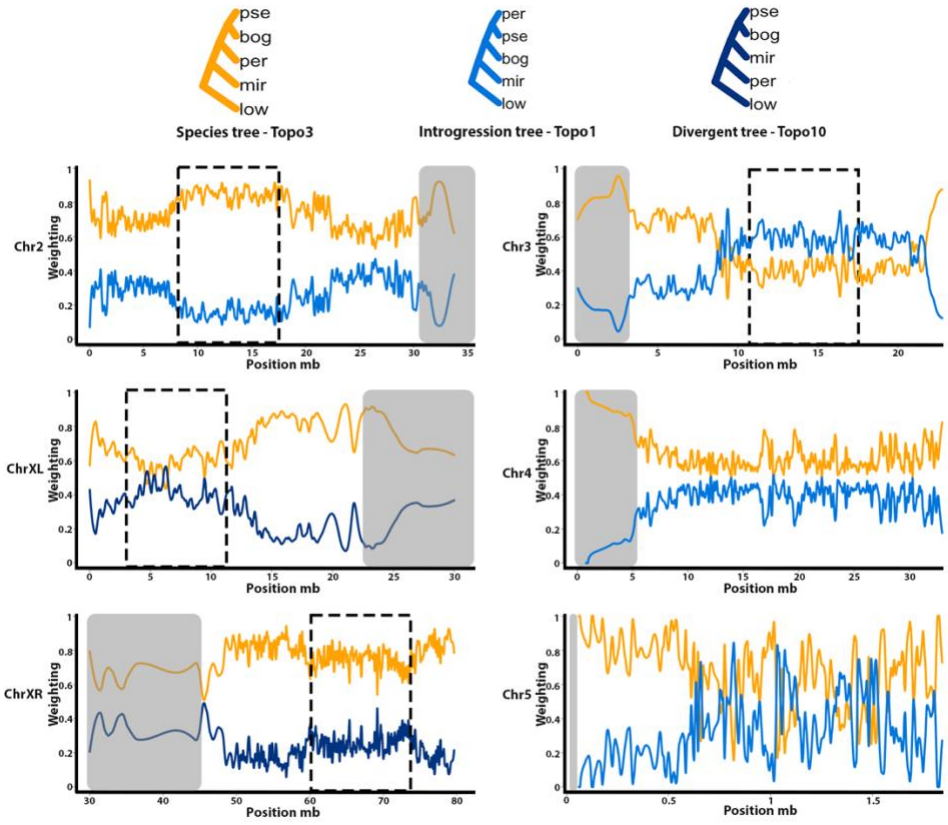


Figure 3. Topology weighting landscape for the three most represented topologies for all chromosomes using 50 SNPs windows. Colors on the line plot correspond to the topologies shown at the top of the figure. Dashed squares highlight the regions with inversion differences between *D. pseudoobscura* and *D. persimilis*. Grey areas highlight centromeric regions.

Interestingly, we observed that for both arms of the chromosome X, topology 10 is the second most represented as opposed to topology 1 (Figure 3; Figure S16 Appendix B). We refer to topology 10 as a 'divergent tree' since the branching order place *D. persimilis* basal to *D. pseudoobscura*, *D. p. bogotana* and *D. miranda*. Moreover, we observed that topology 10 is as represented as the species tree (topology 3) inside the inverted region in the left arm of chromosome X (Figure 3). In the right arm of chromosome X, topology 10 is also the second most common, and while its frequency is slightly higher inside the inverted region, it is not as

common as topology 3 inside the inversion. These findings are consistent with previous observations about unexpected genealogical relationships in few loci from the X chromosome that suggested either the presence of incomplete lineage sorting from the common ancestor of the four species, or a more complex history of divergence of this species group (Machado et al., 2002).

Variation in patterns of introgression across the genome

We evaluated patterns of introgression between *D. persimilis* and *D. pseudoobscura* using the ABBA-BABA test (which estimates Patterson's D) and the admixture proportion (*fd*). Analyses were conducted using the species tree: (((*D. p. bogotana*, *D. pseudoobscura*), *D. persimilis*), *D. miranda*). While Patterson's D is useful for analyzing introgression over large genomic regions, the *fd* statistic is well-suited to estimate admixture frequencies on smaller genomic regions (Martin et al., 2015). We tested for significant excess of ABBA over BABA patterns for each chromosome independently (i.e. excess of derived 'B' alleles shared between *pseudoobscura* and *persimilis* -ABBA- than between *bogotana* and *persimilis* -BABA). We observe significant excess of the ABBA pattern (z-scores > 5) for all chromosomes except for chromosome 3 (Table S5 Appendix B) in agreement with (Korunes et al., 2021). The pattern in chromosome 3 could be the result of confounding factors derived by the inversion polymorphisms within species. Although confounding factors can also arise from chromosome 5 given its non-recombining nature, this chromosome also shows a significant excess of the ABBA pattern, in agreement with a previous result suggesting evidence of introgression at a locus in this chromosome (Larracuente & Clark, 2014; Machado et al., 2002).

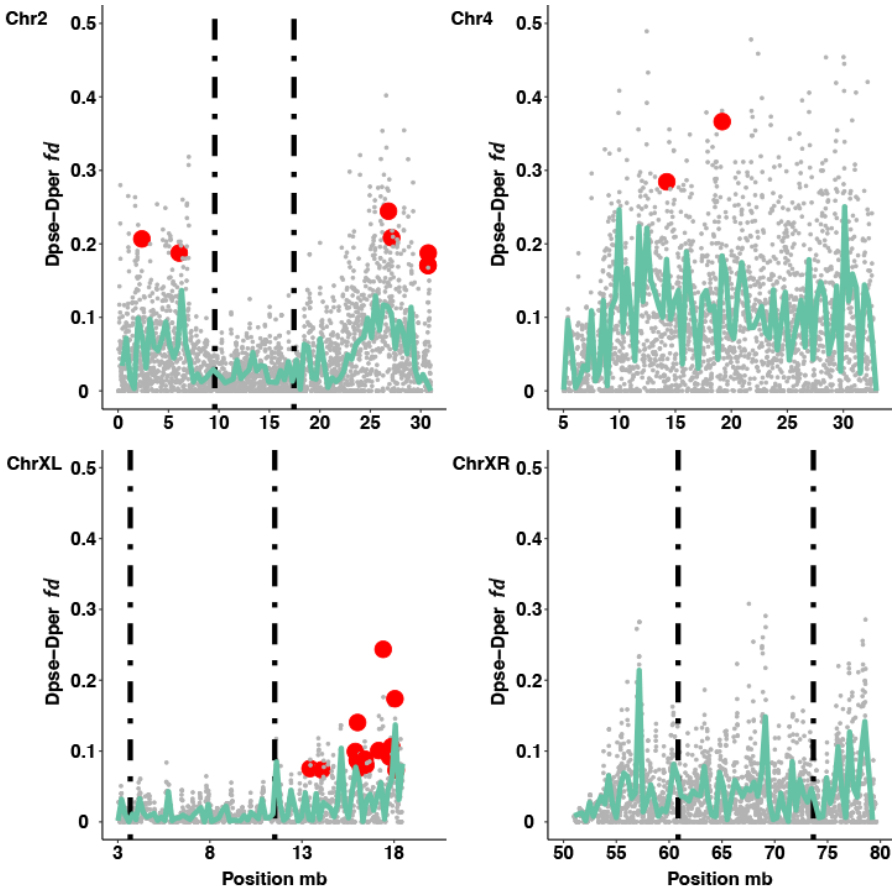


Figure 4. Admixture proportions (*fd*) between *D. pseudoobscura* and *D. persimilis* for chromosomes 2, 4, and X. Each dot represents a 10 kb sliding-window and vertical dashed lines highlight inversion breakpoints. Red dots indicate introgression outliers based on Dxy by the kNN approach (see text).

To explore patterns of introgression between *D. pseudoobscura* and *D. persimilis* at smaller scales we calculated *fd* (the admixture fraction) on 10 kb nonoverlapping windows across all chromosomes. *fd* values vary widely across the genome and appear to be higher in co-linear regions for chromosomes with inversion differences (Figure 4). In addition, we note that not only the signal of introgression is much weaker in chromosome X, but the fully co-linear chromosome 4 shows the highest values of *fd* across the entire chromosome (Figure 4).

Chromosome 5 also show peaks of introgression, in agreement with previous findings from smaller datasets (Larracuente & Clark, 2014; Machado & Hey, 2003).

Recombination rates are negatively correlated with admixture.

We used LDHelmet (Chan et al., 2012) to estimate the population recombination rate (*p*) based on patterns of linkage disequilibrium. Recombination landscapes show wide variability across the genome with several recombination hotspots per chromosome in each species (Figure 5). The landscapes are very similar for both species (Figure 5) and significantly positively correlated (Figure S19 Appendix B), but rates of recombination are about an order of magnitude higher in *D. pseudoobscura* (Figures 5,6). Inversion differences can suppress recombination rate when segregating between populations, and this effect is quite clear in the middle of chromosome 3 which harbors a rich inversion polymorphism in each species (Figures 5,6). In fact, chromosome 3 appears to have the lowest level of recombination (Figures 5,6). Interestingly, we found that regions with fixed inversion differences also have significantly different recombination rates than their corresponding collinear regions in both species (Figure 6). While the inverted region in chromosome 2 has lower recombination rates, both inverted regions in chromosome X have higher recombination rates (Figure 6). These findings are puzzling because it was only in *D. persimilis* that the inversions were polymorphic sometime in the past, but they were thought to have fixed quickly (Machado et al., 2007). Thus, any effects on recombination suppression within species should have occurred in *D. persimilis* and should have been transient.

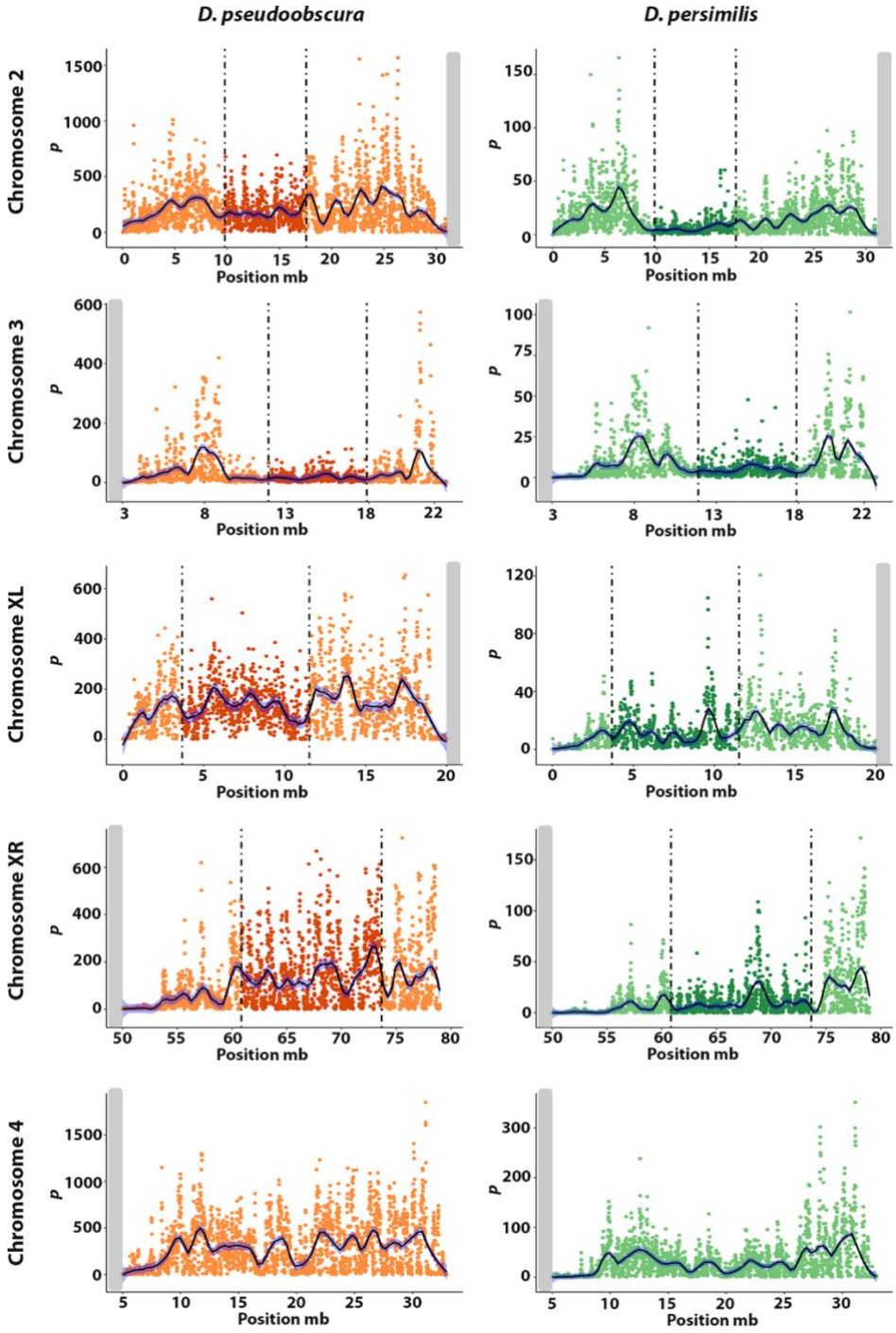


Figure 5. Population recombination rate (*p*) variation across all chromosomes of *D. pseudoobscura* and *D. persimilis*. Each dot represents a 10 kb sliding windows, and solid lines represent the locally weighted average (loess span = 0.8). Windows inside inversion differences are shown in dark orange and dark green; inversion breakpoints are represented with vertical dotted lines. Centromeres are

denoted by grey bars. Note the difference in value on the Y-axis for both species. The pattern for chromosome 5 is shown in Figure S24 Appendix B.

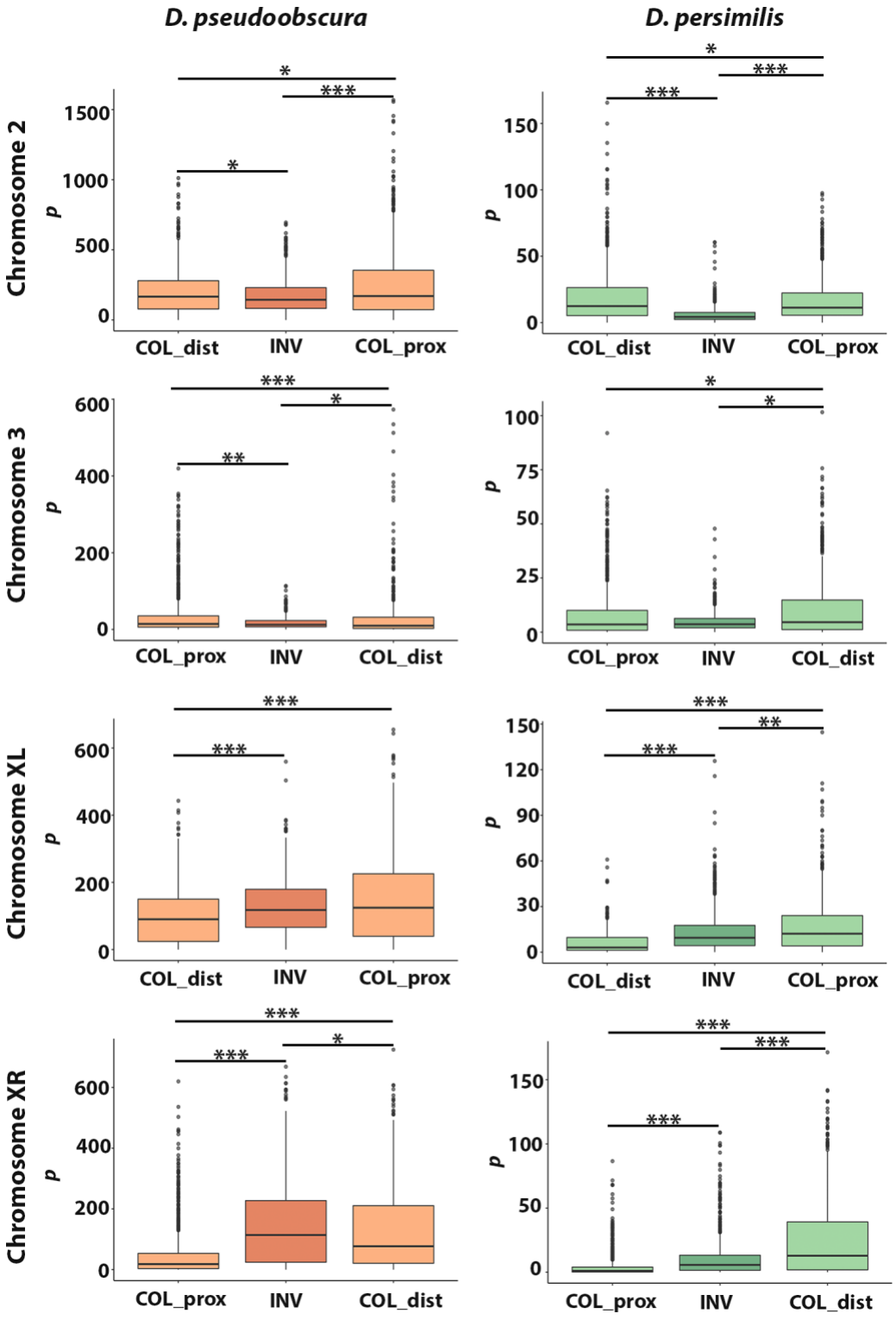


Figure 6. Boxplots of the population recombination rates in inverted regions (INV) and co-linear regions distal or proximal to the centromere (COL_dist, COL_prox). Mann-Whitney test; *** << 0.0001; ** < 0.01; * < 0.05.3

Recombination rate and admixture proportions are positively correlated, and this significant pattern holds for both *D. pseudoobscura* and *D. persimilis* (Figure 7). This pattern is consistent with theoretical predictions about the interaction between recombination rate and selection against introgression that is shaped by species barriers (Aeschbacher et al., 2017; Kirkpatrick & Barton, 2006; Nachman & Payseur, 2012), and has been observed in other species groups (Aeschbacher et al., 2017; Martin et al., 2019; Samuk et al., 2017). The pattern is not driven by data from the inverted regions because the positive correlation is significant even if data from these regions is not included in the analyses. Interestingly, the correlation is negative when comparing *D. persimilis* and *D. p. bogotana* (Figure S25 Appendix B), but we note that any signal of introgression between these lineages predates the split of *D. p. bogotana*.

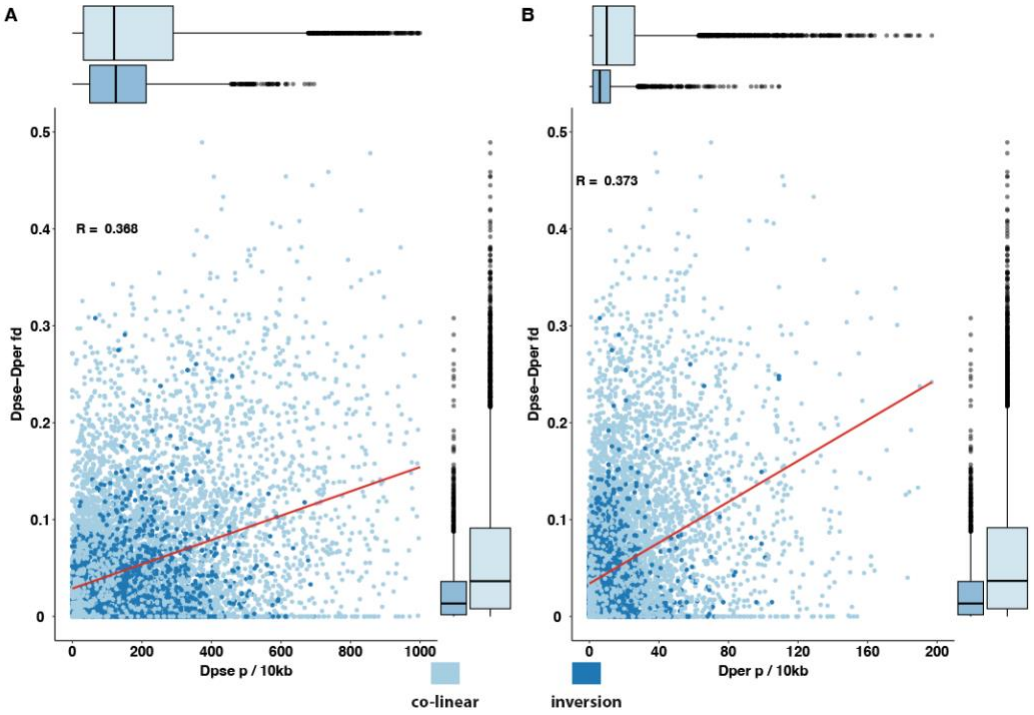


Figure 7. The relationship between recombination rate (*p*), admixture proportion (*fd*) between *D. pseudoobscura* and *D. persimilis.* Each point represents the value of *p* or *fd* for a single 10 kb window. Colors depict structural differences between *D. pseudoobscura* and *D. persimilis*. R values show significant correlation coefficients ($p < 0.05$), and the red lines represent the best model fit for each comparison. The correlation is also significant if data from inverted regions are not included. Margin boxplots show the comparison of *p* (top) or *fd* (side) values between inverted and co-linear regions. Dpse: *D. pseudoobscura*; Dper: *D. persimilis.*

Correlations between recombination rate and absolute divergence (Dxy) are positive and significant for both species pair comparisons (Figure 8A-C): *D. pseudoobscura* versus *D. persimilis* and *D. pseudoobscura* versus *D. p. bogotana*. The strongest correlation is observed for the allopatric *D. pseudoobscura* and *D. p. bogotana*. Conversely, correlations between recombination rate and Fst, a relative measure of divergence, are negative for the two species pair comparisons (Figure 8 D-F). The pattern of increasing Dxy with recombination rate seems

counter intuitive based on the observed positive correlation between admixture proportion and recombination (Figure 7). However, the pattern can be explained by the higher polymorphism levels expected in high recombination regions. Segregating alleles in high recombination regions are more likely to be ancestral, leading to an increase in Dxy at those regions. On the other hand, the observed significant reduction in Fst with increased recombination is the result of the expected lower nucleotide diversity in low recombination regions due to linked selection that inflates Fst values (Charlesworth, 1998). These results point toward the danger of using absolute or relative measures of divergence to identify genomic islands of divergence particularly when patterns of recombination are not taken into account (Cruickshank & Hahn, 2014; M. A. Noor & Bennett, 2009).

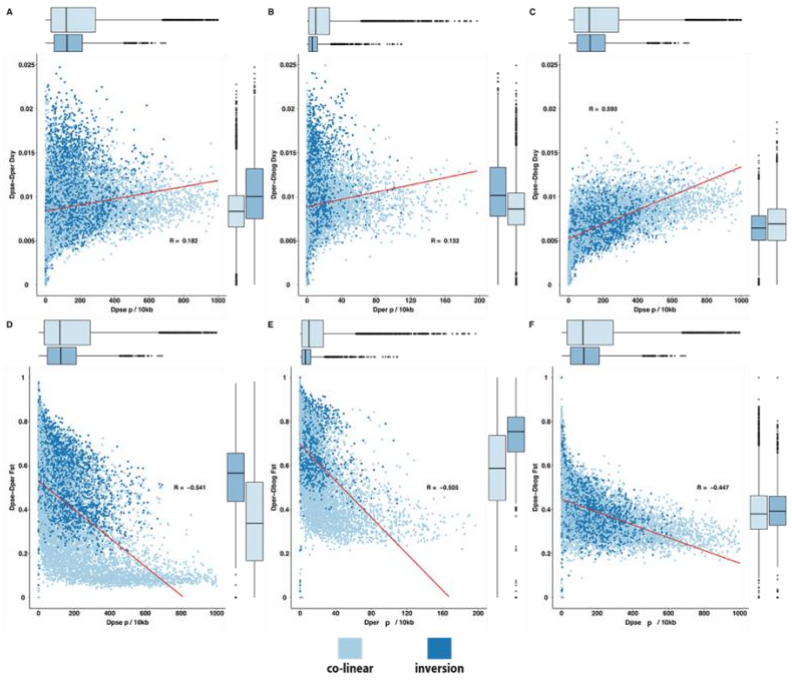


Figure 8. The relationship between Dxy (A,B,C), Fst (D,E,F) and recombination rates for *D. pseudoobscura* (Dpse), *D. persimilis* (Dper) and *D. p. bogotana* (Dbog). R values show significant correlation coefficients ($p < 0.05$) and the red lines represent the best model fit for each

comparison. Margin boxplots show comparisons between inverted and co-linear regions for each pairwise comparison.

Adaptive divergence in collinear regions of the genome occurs more often in regions of low recombination and low introgression.

We identified putative islands of genomic divergence (genomic outliers) between *D. pseudoobscura* and *D. persimilis* using the top 5% Dxy windows (596 outliers) and from results of the kNN analyses (261 outliers). Results suggests that chromosomes with inversion differences have more divergent regions that the rest of the chromosomes (Figure S26 Appendix B). We then used logistic regression to determine if outliers from collinear genomic regions are more likely to be found in regions of low recombination and low introgression (see methods). These analyses were confined to collinear regions of the genome because introgression has occurred preponderantly there, and these regions have not been thoroughly analyzed in other studies of this species group. Similar analyses were conducted between *D. persimilis* and *D. p. bogotana* and we treated those results as null model for similar divergence time with no recent introgression (after the split of *D. p. bogotana*).

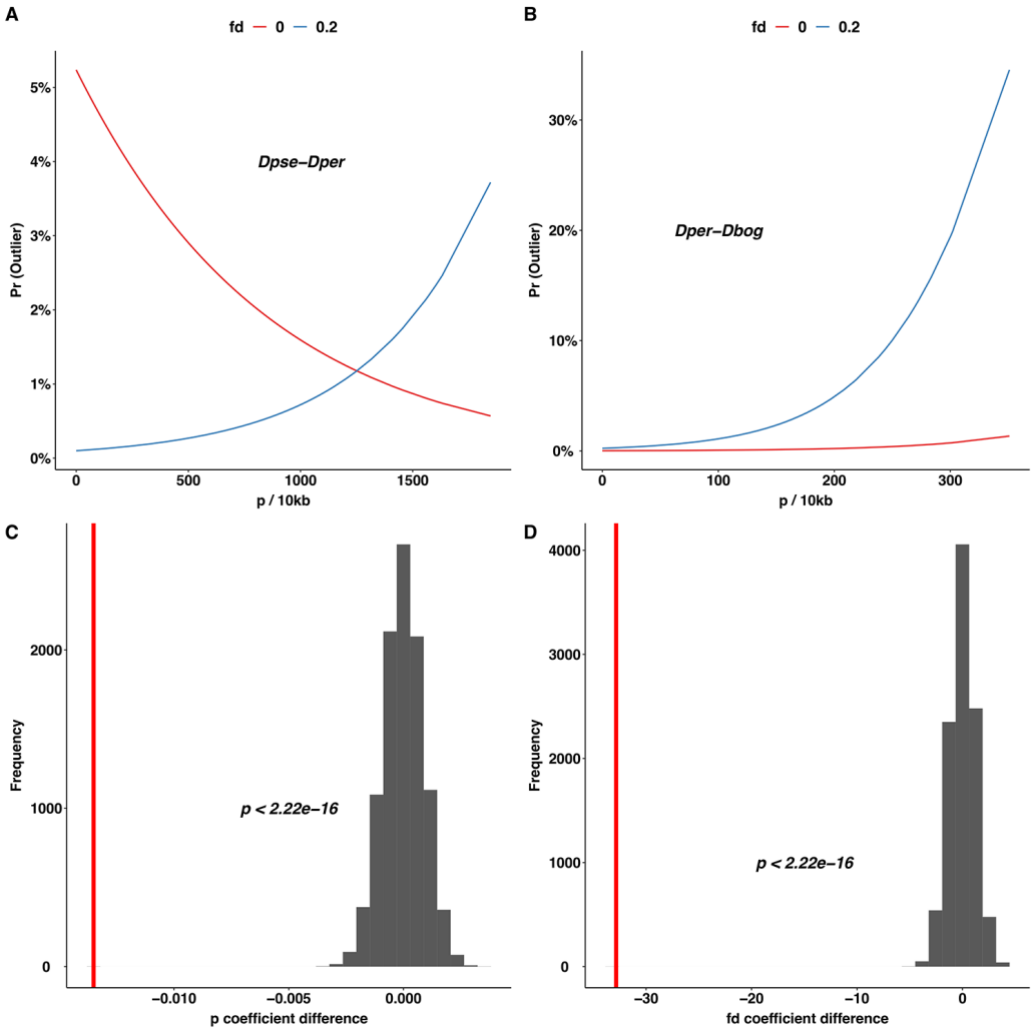


Figure 9. The tendency of the top 5% Dxy outliers to fall in regions of low recombination and admixture. Line plots show the probability of being an outlier as a function of the population recombination rate *p* at two *fd* levels: *fd = 0* and *fd* = 0.2. **A)** *D. pseudoobscura* vs *D. persimilis*; **B)** *D. persimilis* vs *D. p. bogotana*. Barplots show the null distribution of *p* (C) and *fd* (D,) coefficient difference values between the two species comparisons with 10,000 permutations. Two-sided p-values were calculated by calculating the fraction of null distribution values greater than the observed difference value and multiplying by two.

Table 1. Estimated regression parameters for the generalized linear model of the top 5% Dxy outliers between *D. pseudoobscura* and *D. persimilis*. Predictors of interest: p, fd and the interaction of both predictors. GD: Gene density; GC: GC content. Numbers in bold highlight coefficients of interest and significant p-values.

Dpse-Dper	***Estimate***	***Std. Error***	***z value***	**P-*value***
Intercept	-3.842	0.9224	-4.165	3.12E-05
p	**-0.0012**	0.0005	-2.202	**0.0276**
fd	**-20.02**	2.969	-6.743	**1.55E-11**
GD	0.1053	0.0476	2.212	0.027
GC	1.784	2.041	0.874	0.3821
p*:*fd	**0.016**	0.0051	3.133	**0.00173**

Table 2. Estimated regression parameters for the generalized linear model of the top 5% Dxy outliers between *D. persimilis* and *D. p. bogotana*. Predictors of interest: p, fd and the interaction of both predictors. GD: Gene density; GC: GC content. Numbers in bold highlight coefficients of interest and significant p-values.

Dper-Dbog	***Estimate***	***Std. Error***	***z value***	**P-*value***
Intercept	2.131	1.2589	-6.015	1.80E-09
p	**0.0122**	0.0074	1.638	0.101

fd	**12.8261**	1.1418	11.233	**< 2e-16**
GD	0.0816	0.0623	1.31	0.19
GC	-2.5717	2.7507	-0.935	0.35
p*:*fd	**0.0155**	0.0306	0.508	**0.612**

We found that Dxy outliers from collinear regions between *D. pseudoobscura* and *D. persimilis* are more likely to occur in regions of low recombination and low introgression between this species pair (Figure 9). Logistic regression models show significant negative coefficients for both *p* and *fd* (Table 1). Negative coefficients represent a reduction in the probability of being an outlier when the value of recombination rate increases. Although the comparison with *D. p. bogotana* shows the same tendency, correlation coefficients are positive and only significant for *fd* (Table 2). We observed similar trends in chromosomes where the signal of introgression between *D. pseudoobscura* and *D. persimilis* is high (Figure S21 Appendix B). These results suggest that genomic regions that have recently introgressed between *D. pseudoobscura* and *D. persimilis* are divergent between *D. persimilis* and *D. p. bogotana*. We note that the observed trend is exactly the opposite to the genome-wide trend from Figure 8 showing an increase of Dxy with recombination rate, reflecting the contribution of ancestral polymorphisms to the generation of Dxy outliers, and the need for incorporating recombination rate in the identification and exploration of genomic outliers.

As previously reported, genome scans that define outliers based on hard thresholds can be biased by recombination rate variation across the genome. Specifically, low recombination regions can generate Fst outlier regions under neutrality (Booker, Yeaman, & Whitlock, 2020). Thus, we implemented the approach of (Pfeifer et al., 2020) to detect outliers under selection

based on kNN techniques. Logistic regression models for kNN outliers show that recombination rate has a positive coefficient value, which suggest that at higher recombination rates the probability of being an outlier is also higher (Figure S29 Appendix B). However, for the null comparison with *D. p. bogotana* we observed the same result but with lower probabilities of being an outlier at low recombination rates and at moderate levels of introgression and higher probabilities at high recombination rates and at no signal of introgression than the *D. pseudoobscura*/*D. persimilis* comparison (Figure S29 Appendix B). In addition, model coefficients were positive for both comparisons, but only the null model was significant and greater than the *D. persimilis*/*D. p. bogotana* comparison (Tables S6,7 Appendix B). We also conducted permutation analyses to test if *fd* and *p* coefficients of the alternative model are significantly lower than the null model. As expected, results show that *fd* and *p* coefficients for the alternative model are significantly lower that the null model. These results also support the idea that outliers tend to be in regions of both low recombination rate and low signal of introgression.

To investigate if these patterns hold for each chromosome, we applied the same modeling approach with Dxy outliers to each chromosome independently. For this analysis we also excluded inversions and focus only on co-linear regions. Overall, we observed a tendency of Dxy outliers being on regions of low recombination and low introgression in chromosomes 2 and XL with significant difference in the respective coefficients for chromosome 2 and only *fd* for chromosome XL (Figure S20,22 Appendix B). For chromosome XR, we observed the same tendency of Dxy outliers but only the *fd* coefficient is significantly lower (Figure S23 Appendix B). Chromosome 4 shows an opposite tendency compared to the rest of the chromosomes where both the null and the alternative models indicate higher divergence on regions of high

recombination and high signal of introgression. Nonetheless, it appears that *p* is significantly lower in the *D. pseudoobscura* and *D. persimilis* comparison (Figure S24 Appendix B).

We also tested for significant association among genes under positive selection, chromosomes, inversion differences and the most represented weighted topologies (Figure 10). Here we considered all genes under positive selection pinpointed by the MKT and selective sweeps analyses (genes falling in genomic windows with evidence of selection were included in the analyses even if there was no evidence of adaptive protein evolution by the MKT). We observe a higher number of positively selected genes than expected in chromosomes 2, 4 and XL (Figure 10A). For chromosomes with inversion differences our results unveil a significant association of positive selection genes located with inversions in chromosomes 2 and XL only. Interestingly, most of the positively selected genes are significantly associated with the topology representing the species tree for all chromosomes, but we also observe some association with topology 2 and topology 3, which separate both *D. pseudoobscura* and *D. persimilis* (Figure 10C).

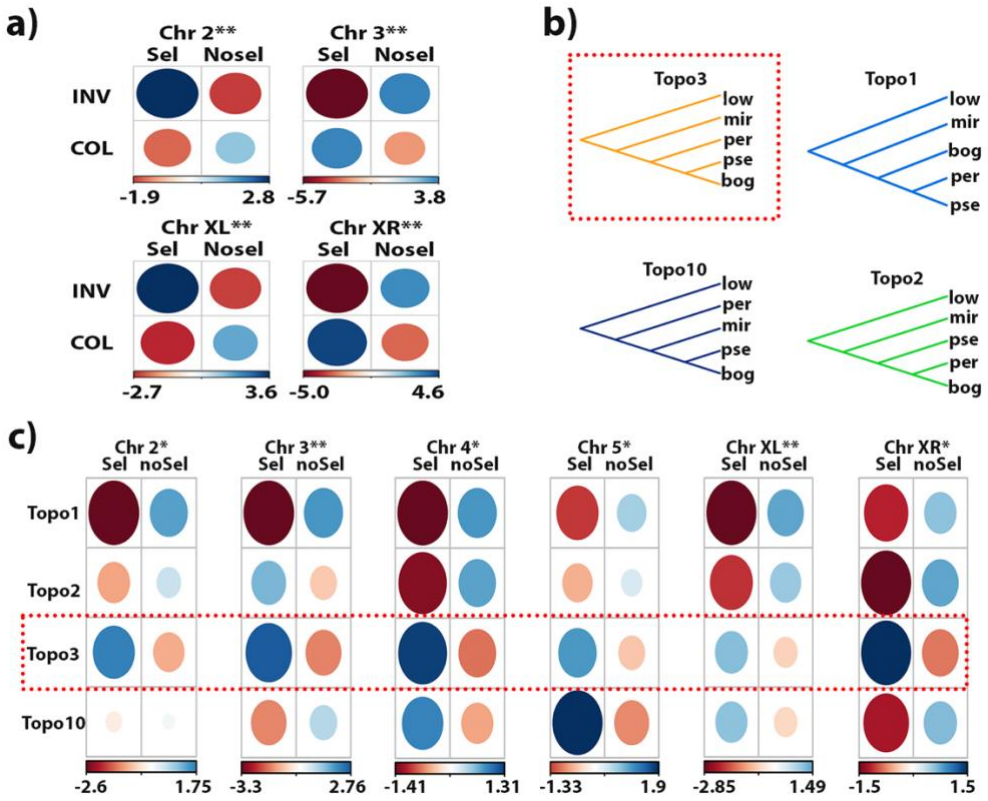


Figure 10. Positive selection is associated with species divergence. Test of independence showing significant association between genes under positive selection and chromosomal rearrangements (a), and the most represented topologies (b, c). Circle size depicts the contribution of that cell to the overall significance of the chi squared test and color shows the correlation strength between columns and rows of the contingency table. Sel = genes under selection; noSel = genes with no selection evidence. INV = inversion regions; COL = co-linear regions. *p < 0.05; ** p < 0.001.

DISCUSSION

Species divergence at the genomic level is a complex process where several factors, such as recombination rate and introgression, can come into play. Nonetheless, it is still poorly understood how the interaction of genomic intrinsic factors can drive species divergence when gene exchange is occurring between diverging lineages. Here we revisited patterns of divergence

between species of the *Drosophila pseudoobscura* group in the context of introgression and recombination rate. To our knowledge this is the first study in Drosophila that explicitly models the contribution of both recombination rate and introgression in generating islands of genomic divergence. In addition, we report new high quality genome assemblies for *D. pseudoobscura* and *D. persimilis* that we used to re-estimate patterns of divergence using a comprehensive population data set for each *D. pseudoobscura*, *D. persimilis* and *D. p. bogotana* species. These data add a valuable genomic resource for further comparative genomic studies.

Patterns of divergence revisited

We re-calculated genomic divergence statistics using a new HiFi genome assembly for *D. persimilis* as reference. Although patterns of divergence between *D. pseudoobscura* and *D. persimilis* have been extensively studied previously (Korunes et al., 2021; Machado et al., 2007; Machado & Hey, 2003; Machado et al., 2002; M. A. F. Noor et al., 2007), we decided to revisit such data not only by including a higher number of samples for each species but also by incorporating high quality genome assemblies from PacBio HiFi technology. We use *D. persimilis* as a reference because it was the most contiguous genome. Having a high-quality genome from HiFi reads helps to reduce mapping biases (Degner et al., 2009). In addition, to further control for mapping biases, we also used the *D. persimilis* genome assembly since the *D. pseudoobscura* and *D. p. bogotana* are closer to each other than these three species to *D. miranda*. Although conservative, we also controlled for calculation biases by excluding sites with unknown data. Hence, even though our results show similar patterns as previously described, we noticed that our Dxy estimates are lower than those reported by (Korunes et al., 2021). In addition, we also detected a more pronounced difference between co-linear and inverted regions in our Dxy estimates. Although these differences may not lead to different

conclusions about patterns of divergence of this species group, these can affect results where such estimates of divergence are used for downstream analyses.

Estimation of introgression and recombination rates

Patterns of genome-wide recombination rates have been estimated previously in this group by constructing genetic maps for single strains of both *D. pseudoobscura* and *D. persimilis* (Kulathinal, Bennettt, Fitzpatrick, & Noor, 2008; Samuk, Manzano-Winkler, Ritz, & Noor, 2020; Stevison et al., 2011; Stevison & Noor, 2010). These estimations were conducted using either the first release of the *D. pseudoobscura* genome assembly (Richards et al., 2005) or the release 3.04 deposited in FlyBase (Thurmond et al., 2019). This genome assembly belongs to *D. pseudoobscura* line MV-225, which is different than the one sequenced in this study. Hence, transferring such estimates to our genome assemblies results in a non-trivial task prone to errors. Furthermore, recombination maps are specific to the strain used and there is evidence of variation in recombination rates among populations and strains of *D. pseudoobscura* (Samuk et al., 2020). We estimated levels of recombination for both species using a model-based population genetic method to estimate genome-wide population recombination rates (Chan et al., 2012). Although LD-based methods to estimate population recombination rate can be biased if there is gene-flow between populations (Samuk & Noor, 2022), our estimations were performed at a species level, where both species appear to be panmictic (Machado et al., 2002). In addition, we used a large number of samples for each species, which is another important factor when using LD-based methods. Thus, we believe it is reasonable to think that our recombination rate estimations meet the basic assumptions of LD-based methods (Chan et al., 2012; Samuk & Noor, 2022).

We found that the recombination landscapes are very similar for both species, but rates of recombination are about an order of magnitude lower in *D. persimilis* (Figures 5,6; Figure S19 Appendix B). Chromosomal rearrangements can suppress recombination when segregating within a species, and this effect is quite clear in the middle of chromosome 3, which has the lowest overall level of recombination, and is known to harbor a well know inversion polymorphism in each species (Schaeffer et al., 2003). One intriguing finding was the lower estimate of recombination rate in the inverted region from chromosome 2 in both species relative to collinear regions of the same chromosome. This is intriguing because there is no polymorphism for these inverted regions within species; the fixed inversion in this chromosome arose and fixed in the *D. persimilis* lineage and it was polymorphic in this species probably for a short period of time and a long time ago (>1 Mya). This observation brings up the possibility of the evolution of genetic modifiers of recombination (Ortiz-Barrientos et al., 2016) linked to adaptive alleles spread through this inversion in both species.

We also revisited patterns of introgression occurring between *D. pseudoobscura* and *D. persimilis*. As expected, and in agreement with previous estimates (Korunes et al., 2021), we observed that levels of introgression are significantly higher in co-linear regions compared to inversions. We observed, nonetheless, some discrepancy on the levels of introgression for the inversion of the XL chromosome. For example, (Korunes et al., 2021) reported that a peak of introgression in the middle of the fixed inversion in the XL chromosome, whereas we do not observed such peak (Figure 4). Moreover, results from the topology weighting approach (Martin et al., 2019; Martin & Van Belleghem, 2017) show that the topology supporting introgression events between *D. pseudoobscura* and *D. persimilis* (topology 1) is the third most represented in this genomic region, a pattern that is different from the rest of the chromosomes (Figure S18

Appendix B). Although we detect signal of introgression in the X chromosome, the two approaches implemented in this study suggest this chromosome harboring the strongest species barriers. Thus, we believe that implementing multiple approaches can help to further elucidate evolutionary process occurring between species.

The interplay between recombination rate and introgression in driving the divergence of *D. pseudoobscura* and *D. persimilis*.

In this study we explicitly tested the hypothesis that recombination and introgression shape the genomic divergence landscape between *D. pseudoobscura* and *D. persimilis.* At early stages of speciation genomics research, efforts were focused on identified genomic regions with evidence of high differentiation across populations or species. The initial assumption was that highly divergent regions (detected with the Fst statistic) are the result of strong species barriers driven by selection and a reduction of gene flow (Barrett & Hoekstra, 2011; Nosil & Feder, 2012). Nevertheless, studying the genomic landscape of speciation to detect evidence of adaptive divergence requires the consideration of several evolutionary forces. Disentangling the effects of each force is not a trivial task since many factors can generate confounding signals of divergence between species (Cruickshank & Hahn, 2014; Ravinet et al., 2017). Factors such as recombination rate, gene flow, gene density and selection are amongst the most important factors that shape the genomic patterns of divergence populations (Ravinet et al., 2017). For example, in the absence of gene flow peaks of differentiation can be the result of background selection or selective sweeps. However, identifying which have become species barriers is not trivial since only those regions preventing gene flow can act as potential barriers (Charlesworth, Morgan, & Charlesworth, 1993; Cruickshank & Hahn, 2014; M. A. Noor & Bennett, 2009). Here we

compared divergence patterns across two scenarios of divergence to test the tendency of finding divergent loci on regions of low recombination and low signal of introgression.

Understanding the relationship among recombination rate, introgression and selection in shaping species divergence has drawn more attention in the recent years (Aeschbacher et al., 2017; Samuk et al., 2017). However, to our knowledge, this is the first attempt to shed light on the contribution of introgression and recombination rate in shaping the genomic divergence in Drosophila, and certainly between *D. pseudoobscura* and *D. persimilis.* The contribution of recombination rate to sequence divergence has been extensively studied (Ravinet et al., 2017), but it is still not clear at what extent recombination rate and introgression variation across the genome influences the probability of generating divergent regions. It is predicted that signals of species barriers could occur in regions in low recombination with high levels of genetic differentiation (Fst) and divergence (Dxy) (Cutter & Payseur, 2013; Nachman & Payseur, 2012; Stephan, 2010). In this study we demonstrate that in fact, regions of low recombination have a higher probability of being divergent regions than regions with high recombination rate, which appears to be also influenced by the action of introgression between *D. pseudoobscura* and *D. persimilis*. Although recent studies have shown that even under neutrality, regions of low recombination can generate peaks of genetic differentiation and divergence (Booker et al., 2020) between populations. Our implementation of kNN techniques can account for those recombination rate effects given that it is able to detect local divergent outliers and does not depend on hard thresholds coming from genome-wide estimates. Moreover, the use of Dxy to detect divergent regions is also less susceptible to the effects of background selection, which can inflate other Fst given the reduction of nucleotide diversity (Cruickshank & Hahn, 2014; Nachman & Payseur, 2012). Still, we believe that our results can be complemented by simulation

analyses that consider recombination rate variation and introgression, which would provide further support to the identified potential species barriers between *D. pseudoobscura* and *D. persimilis.*

We report that, in colinear regions of the genome, significantly divergent genomic regions ("islands of genomic divergence") between these species tend to be located on regions of low recombination and low introgression. Although inversion differences between *D. pseudoobscura* and *D. persimilis* harbor the loci of major effect involved in hybrid incompatibilities (M. A. Noor et al., 2001; M. A. F. Noor et al., 2001), our results show that co-linear regions also harbor genomic outliers in regions where recombination rate and introgression are low. These results suggest that the genetic architecture of species divergence between these species also includes loci in collinear regions of the genome. Identifying and understanding the contribution of those loci to the genetics of species divergence should be the focus of future research in this species group.

ACKNOWLEDGEMENTS

I thank the members of the Machado Lab for helpful discussions and technical advice. Research supported by National Science Foundation grants MCB-1716532 and DEB-1754572 to Carlos Machado.

CHAPTER 3: Gene co-expression network re-wiring between *Drosophila pseudoobscura* and *D. persimilis*

ABSTRACT

Network analysis techniques provide a useful novel approach to study species divergence as they may help elucidate evolutionary forces acting on genes from common pathways. Most speciation genomics studies have focused on identifying speciation drivers based on patterns of sequence divergence. Nonetheless, comparing species using gene co-expression networks may provide novel and important insights into the process of species divergence. Here we present and compare robust gene co-expression networks from two closely related species of Drosophila and explore co-expression divergence between species. We use a comprehensive RNA-seq data set to construct independent networks for each species to identify modules of genes with conserved and divergent co-expression profiles. Our network comparisons reveal significant module conservation between species. At a finer scale, genes with conserved co-expression interactions in both species exhibit higher connectivity densities, lower sequence divergence and lower levels of differential expression compared to genes with species-specific co-expression interactions. We report strong signals of selection acting on genes with divergent co-expression profiles between species, although network co-expression rewiring also occurs in genes with no evidence of selection or differential expression. Overall, our results suggest a strong interplay between sequence divergence and gene co-expression divergence highlighting the relevance of network analysis techniques to identify drivers of species divergence not identified by conventional methods.

INTRODUCTION

Studying species divergence and speciation requires the integration of multi-omics data to fully unveil fixed genomic changes between species and their functional effects on species divergence. Speciation can be studied from different angles (Cooper & Phadnis, 2016; Orr, 1995; Pavey, Collin, Nosil, & Rogers, 2010; Nitin Phadnis & Orr, 2009; Pinho & Hey, 2010; Presgraves, 2010; Ravinet et al., 2017) but studying speciation using genomic scale data has become one of the most common approaches. The usual approach is to compare multiple genomes of different species and scan for highly differentiated regions often considering intrinsic genomic properties such as the rate of genomic recombination and/or demographic parameters (Choi, Purugganan, & Stacy, 2020; Cruickshank & Hahn, 2014; Izuno et al., 2022; Saetre, 2014; Wolf & Ellegren, 2017). The result of such approaches often reveals loci that are highly differentiated between the focal species. Highly divergent loci between species typically unveil clues about the molecular processes involved in species divergence. Although this basic approach has been successful at pinpointing candidate genes and genomic regions involved in species divergence, we still need a unified framework to connect those loci with interacting loci across the genome to unveil patterns of genomic divergence in a functional context.

One potentially useful novel approach to study species divergence and speciation is to use network analyses (Bertranpetit et al., 2015; S. Chakraborty & Alvarez-Ponce, 2016; Filteau, Pavey, St-Cyr, & Bernatchez, 2013; Hu et al., 2016; Schwarzer, Misof, & Schliewen, 2012; Tsaparas, Marino-Ramirez, Bodenreider, Koonin, & Jordan, 2006). A network is a structure representing relationships among objects, and network analysis helps us understand those relationships. In biology, network analysis has been applied across multiple areas, from spatial

ecology to studies of the physical interactions between molecules inside a cell (Fortin, Dale, & Brimacombe, 2021). Hence, in the past decade, network theory has been widely used to identify genes or clusters of genes involved in complex phenotypes (Costanzo et al., 2019). For example, network analysis has been applied to cancer data to find genes or clusters of genes that act as key regulators in cancer phenotypes (Creixell et al., 2015; Reyna et al., 2020). Similarly, network analyses have also been applied to study gene essentiality in terms of the number of interactions where a correlation exist between the number of interactions and the rate of protein evolution (Alvarez-Ponce, Feyertag, & Chakraborty, 2017; Bertranpetit et al., 2015; S. Chakraborty & Alvarez-Ponce, 2016; Jalili et al., 2016). Moreover, network analyses have also been useful in predicting Chip-Seq targets and survival indicator genes in functional genomics experiments (Jiang et al., 2015).

Although less common, network analysis has also been applied to compare networks between species and study the rewiring of gene clusters to find key drivers of evolutionary change (Ovens, Eames, & McQuillan, 2021). The use of gene co-expression networks within species comparisons have been used to identify sex-biased gene clusters across development (Rago, Werren, & Colbourne, 2020). Topological properties of gene co-expression networks of different brain tissues have also been compared between humans and chimpanzees (Oldham, Horvath, & Geschwind, 2006), finding that gene cluster conservation is significantly weak in the brain between species even though all analyzed genes were orthologous. More recently, comparisons between humans and mice revealed that co-expression connectivity is negatively correlated with sequence divergence (Monaco, van Dam, Ribeiro, Larbi, & de Magalhaes, 2015). This highlights the potential relationship of sequence divergence and network rewiring between species. In addition, a combination of protein-protein interactions and gene co-expression meta-

analyses across phyla suggest that the probability of young genes to create new connections to other genes decreases with phenotypic complexity (W. Wei et al., 2016). Thus, the implementation of network theory underlines its importance for unlocking hidden patterns to explain complex biological phenomena. In the context of speciation, network analyses can help elucidate evolutionary forces targeting groups of genes from common pathways, allowing to place the process of species divergence in a functional context.

One classic model for studying species divergence is the *Drosophila pseudoobscura* subgroup which includes the pair of closely related species *D. pseudoobscura* and *D. persimilis*. Reproductive isolation among these species is not complete; although F1 hybrid females are fertile, F1 hybrid males are sterile regardless of the direction of the cross. These two species are partially sympatric in the Western US (*D. pseudoobscura* has a larger geographic range), and there is evidence of low rates of hybridization occurring in nature (Theodosius Dobzhansky, 1973) and of a pattern of historical introgression based on population genetic and genomic data (Korunes et al., 2021; Machado & Hey, 2003; Machado et al., 2002).

The genome of these species is organized in 5 chromosomes (2, 3, 4, 5, and X, which is divided in chromosome arms XL and XR), and patterns of genomic variation have revealed that divergence between *D. pseudoobscura* and *D. persimilis* is heterogeneous across the genome (Korunes et al., 2021; Kulathinal et al., 2008; Kulathinal et al., 2009; Machado et al., 2002; M. A. F. Noor et al., 2007). The most important genomic differences between these species are fixed chromosomal inversions in 3 chromosomes (XL, XR, 2) that are derived in *D. persimilis* and that capture most genetic differences involved in species divergence phenotypes (Anderson et al., 1977; Machado et al., 2007; Moore & Taylor, 1986; M. A. F. Noor et al., 2007; Schaeffer et al., 2008; Tan, 1935).

Divergence patterns between these species have been described extensively and have revealed that loci located in regions with introgression barriers are highly divergent between species (Korunes et al., 2021). Introgression barriers between these species are stronger in regions with inversion differences, which also correlates with the higher divergence levels occurring in the same regions (Korunes et al., 2021; Machado & Hey, 2003; Machado et al., 2002; M. A. Noor et al., 2001; M. A. F. Noor et al., 2001). In addition, recent results suggest that selection is a key driver of adaptive divergence in these species, which mostly acts on introgression barrier loci (see Chapter 2). Although divergence patterns between *D. pseudoobscura* and *D. persimilis* have been widely studied, the functional context of species divergence has yet to be elucidated. Unveiling the patterns of interactions among divergent genes and their interacting genes is pivotal to understanding the contribution of genomic divergence to gene co-expression divergence.

In Drosophila, network theory has been applied to identify clusters of genes responsible for complex phenotypes and to predict potential function to unknown or novel genes within and between species (Lau et al., 2020; Marco, Konikoff, Karr, & Kumar, 2009). As networks have been constructed mostly for *D. melanogaster*, interaction data for other Drosophila species is still limited and little is known about the evolution of gene co-expression in a network theory framework.

Here we present robust gene co-expression networks for both *D. pseudoobscura* and *D. persimilis*. The networks presented here are constructed from a comprehensive RNA sequencing data set that includes four developmental time points and tissue-specific samples for both males and females. To our knowledge this is the first report of such networks for Drosophila species other than *D. melanogaster* and, more importantly, it is also the first study comparing networks

from different Drosophila species. Using these data, we link differential gene expression and sequence divergence across species to test the hypothesis that expression and sequence divergence between species are associated with network rewiring. In addition, given that gene connectivity appears to be correlated with the probability of being under purifying selection (Alvarez-Ponce et al., 2017; Bertranpetit et al., 2015; S. Chakraborty & Alvarez-Ponce, 2016), we hypothesize that natural selection is associated with gene essentiality on species specific interactions (Rancati, Moffat, Typas, & Pavelka, 2018).

MATERIALS AND METHODS

Orthologous gene identification

Orthologous genes between *D. pseudoobscura* and *D. persimilis* were established using OrthoFinder v2.4.0 (Emms & Kelly, 2019) with default parameters. In addition, *D. melanogaster* genes were also included to transfer gene names and GO annotations to homologous genes of the pseudoobscura subgroup. Strictly 1-to-1 orthologous were retrieved and used for both differential expression and network analyses.

RNAseq samples

Four developmental stages samples were collected for 5 different lines of both *D. pseudoobscura* and *D. persimilis* (Table S2 Appendix C), for both whole-body males and females using the approach described in (Nyberg & Machado, 2016). The four developmental stages included: first instar larvae (1L, not sexed), third instar larvae (3L, sexed), mid-stage pupae (Pup, sexed) and 6-day post-eclosion adults (Ad, sexed). Non-adult samples were sexed by PCR using Y-chromosome primers (Nyberg & Machado, 2016). Tissue-specific samples (ovary,

testes, carcass) were also collected for 5 different lines of *D. pseudoobscura* and 5 lines of *D. persimilis*. All samples had two biological replicates. A total of 75 female and 77 male samples were collected for *D. pseudoobscura* and 62 female and 62 male samples were collected for *D. persimilis*.

For the construction of gene co-expression networks, expression levels were quantified using salmon v1.5.2 (Emms & Kelly, 2019). Raw quant files were processed using tximport v1.24.0 (Soneson, Love, & Robinson, 2015) to obtain a final matrix with raw counts containing all the samples for each sex group for each species. Normalization and data transformation of raw counts were conducted following the recommendations of (Johnson & Krishnan, 2022) for constructing gene co-expression networks. Briefly, first lowly expressed genes were removed and remaining raw counts were normalized to counts adjusted with TMM factors (CTF) and resulting normalized data was transformed using the hyperbolic arcsine (asinh) transformation. To make networks comparable between species, genes with strictly 1-to-1 orthologous relationships between species were selected to be included in the final networks (see below). Normalized and transformed expression matrices were then used as input for the co-expression network construction.

Differential expression analyses

Since transcript expression quantification was conducted for each species using their corresponding genome annotations, 1-to-1 orthologous genes were only considered for differential expression analyses between species. Species-level differential expression analyses were conducted for each developmental stage independently and for tissue specific samples. Differential expression between species was assessed for males, females, ovaries, testis and carcasses using DESeq2 v1.30.1 (Love et al., 2014) (design = ~species). Significant differential

expression for each comparison was considered for genes with log2FoldChange > 1.5 and padj < 0.005.

Network construction

Adjacency matrices for both species were constructed using the expression data of 8,680 and 9,903 genes for females and males, respectively. Final adjacency matrices were constructed using a consensus approach as described in (Monti, Tamayo, Mesirov, & Golub, 2003; Shahan et al., 2018; L. F. Wu et al., 2002). Briefly, subsamples of genes were conducted 1,000 times with randomized parameters for network construction and module identification for each iteration: power transformation, minModuleSize, and merge on eigengenes. Then a consensus adjacency matrix was constructed considering the number of times gene *i* is clustered with gene *j* divided by the number of times gene *i* is subsampled with gene *j* (Shahan et al., 2018). The idea of a consensus adjacency matrix is to ensure module reproducibility and to reduce spurious clustering. All the consensus network construction pipeline was implemented using modified R scripts from https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/ and (Shahan et al., 2018).

Module identification and module preservation analyses

Groups of highly interconnected genes, called modules, were identified based on co-expression patterns across samples using the WGCNA package. Using the consensus adjacency matrices for each species for each sex, hierarchical clustering was implemented for each matrix using the hclust function (method="average").

Hierarchical dendrograms produced by hclust were used as inputs to implement the module identification analysis using the cutreeDynamic function (deepSplit=2,

pamRespectsDendro=FALSE, minClusterSize=100), which implements an adaptive branch pruning algorithm. After having all modules identified for each adjacency matrix, modules with similar expression patterns were merged into the same module using the mergeCloseModules function (cutHeight=0.25). Metadata information on developmental stage and tissue were incorporated to assign module-trait relationships to each module for each network. Final modules are represented by colors independently for each of the four networks and same color name for more than one network does not represent similarity.

To assess module correspondence, module overlap was conducted using the function overlapTable, which calculates gene overlap counts and assess significance using the Fisher exact test for two module sets. For this analysis, comparisons were made for *D. pseudoobscura* vs *D. persimilis* for each sex data set independently.

To test if modules across species are conserved, we calculated the Zsummary and median rank statistics for each module. These statistics estimate module preservation based on network connectivity and density statistics across species. Module preservation is typically used to asses module reproducibility across different groups of samples of the same condition (Langfelder, Luo, Oldham, & Horvath, 2011). Nonetheless, this type of analysis is also suitable to test if modules are preserved across species (Du et al., 2021; Oldham et al., 2006; Ovens et al., 2021; Pembroke, Hartl, & Geschwind, 2021). Since our sampling in both species is similar, we tested for module preservation between *D. pseudoobscura* and *D. persimilis* comparing males and females separately.

Median rank and Zsummary statistics were calculated in a two-way fashion: first using *D. pseudoobscura* modules as a reference and then using *D. persimilis* as a reference. In addition, a consensus gene co-expression network was constructed for each sex to identify consensus

modules between species. Consensus module analyses were implemented with adapted scripts from

https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/

Differential co-expression analyses and gene essentiality

Differential gene co-expression was analyzed using the package csdR v1.2.0 (Pettersen & Almaas, 2022), which implements the CSD approach (Voigt, Nowick, & Almaas, 2017). CSD is a composite approach that aims to categorize gene pair co-expression comparing two conditions. For the present work, differential co-expression was assessed between species for each sex independently. Under the CSD approach, a pair of co-expressing genes are conserved when they co-express in both species. Species-specific or divergent co-expression occurs when significant co-expression between the two genes is observed in one species but not in the other. Lastly, divergent co-expression (or "conserved-divergent") occurs when a pair of genes co-express in one species but with a significant negative correlation on the other condition/species.

Using the orthology relationships, essentiality status for each gene in the *pseudoobscura* subgroup was transferred from assessments in *D. melanogaster*. Gene essentiality classes were retrieved from the OGEE database (W. H. Chen, Lu, Chen, Zhao, & Bork, 2017; W. H. Chen, Minguez, Lercher, & Bork, 2012). Gene essentiality classification from the OGEE database includes four classes: essential (E), non-essential (NE), essential conditional (C) and undetermined.

Hub identification

Global hubs were considered as the top 5% of genes with the highest kTotal values. kTotal is a measure of connectivity based on the correlation values of the whole network. Global hubs were identified for each network. In addition, module hubs were also retrieved by selecting

the top 5% genes with the highest kWithin and module membership values. kWithin is a measure of intramodular connectivity while module membership represents the correlation of the gene expression profile with the module eigengene of a given module. The top 5% genes with the highest values of gene significance for development and tissue gene expression profiles were also considered for hub classification.

RESULTS

Module similarity across species

Analyses of our novel co-expression networks for *D. pseudoobscura* and *D. persimilis* show that the number of modules is similar but not identical between species. We first generated four adjacency matrices representing single networks for males and females for each Drosophila species. A total of 8,680 1:1 orthologous gene for female samples and a total of 9,902 1:1 orthologous genes for male samples were included in the analyses after considering only orthologous genes with expression signal in all samples for both species. Adjacency matrices were used to identify gene modules (groups of genes showing coordinated expression across samples) and compare their preservation between species. For females, WGCNA identified a total of 39 and 36 initial modules for *D. pseudoobscura* and *D. persimilis* respectively. After merging modules whose expression profiles were similar across samples and conditions, a total of 12 and 10 final modules were identified for *D. pseudoobscura* and *D. persimilis* females, respectively (Table S1; Figure S1 Appendix C). For males, WGCNA identified 38 and 35 initial modules for *D. pseudoobscura* and *D. persimilis*, respectively. After the merging step, a total of 11 and 9 final modules were identified for *D. pseudoobscura* and *D. persimilis* males,

respectively (Table S1; Figure S1 Appendix C). The final merged modules are labeled with different colors, but the same colors across samples (species, sexes) do not imply the same module identity. Although we observe a lower number of modules in *D. persimilis* for both sex-specific networks, the overall number of modules is similar in both species. Interestingly, for males we observe one less module than females in both species even though the number of expressed orthologous genes is substantially higher. These observed slight differences in modularity suggest network rewiring occurring across species.

The gene overlap of modules across different species and the expression profiles of those modules are significantly correlated with the expression profiles observed across various developmental stages and tissues. We tested if modules detected for each network show significant ($p < 0.05$) overlap between species using the overlapTable function of WGCNA. In terms of the number of shared genes across modules, results show significant overlap between species in 11 modules of *D. pseudoobscura* females and 9 modules of *D. persimilis* females (Figure 1). The modules showing the least significant overlap between species are the purple modules for both species, which could represent species-specific modules. For males, we observe that all modules in *D. pseudoobscura* show significant overlap with at least 1 module in *D. persimilis*. Only one module from *D. persimilis* males appears to be species specific (black module).

Module-trait relationships were then assessed to determine how module expression profiles correlate with expression profiles across developmental stages, tissue, and species lines. We did not observe significant correlation between modules and species lines, for both species and both sexes (Figure S1 Appendix C). This suggests lack of significant intraspecific variation in gene co-expression profiles. We also observed that the largest modules in all networks

significantly overlap across species for both sexes (Figure 1). These large modules show significant correlation ($p < 0.05$) with both developmental stage and tissue expression profiles for all networks. However, we noticed inconsistent patterns in the smallest modules (e.g. purple modules, see Figure 1). The purple module in *D. pseudoobscura* females (153 genes) shows positive correlation with developmental stages but negative correlation with tissue. In *D. persimilis* females, however, the purple module (123 genes) shows negative correlation for both traits. Although the number of modules is different between species, significant gene overlap is still observed across most modules. Our results also suggest that modules with the highest degree of overlap share similar expression profiles across development and tissue in both species (Figure S1 Appendix C).

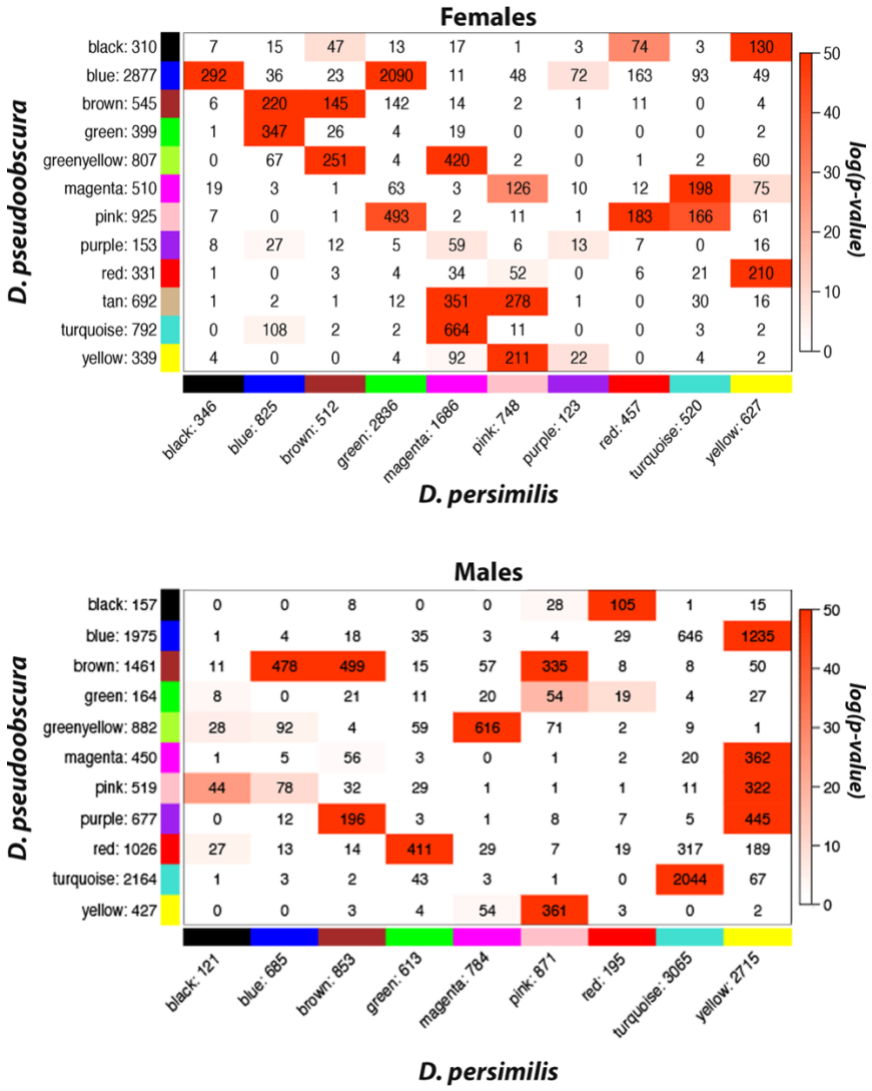


Figure 1. Module correspondence for both females (top) and males (bottom) between *D. pseudoobscura* and *D. persimilis*. X-axis shows *D. persimilis* modules and Y-axis shows *D. pseudoobscura* modules. Number of genes for each module are shown besides its name (color). Numbers in each cell represent the number of common genes between modules. Heatmaps represent the degree of overlap between modules between species. Cell colors represent the -log(p-value): white (N.S; $p = 1$); dark red (extremely significant; $p < 0.0001$); Fisher exact test.

Co-expression networks are conserved between species

Module preservation analyses reveal gene co-expression conservation between *D. pseudoobscura* and *D. persimilis*. Using *D. pseudoobscura* as reference, we find that the Zsummary statistic is significant (> 10) for all female modules, suggesting that all *D. pseudoobscura* modules have high probability of being present in the *D. persimilis* co-expression network (Figure 2; Table S3 Appendix C). For males, the Zsummary statistic is also significant for all modules except one (green, Zsummary = 8.6). This lower Zsummary value indicates that the green module is only moderately preserved between males of both species (Figure 2, Table S4 Appendix C) (Langfelder et al., 2011). None of the modules present Zsummary values below 2, which indicate that all identified modules are preserved between species and have a low probability of being the result of technical artifacts.

Module conservation evaluated with the media rank statistics further supports module conservation between species. Nevertheless, the black module in males appears to be more conserved than the value suggested by the Zsummary statistic (Figure 2; Table S3 Appendix C). This discrepancy can be explained by the tendency of Zsummary to be affected by module size. It is noteworthy, however, that the green module remains one of the least conserved modules in males (Figure 2; Table S4 Appendix C).

Discrepancies observed in the two least conserved modules reflect patterns of gene overlap across modules between species. While most of the genes in the black module of *D. pseudoobscura* males show significantly overlap with the red module of *D. persimilis*, the green module of *D. pseudoobscura* males shows significant gene overlap with at least three modules in *D. persimilis* (Figure 1).

As expected, reciprocal analyses using *D. persimilis* as a reference suggest similar patterns of module conservation between species (Figure S2 Appendix C). In addition,

Zsummary and median rank statistics suggest that modules purple and yellow are the least preserved modules in females. Moreover, modules red and black appear to be the least preserved modules in males (Figure S3 Appendix C). Overall, preservation statistics suggest significant module conservation not only in terms of the number of connections but also in the connection similarity in both species.

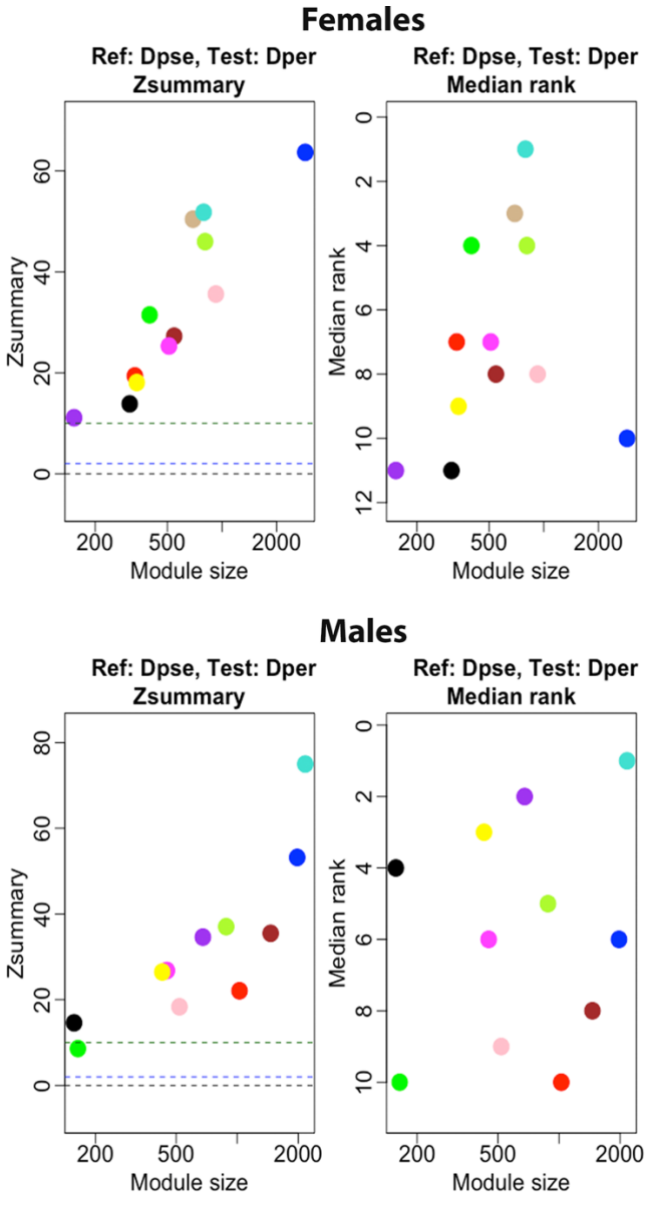


Figure 2. Module preservation analysis for female and male networks using *D. pseudoobscura* as a reference. Module preservation Zsummary and Median rank summary statistics for females (top) and males (bottom) networks. Each colored dot represents a single module. Zsummary horizontal lines indicate preservation thresholds: black (Zsummary = 0); blue (Zsummary = 2); green (Zsummary =10). Similar results using *D. persimilis* as reference are shown in Figures S2 and S3 Appendix C.

Differential co-expression reveals connectivity differences within and between species.

To further compare patterns of gene co-expression, we characterized species-specific and conserved gene co-expression links across species. We implemented a module consensus analysis to create consensus modules across species. The rationale for this approach is to create new modules that reflect patterns of co-expression occurring in both species with the purpose of reducing species-specific co-expression bias. We find that all genes in the female networks have a consensus counterpart (Figure S4 Appendix C). In addition, we observed that consensus module eigengenes show similar direction of correlation with the traits considered in the analyses (Figure S5 Appendix C). We only observed, however, opposite correlation between species in three module eigengenes for both males and females (Figure S5 Appendix C). These 3 module eigengene have significant overlaps with modules that show signs of low preservation between species (Figure S4 Appendix C). For example, in females the consensus black module does have significant gene overlap with the purple module in *D. persimilis*, which is the least preserved module in this species (Figure S4 Appendix C). In males, the consensus midnight blue module significantly overlaps with the green module in *D. psedudoobscura*, which is also the least conserved in this species (Figure S4 Appendix C).

Gene expression profiles in consensus modules show similar correlation patterns with development and tissue across species. We further assessed congruence of consensus modules across species by evaluating the correlation of species-specific expression profiles with development and tissue. Evaluation of module eigengene expression profiles for each species suggest congruency in correlation patterns between species (Figure S5 Appendix C). In females, both species show similar correlation patterns across development and tissue for most consensus modules (Figure S5 Appendix C). The only exceptions are modules black, grey60 and darkorange where *D. pseudoobscura* shows positive correlation with development while in *D. persimilis* the correlation is negative (Figure S5 Appendix C). In males, discrepancies are observed in consensus modules purple, and blue. These results indicate that consensus modules capture similarities in gene co-expression across species, which highlight the validity of their use to measure differential co-expression profiles across species.

We used the consensus modules as reference to run differential co-expression analysis for each module independently (see methods). As in (Pettersen & Almaas, 2022; Voigt, Nowick, & Almaas, 2017), we classified network edges based on significant correlation among gene pairs (see methods). We classified genes in three main classes: conserved, divergent (or species-specific) and conserved-divergent. The conserved class contains genes that have co-expression links that are observed in both species' networks. The divergent or species-specific class contains genes with co-expression links in one species but strong negative correlation in the other. The conserved-divergent class contain genes that have conserved and species-specific co-expression links between species (see methods).

Differential co-expression results reveal connectivity differences across conservation classes between species. After running csdR, we found a total of 71 divergent, 1,570 conserved-

divergent and 5,224 conserved genes in females. For males, we found 124 divergent, 2,109 conserved-divergent and 4,413 conserved genes. To investigate differences across co-expression classes and between species, we evaluated if the number of connections for each gene significantly differs across conservation classes and between species. Considering global network connectivity (KTotal) we observed significant difference across conservation classes both within and between species and between sexes (Figure 3A,B). For conservation class, we observe higher global connectivity in *D. pseudoobscura* than in *D. persimilis*. In addition our results indicate a higher average connectivity for genes with conserved co-expression across species.

We also explored global connectivity across different essentiality classes retrieved from *D. melanogaster* (see Methods). Although we assume that essentiality definitions in *D. melanogaster* hold in our species pair, we still observed significant differences in KTotal when genes are grouped by essentiality classes (Figure S6 Appendix C). Essential genes overall exhibit higher connectivity levels than non-essential genes. Hence, these results indicate that co-expression conservation can be stronger for genes having higher number of global interactions in the network. Moreover, the lower number of global connections observed in *D. persimilis* can be the associated with the lower number of modules constructed.

Results from measures of intramodular connectivity indicate significant differences in KWithin across co-expression conservation classes (Figure 3C,D). We observe that genes with both types of co-expression connections show the highest values of intramodular connectivity. Moreover, contrary to what is observed for kTotal, KWithin appears to be higher in *D. persimilis* than in *D. pseudoobscura*. Similar to KTotal we observed that genes with both types of co-expression connections also exhibit the highest values of intramodular connectivity (Figure

3C,D). Higher intramodular connectivity in *D. persimilis* is expected since this species has a lower number of modules, which translate into a larger number of genes per module. Similar to KTotal, KWithin comparisons for gene essentiality classes show higher levels for essential genes than non-essential (Figure S6 Appendix C). Overall, our results indicate that intramodular connectivity is associated with co-expression conservation across species.

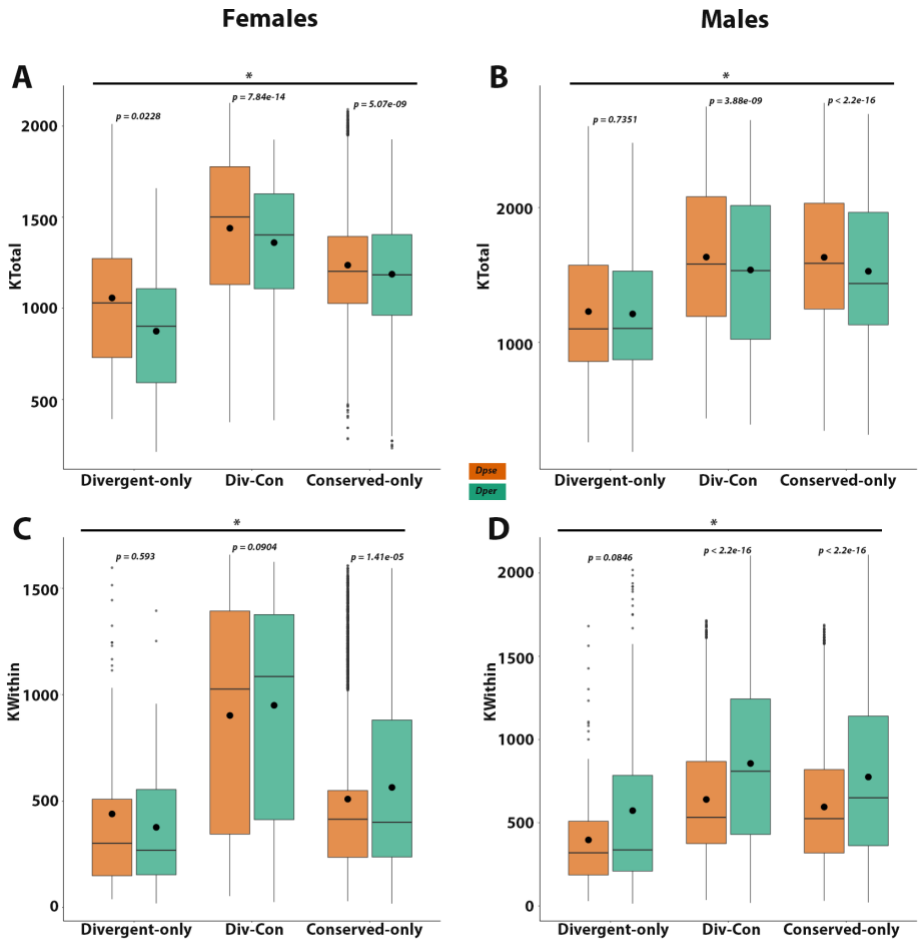


Figure 3. Gene connectivity comparisons across three edge classes between species. Panels A and B show KTotal, which is a connectivity measure based on the whole network connectivity. Panels C and D show KWithin, which is an intramodular connectivity measure. P-values (Wilcoxon rank sum test) are shown for each species comparison across edge classes. The top horizontal line with asterisk indicates significant differences across edge classes for each

species ($p < 0.001$; Kruskal-Wallis rank sum test). Orange: *D. pseudoobscura;* Green*: D. persimilis.*

Sequence and expression divergence are associated with co-expression divergence.

To further understand patterns of co-expression network divergence we tested whether sequence and expression divergence differ across co-expression conservation and essentiality classes. We also tested if signals of positive selection are also associated with co-expression conservation classes and gene essentiality by retrieving divergence and selection estimates from Chapter 2. Our results show that genes with only divergent co-expression interactions between species show the highest levels of sequence divergence (Dxy) for females (Figure 4A). For males, however, we observed higher levels of sequence divergence on genes that fall in the conserved-divergent category (Figure 4B). Interestingly, these results support observed patterns of within species diversity: genes exhibiting divergent co-expression have higher levels of nucleotide diversity in both species (Figure S7 Appendix C). Moreover, we explored levels of sequences divergence across genes grouped by gene essentiality class. We observe that essential genes show the lowest levels of sequence divergence, while levels of divergence for genes classified as non-essential are significantly higher. Overall, our results suggest that genes with higher sequence divergence across species also tend to diverge at the co-expression level presumably by forming new connections with other genes generating species-specific clusters.

It is of particular interest to understand if adaptive divergence across species can be explained at the gene co-expression level. To shed light into the association of selection and gene expression divergence with gene co-expression differences across species, we tested for significant association across all different gene co-expression divergence and gene essentiality

classes. Our tests of independence reveal that in males (Figure 5A) and females (Figure S8 Appendix C) differential expression between species either across development or between tissues is associated with differential co-expression divergence. Differential expression also appears to be associated with gene essentiality but only in males (Figure 5B). Nonetheless, our results show only strong association of differential expression with genes that have not been classified in terms of essentiality.

We found that genes that exhibit co-expression divergence are more likely to be under selection at the nucleotide level (Figure 5C). Although significance is above the significant threshold ($p = 0.06$), it is noteworthy to mention that genes with evidence of differential co-expression in males show more evidence of selection. Moreover, we also tested if essentiality extrapolated from *D. melanogaster* data is associated with the different classes of co-expression divergence across species. In both males and females (Figure 5D; Figure S8B Appendix C), genes classified as non-essential also tend to be genes that exhibit co-expression divergence between species. Altogether, these results strongly suggest a significant association of sequence divergence with expression divergence. Moreover, it is noteworthy that genes with evidence of selection tend to be also genes with significant co-expression divergence across species.

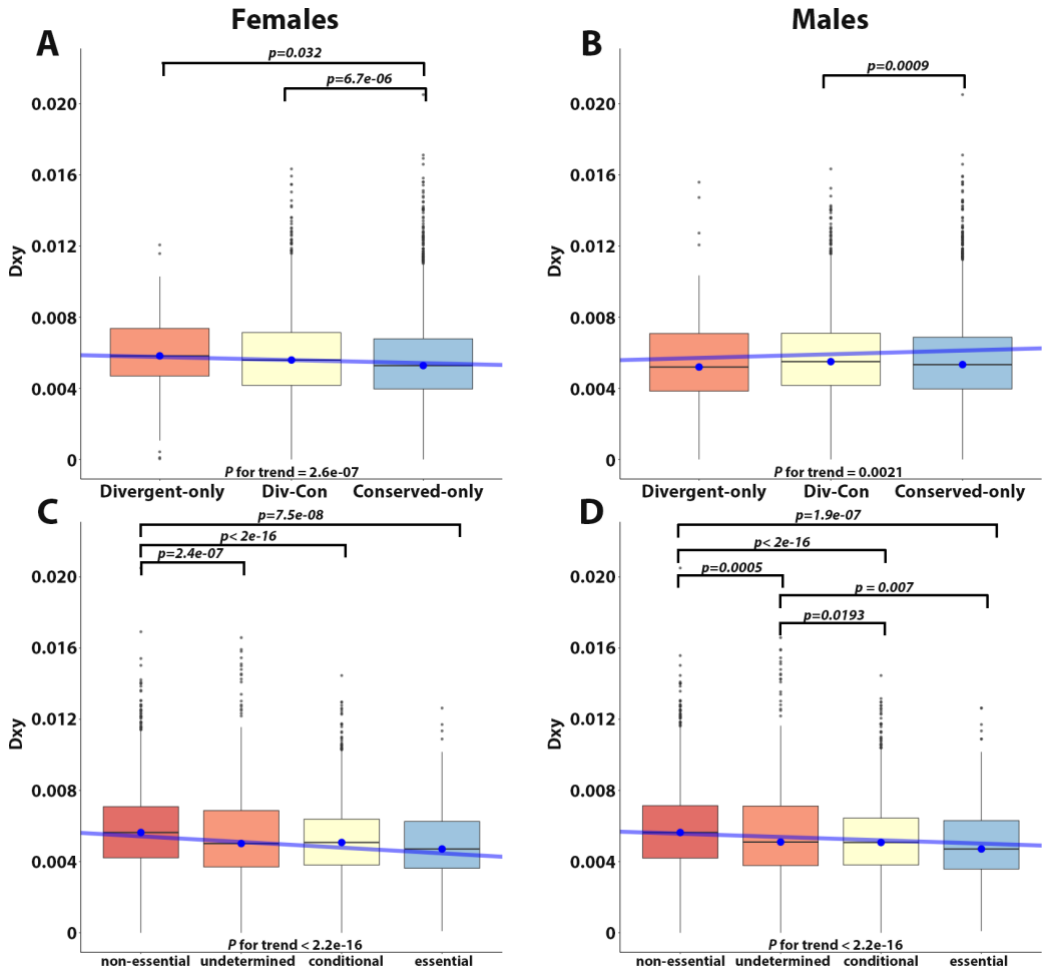


Figure 4. Divergence (Dxy) across genes grouped by edge classes and essentiality classification. Panels A (Kruskal-Wllis rank sum test; p = 1.769e-06) and B (Kruskal-Wllis rank sum test; p = 0.0015) show Dxy differences across edge classes for males and females. Panels C (Kruskal-Wllis rank sum test; p < 2.2e-16) and D (Kruskal-Wllis rank sum test; p < 2.2e-16) show Dxy differences across essentiality classification based on *D. melanogaster*. P-values are shown for each significant pairwise comparisons. Blue curves represent the linear regression of Dxy and gene classification. *P*=Wilcoxon rank-sum test for trend.

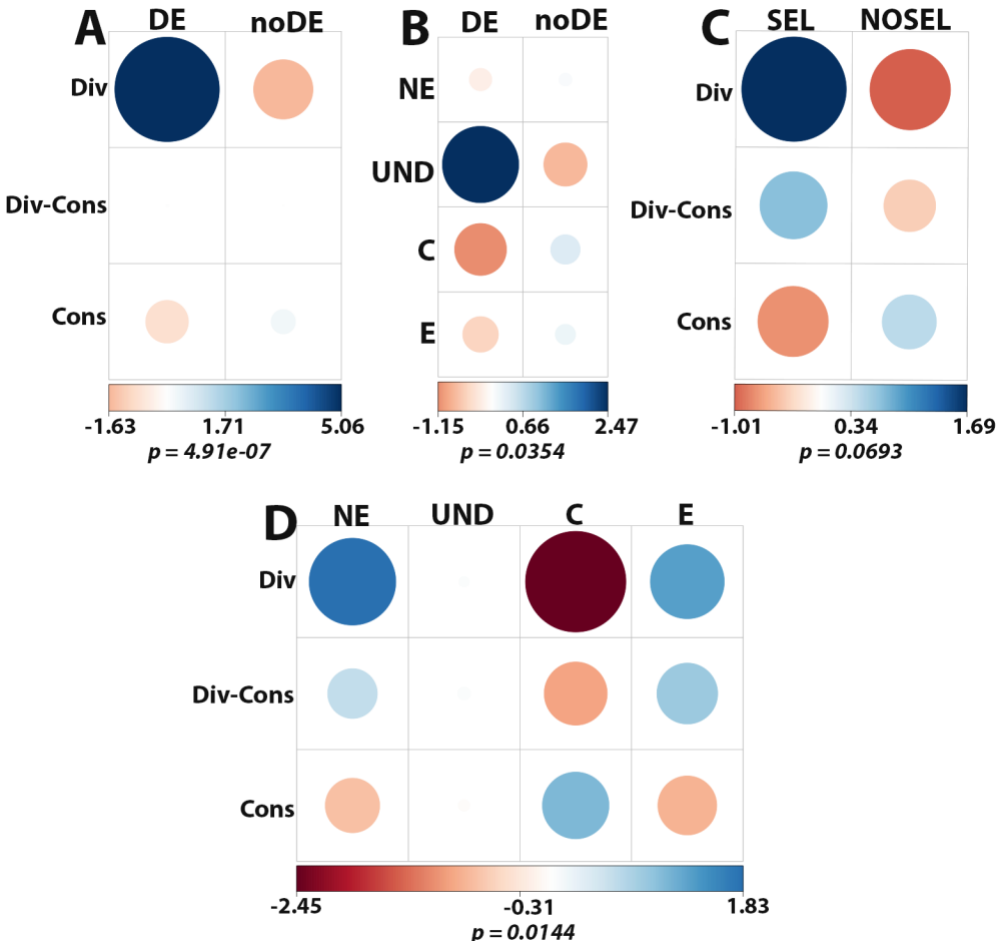


Figure 5. Test of independence comparing associations across differential expression status, evidence of selection, essentiality status and edge classes in males. A) Association between differential expression (DE) and edge classes. **B)** Association between differential expression status and gene essentiality. **C)** Association between evidence of selection at the nucleotide level and edge status. **D)** Association between gene essentiality and edge classes. Color of the circles indicates either positive association (blue) or negative association (red) between corresponding rows and column variables. The size of the circle is proportional to the amount of cell contribution. Chi-squared p-values are shown at the bottom of each contingency table. DE: differential expression; noDE: no evidence of differential expression; SEL: evidence

of selection; NOSEL: no evidence of selection; NE: non-essential genes; UND: not classified genes; C: conditional essentiality; E: essential. Div: genes with only divergent edges; Cons: genes with only conserved egdes; Div-Con: genes with both divergent and conserved edges.

Species-specific hubs in *D. pseudoobscura* and *D. persimilis*

We identified species specific as well as highly interconnected conserved genes ("hub genes") to identify potential targets of evolutionary change across species. For each module for each species, we took the top 5% genes with the highest values of KWithin, for males and females independently. In females, we identified a total of 422 hubs specific for *D. pseudoobscura* and 434 hubs for *D. persimilis* and a total of 130 hubs shared across species. For males, we observe a total of 445 hubs for *D. pseudoobscura*, 473 hubs for *D. persimilis* and 132 conserved hubs between species. For the total number of hubs identified, 130 hubs in *D. pseudoobscura* and 105 hubs in *D. persimilis* show evidence of differential expression between species.

GO enrichment analyses for species specific hubs show a variety of enriched biological functions. For *D. pseudoobscura* females, mRNA splicing (GO:0000398) is the most represented biological function. For *D. persimilis*, regulation of transcription (GO:0006357) is the most enriched biological function. In males, *D. pseudoobscura* hubs are enriched for mitotic cell cycle (GO:0000278) and axon guidance (GO:0007411). For *D. persimilis* males, hubs are enriched for protein ubiquitination (GO:0016567). Conserved female hubs, compound eye morphogenesis (GO:0001745) is the most enriched biological function. Moreover, for male conserved hubs, sperm axoneme assembly (GO:0007288) was enriched.

Although, no significant association of differential expression and evidence of selection was found across species-specific or conserved hubs, we examined species-specific gene hubs

with evidence of selection and differential expression. Nonetheless, several species-specific gene hubs show evidence of differential expression and selection. In *D. pseudoobscura* females, nine species-specific hubs show evidence of differential expression and selection. Interestingly, these include genes associated with the development of sensory neurons (*TfAP-2*, *beat-lia*, *kat-60L1*, *Rab4*) and sexual reproduction (CG15117). For *D. persimilis* females, only four genes exhibit evidence of differential expression and selection. It is also noteworthy that two genes are involved in neural development (*swm*) and in the perception of mechanical stimuli (*Cirl*). Moreover, one gene is predicted to be a transcription factor (CG3918) and one gene has no homology with *D. melanogaster* genes. Most of these genes show upregulation in *D. pseudoobscura* except for genes *beat-Iia*, *kat-60L1* and *swm*, which show upregulation in *D. persimilis*. In addition, these same genes show signs of selection in form of selective sweeps in *D. pseudoobscura* (*beat-lia*), in *D. persimilis* (*TfAP-2*, *kat-60L1, Rab4,* CG3918) or both (CG15117, *swm*, *Cirl*) (Table S5 Appendix C).

For males, we identified five *D. pseudoobscura* specific hubs. One of these genes is a transcription factor involved in olfactory behaviour (*Aef1*) and we observe a second gene involved in cell fate specification of sensory organ precursors (*Arpc1*). For *D. persimilis*, we identified a total of eight species-specific hubs. Hubs in these species show a wider variety of functions but we still observed that one gene is involved in memory acquisition (*SA1*) and two genes involved in neuronal development (*chinmo*, *Cul1*). *Arpc1*, SA1 and *Cul1* show upregulation in *D. persimilis* while *chinmo* upregulates in *D. persimilis*. Gene *Aef1* presents more complex patterns of expression. This gene is upregulated in ovaries in *D. pseudoobscura* but is upregulated in pupa and testis in *D. persimilis*. Interestingly, these species-specific hubs in males

show evidence of selective sweeps in both species, except for *Cul1*, which shows evidence of selective sweep on *D. pseudoobscura* only (Table S5 Appendix C).

We also examined shared hubs across sexes for each species to reveal potential network rewiring occurring in both sexes in each species. We identified a total of 30 hubs in *D. pseudoobscura* and 49 hubs in *D. persimilis*. Although none of those hubs show evidence of both differential expression and selection in the same gene, we still found genes for genes involved in response to stress and heat (*Hsp110*, *Hsp67Ba*), to immunity gene expression (*Mkk4*), sex determination alternative splicing (*tra2*) and oogenesis defects (*α-Cat*). *Hsp110*, *Hsp67Ba*, *Mkk4* are classified as *D. pseudoobscura* hubs whereas *tra2* and *α-Cat* are hubs in *D. persimilis*. Among these genes, only *Mkk4* and *tra2* show evidence of selective sweeps in *D. pseudoobscura* (Table S6 Appendix C).

Overall, these results suggest that co-expression rewiring across species can occur on genes involved in neural development associated to sensory pathways not only over developmental stages but also across tissues.

DISCUSSION

Using gene co-expression networks to study how genes interact with each other during development and how those interactions change during species divergence provides a new important approach for understanding the process of species divergence. We reasoned that by comparing patterns of gene interactions between species it is possible to retrieve a unifying picture of evolutionary processes driving species differentiation at the genomic level. Here we seek to compare gene co-expression networks between two closely related Drosophila species to

detect significant network rewiring and the potential implications in processes of adaptive divergence between species. The present work constitutes a novel use of network-based system biology approaches to establish conservation and divergence at the level of co-expression networks between species.

The main limitation for studying species divergence at the level of gene co-expression networks is the lack of comprehensive expression datasets from multiple species; such datasets are needed for the construction of robust co-expression networks. Here we collected a comprehensive RNA-seq data set that not only includes a developmental time series but also tissue-specific samples for multiple populations of both *D. pseudoobscura* and *D. persimilis* (see methods). Although other expression datasets are available for few Drosophila species, they are not suitable for co-expression analyses and are not comparable to the one we have generated for this study (Krause, Overend, Dow, & Leader, 2022; Lau et al., 2020; J. Liu et al., 2022; Paris, Villalta, Eisen, & Lott, 2015; Thurmond et al., 2019). This study is thus the first of its kind, reporting gene co-expression networks constructed from comprehensive RNA-seq data sets from multiple developmental stages, tissues and populations from two closely related species.

The co-expression networks in the present study were constructed using only 1:1 orthologous genes (8,680 from females, and 9,902 from males), leaving out of the analyses more than 3,000 genes for each species. Although this might posit a limitation for studying the whole interactome in both species, significant co-expression differences observed between species still have the potential to reveal groups of genes driving species divergence. For example, foundational studies in gene co-expression network comparisons revealed significant network rewiring across brain tissues within humans and between humans, chimpanzees, and mice (Langfelder & Horvath, 2007; Oldham et al., 2006). These studies were able to capture tissue

specific modules by considering a total of no more than 3,000 genes. Our sample is more comprehensive, allowing us to implement network analyses approaches that are suitable for capturing key co-expression differences and network rewiring between species.

Modules in the context of WGCNA are interpreted as groups of genes that are highly co-expressed with each other (Langfelder & Horvath, 2007). One of the main features of WGCNA is to find modules of genes that can be functionally related. Here we used WGCNA to identify modules for each sex between species and implemented the full tool set to test module conservation between species. This approach has been applied to a variety of systems (Du et al., 2021; Oldham et al., 2006; Ovens et al., 2021; Pembroke et al., 2021), but her we present the first module comparisons across Drosophila species. First, we observed that the number of modules differs slightly for both sexes (Figure 1). *D. persimilis* appears to be the species with lower number of identified modules in both males and females (Table S1 Appendix C). Although it is possible that this module number difference is due to technical artifacts, our conservation statistics (Figure 2) suggest that module size and number can be interpreted as real differences between species.

The lower number of modules observed in *D. persimilis* do not imply the presence of more species-specific modules in *D. pseudoobscura*. We detect a common pattern where two or more modules in one species significantly overlap to a single module in the other (Figure 1). For example, the magenta module in *D. persimilis* females contains genes from three modules in *D. pseudoobscura* (tan, turquoise, and greenyellow; Figure 1). The female magenta module in *D. persimilis* contains a variety of genes with diverse functions that range from courtship and feeding behavior to organogenesis across development. In *D. pseudoobscura*, these same functional categories appear to be represented in 3 different modules. These differences indicate

potential new co-expression connection in one species or the loss of co-expression connection in the other species. Moreover, these differences can also be interpreted as differences in gene expression profiles occurring on genes that might be functionally related as seen for other systems. (Filteau et al., 2013; Oldham et al., 2006; Shahan et al., 2018; Tsaparas et al., 2006; Voigt et al., 2017).

Full characterization of such genes can unveil functional pathways associated with the process of species divergence. Although this study focused on detecting gene co-expression differences between species, considerable module conservation still occurs between species. This conservation is mostly observed on the largest modules for each species, which suggests that conserved functional pathways tend to show similar expression profiles between species.

Although WGCNA analyses suggest conservation in connectivity patterns, identifying which gene pairs are conserved across species and which genes are forming new connections to other genes is not trivial under the WGCNA framework. Hence, we implemented the approach of (Voigt et al., 2017) to identify conserved and divergent co-expression interactions across species. This idea has been mostly applied to identify co-expression differences to identify target genes for human disease (van Dam, Vosa, van der Graaf, Franke, & de Magalhaes, 2018) and for the description of complex phenotypes in plants (Aoki, Ogata, & Shibata, 2007). A similar approach was also implemented to study sex-bias expression differences across the development of *Nassonia vitripennis* (Rago et al., 2020). Here the implementation of this approach revealed that genes holding only conserved co-expression connections across species have high number of connections within species networks (Figure 3). This is consistent with the idea that thigh conserved connections can include conserved essential pathways making them less likely to change. Nonetheless, we also observed that genes holding a combination of conserved and

divergent connections have even higher overall numbers of connections with other genes. Although this pattern has not been described before, we hypothesize that some conserved hubs could be forming connections with less conserved genes from the network periphery (Voigt et al., 2017). This idea is also consistent with what we observed for genes with divergent co-expression connections across species (Figure 3). These genes show the lowest number of connections, consisten with expectation that less connected genes are more likely to diverge. Hence, genes with evidence of co-expression divergence can potentially be drivers of phenotypic divergence.

Interestingly, genes with higher co-expression divergence also show higher levels of sequence divergence in females (Figure 4), indicating potential associations among them. This is consistent with previous studies in protein-protein interaction networks that show higher sequence variability across genes with a low number of connections, making them potential targets of selection (Alvarez-Ponce et al., 2017; S. Chakraborty & Alvarez-Ponce, 2016; Luisi et al., 2015). Nonetheless, we do observe an opposite pattern in males (Figure 4). Males tend to have slightly higher levels of sequence divergence for genes with conserved co-expression connections. Although this seems counterintuitive, this observation suggests that in males, conserved pathways are evolving faster than in females possibly under the influence of selection (G. W. Liu et al., 2015; Luisi et al., 2015). These results are also consistent with what we observed in our association tests, where strong associations among co-expression divergence, differential expression, and evidence of selection, were only observed in males. Hence, even though in females we observed significant network rewiring between species, it appears that genes with high co-expression divergence are not more likely to be targets of selection at the sequence or gene expression levels.

One of the main goals of speciation genomics if to identify potential drivers of species divergence. In *Drosophila,* few genes involved in reproductive incompatibilities have been identified (Patlar & Civetta, 2021; C. I. Wu & Ting, 2004). A common feature of these genes is that all appear to be male-specific genes that express in male reproductive tissue. For example, *Hmr* (hybrid male rescue) and *Lhr* (lethal hybrid rescue) induce male lethality in *D. melanogaster and D. simulans* (Barbash, Roote, & Ashburner, 2000; Barbash, Siino, Tarone, & Roote, 2003). *Ovd* is involved in hybrid male sterility between *D. pseudoobscura* and its subspecies *D. p bogotana (N. Phadnis, 2011; Nitin Phadnis & Orr, 2009)*. Although orthologs of these genes were included in our co-expression networks, we do not observe co-expression differences between *D. pseudoobscura* and *D. persimilis*. Our study, however, does not rule out the possibility that these genes could be involved in species incompatibilities in our species given our conservative threshold approach for defining co-expression classes.

To our knowledge this study presents the first effort to shed light on candidate genes involved in the divergence of *D. pseudoobscura* and *D. persimilis*. Despite much focus on the role of inversions in species divergence (Korunes et al., 2021; Machado et al., 2007; M. A. Noor et al., 2001), nothing is known about the identity of the genes driving adaptive divergence between this species pair. Our results can start providing clues about the involvement of a variety of genes and pathways on species divergence. For instance, results showing that species-specific hub genes are mostly related to the development of neural tissue involved in sensory pathways points towards the importance of behavioral differences between these species. Although very little is known about the natural history and ecology of this species pair, some of the few measurable differences between these species are related to behavior (Lindsay, 1958; Matsuo, Nose, & Kohsaka, 2021; Mohn & Spiess, 1963; M. A. F. Noor & Aquadro, 1998).

This study underlies the usefulness of implementing system biology approaches to address questions in evolutionary biology. We demonstrate that studying species differences in the context of gene co-expression networks provides a novel approach for the identification of drivers of speciation.

ACKNOWLEDGMENTS

I thank the members of the Machado Lab for helpful discussions. Also, to Dr. Michelle Girvan, Dr. Daniel Serrano, and Dr. Hector Corrada-Bravo for helpful discussions and technical advice. I also thank the COMBINE program for the training in network analyses. I thank Dr. Kevin Nyberg for generating the RNA-seq data used in this part of the dissertation. Research supported by National Science Foundation grants MCB-1716532 to Carlos Machado.

Appendix A: Supplementary material for Chapter 1

Supplementary Figures

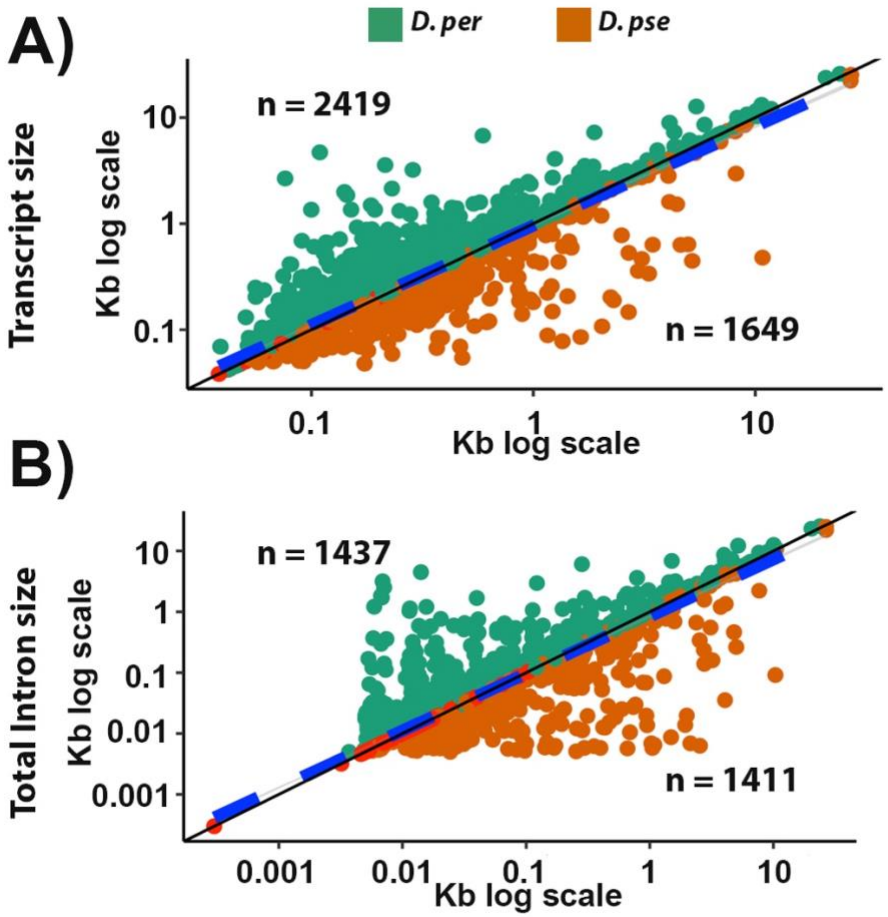


Figure S1. Correlation plots for transcript size (A) and intron size (B) between *D. pseudoobscura* and *D. persimilis* for the 4,613 orthologous genes with the same aminoacid length. 'n' depicts the number of genes with larger transcript size (A) or larger intron size (B) for each species (above or below the curve). Black solid line represents the 1:1 expectation between species; blue dashed line depicts the implemented LM: $R^2 = 0.8694$; Intercept = 0.2358; Slope = 0.9363 (A) and $R^2 = 0.8756$; Intercept = 0.1736; Slope = 0.9372 (B).

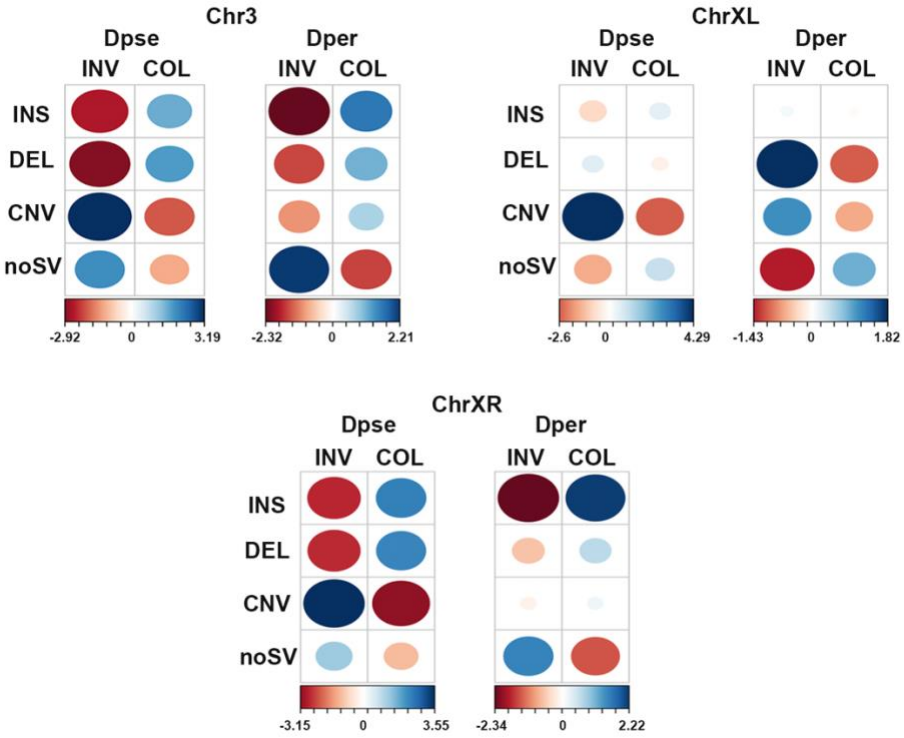


Figure S2. Correspondence analysis showing the association between genes (including the 10kb upstream region) and SVs, for chromosomes 3, XL and XR. Circle sizes depict the number of genes, and color depicts correlation values (bottom of contingency table). INS: insertions; DEL: deletions; CNV: copy-number variants; noSV: genes not associated with SVs.

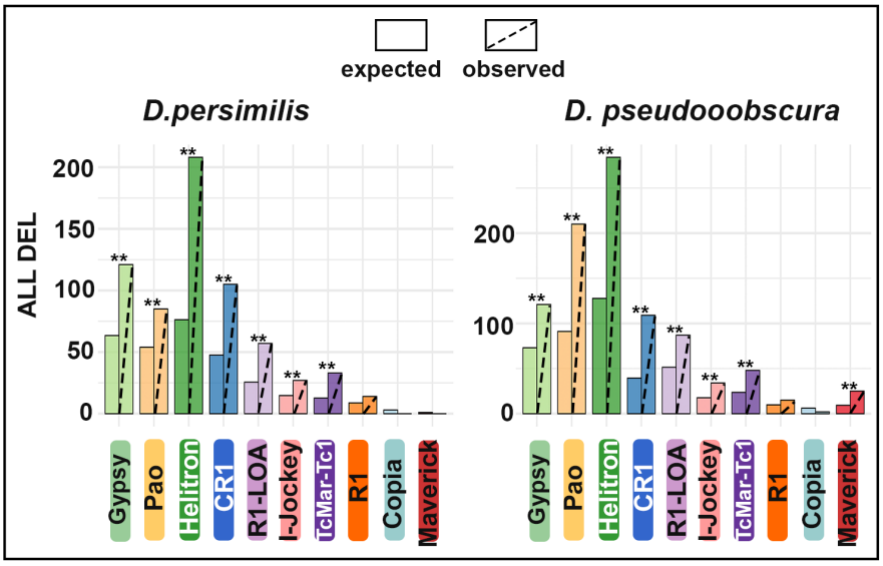
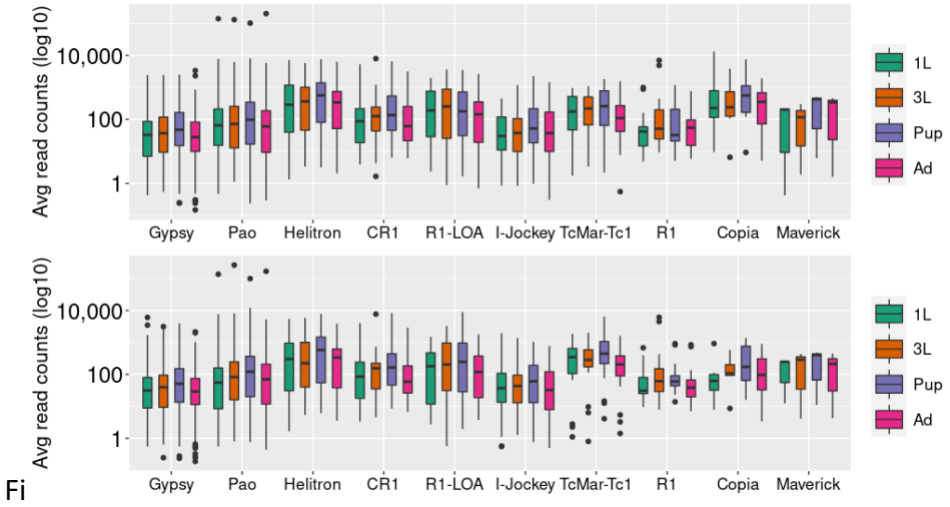


Figure S3. Permutation analysis of DEL overlapping TE annotations; ** $p < 0.01$ significant difference between observed and expected counts.



Fi

Figure S4. Expression levels of the 10 most abundant TE families in both *D. pseudoobscura* and *D. persimilis* (x-axis). Colors depict four developmental stages and average read counts are represented in log scale (y-axis). Up graph show read counts using *D. pseudoobscura* genome as a reference and bottom graph is the reciprocal analysis using *D. persimilis* genome as a reference.

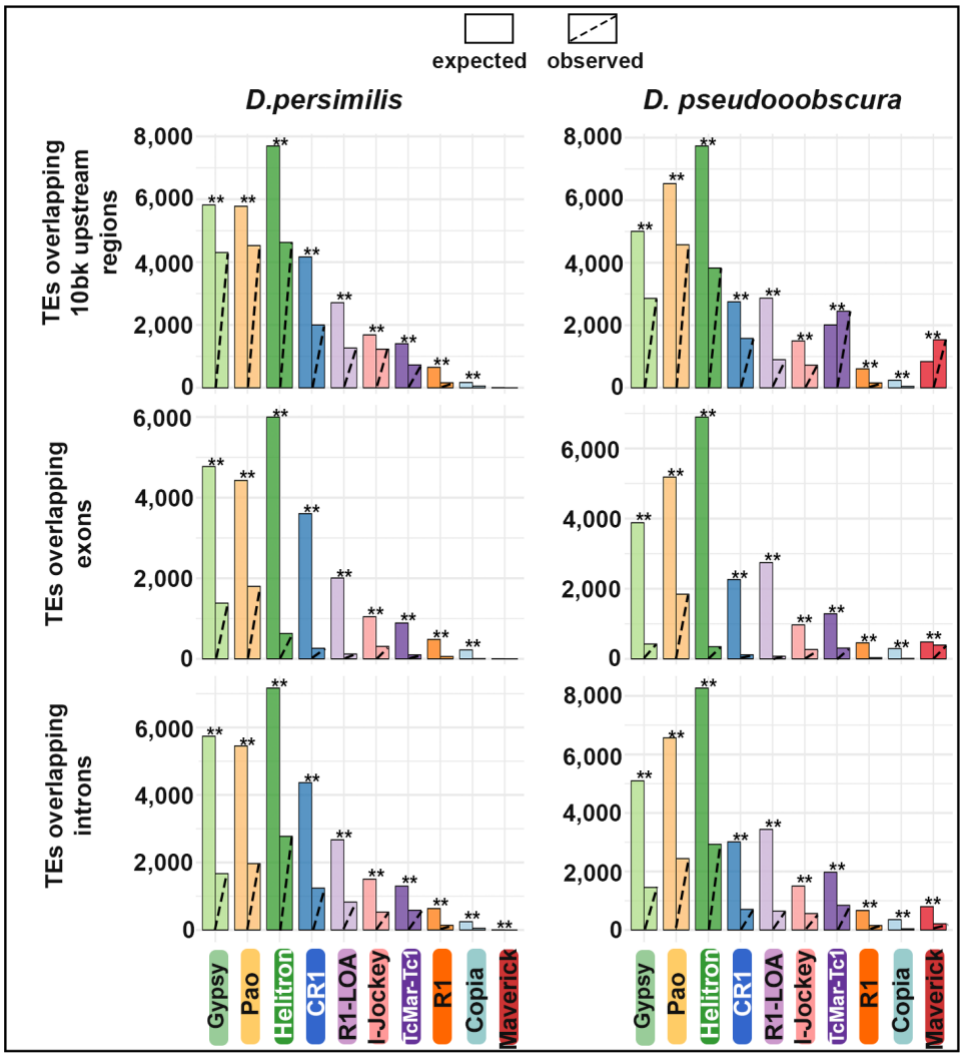


Figure S5. Permutation analysis of TEs overlapping annotated gene regions; ** $p < 0.01$ significant difference between observed and expected counts.

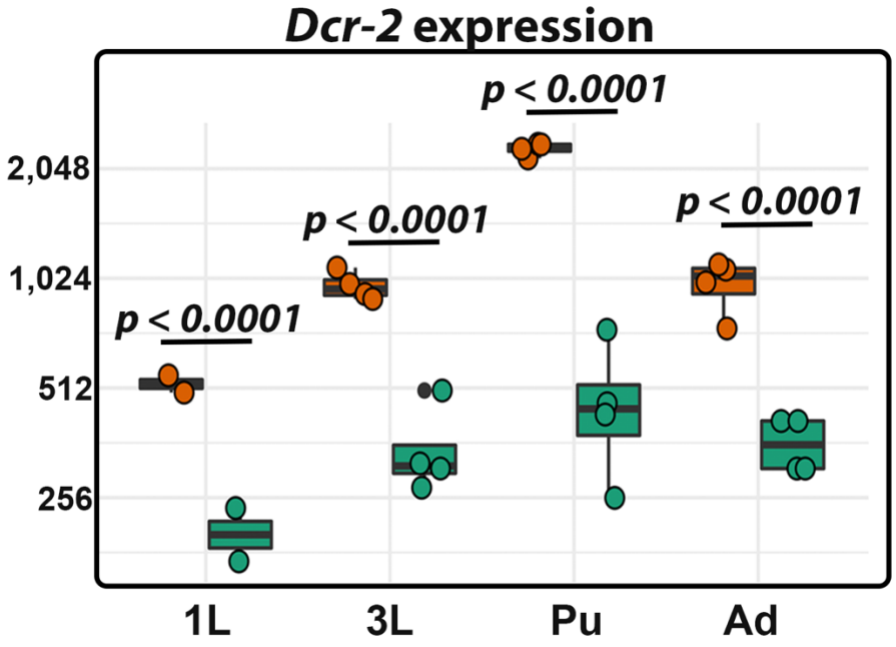
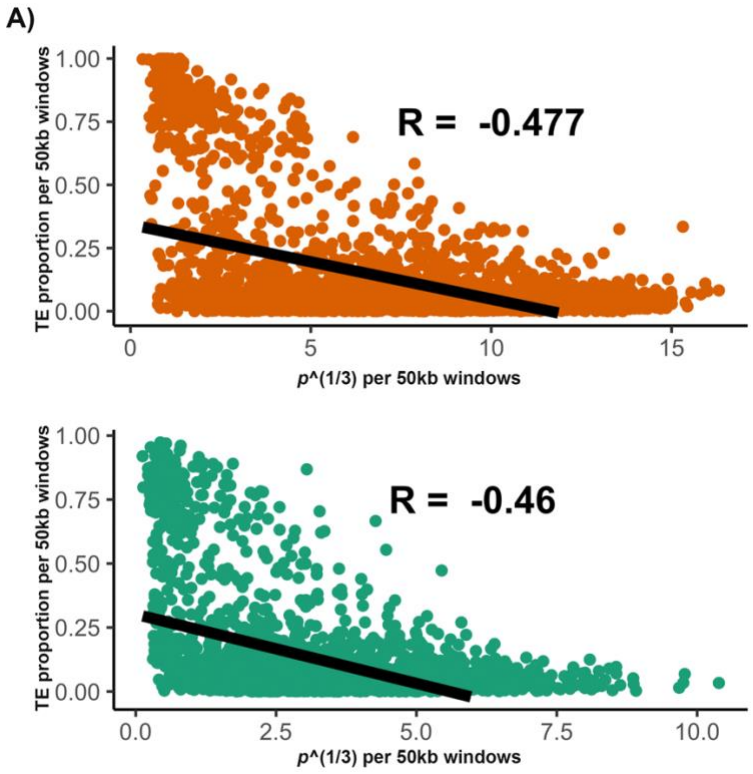


Figure S6. Boxplots showing the DESeq2 normalized read counts in log scale for *Rhino* and *Dcr-2* over four developmental stages: 1L: first instar larvae, 3L: third instar larvae, Pup: Pupae, Ad: Adult, between *D. pseudoobscura* (orange) and *D. persimilis* (green).



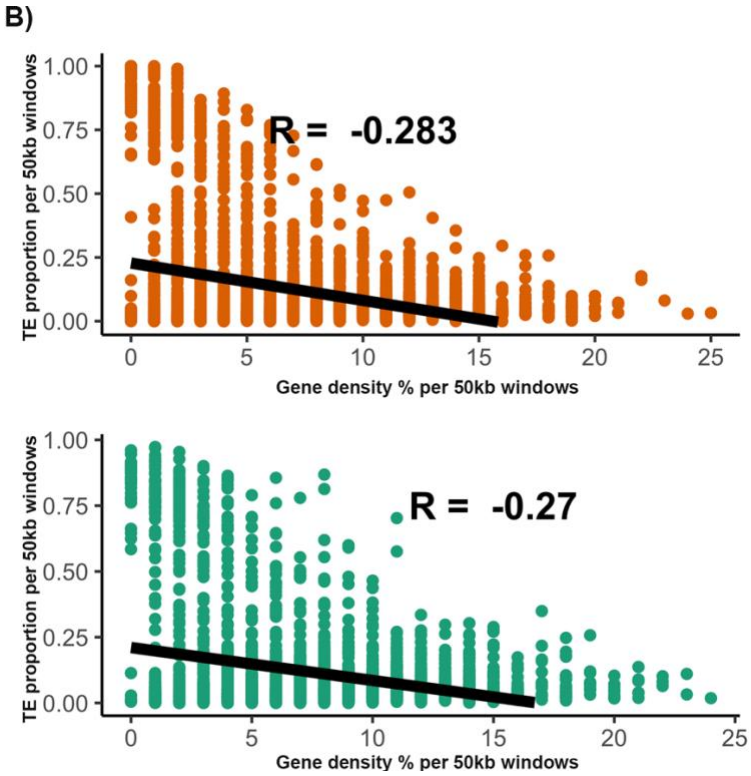


Figure S7. Negative correlation between TE proportion recombination rates (A) and gene density (B) for *D. pseudoobscura* (orange) and *D. persimilis* (green).

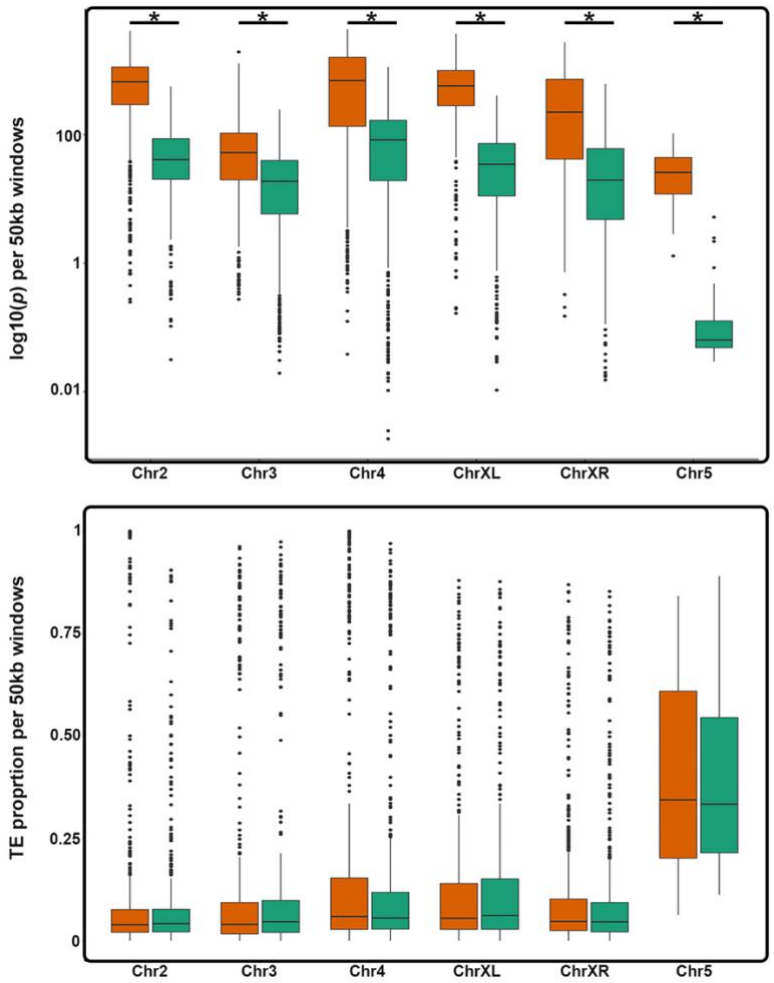


Figure S8. Recombination rate (top) and TE content (bottom) estimates for each chromosome for both *D. pseudoobscura* (orange) and *D. persimilis* (green). * Recombination rate p-values (Wilcoxon test) $p < 2.2e\text{-}16$ for chromosomes 2, 3, 4, XL and XR. $p = 3.564e\text{-}13$ for chromosome 4.

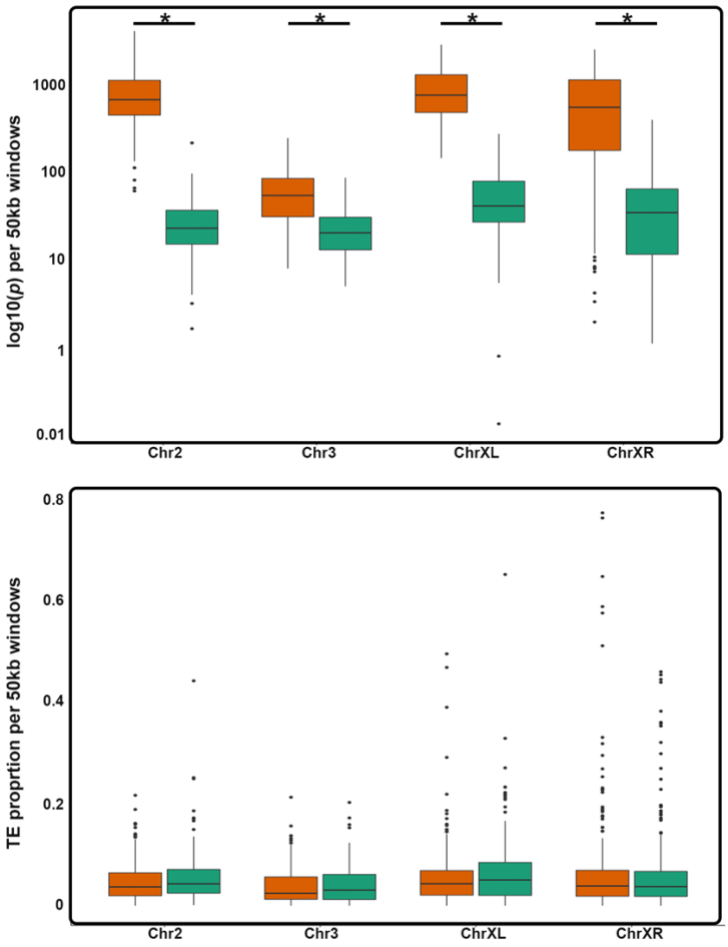


Figure S9. Recombination rate (top) and TE content (bottom) estimates for inverted regions for both *D. pseudoobscura* (orange) and *D. persimilis* (green). * Recombination rate p-values (Wilcoxon test) $p < 2.2e\text{-}16$.

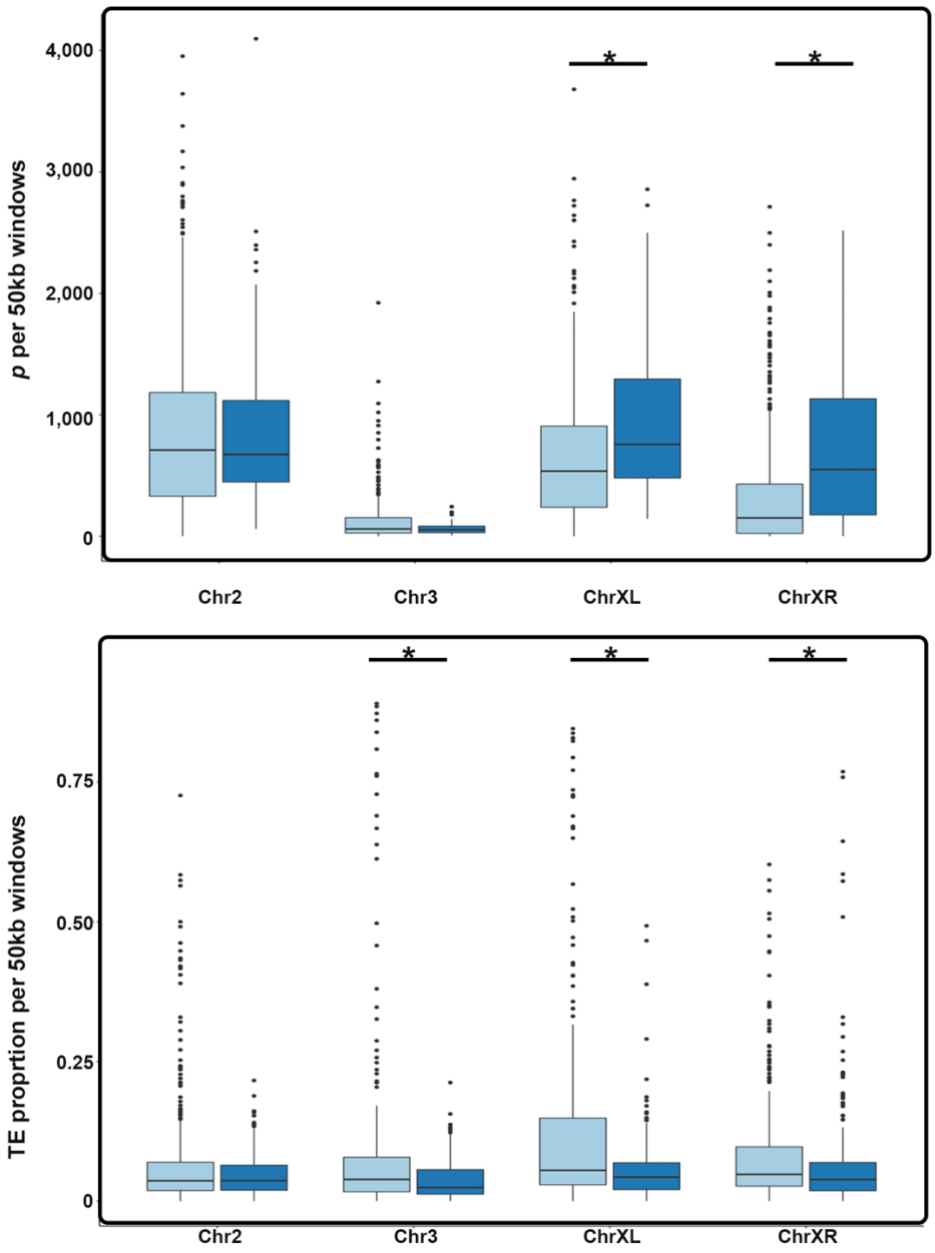


Figure S10. Recombination rate (top) and TE content (bottom) estimates comparing inverted (dark blue) and co-linear (light blue) regions for *D. pseudoobscura*. * Recombination rate p-values (Wilcoxon test): ChrXL $p = 2.172e-06$, ChrXR $p < 2.2e-16$. * TE proportion p-values (Wilcoxon test): Chr3 $p = 0.00226$, ChrXL $p = 0.000302$, ChrXR $p = 0.005167$.

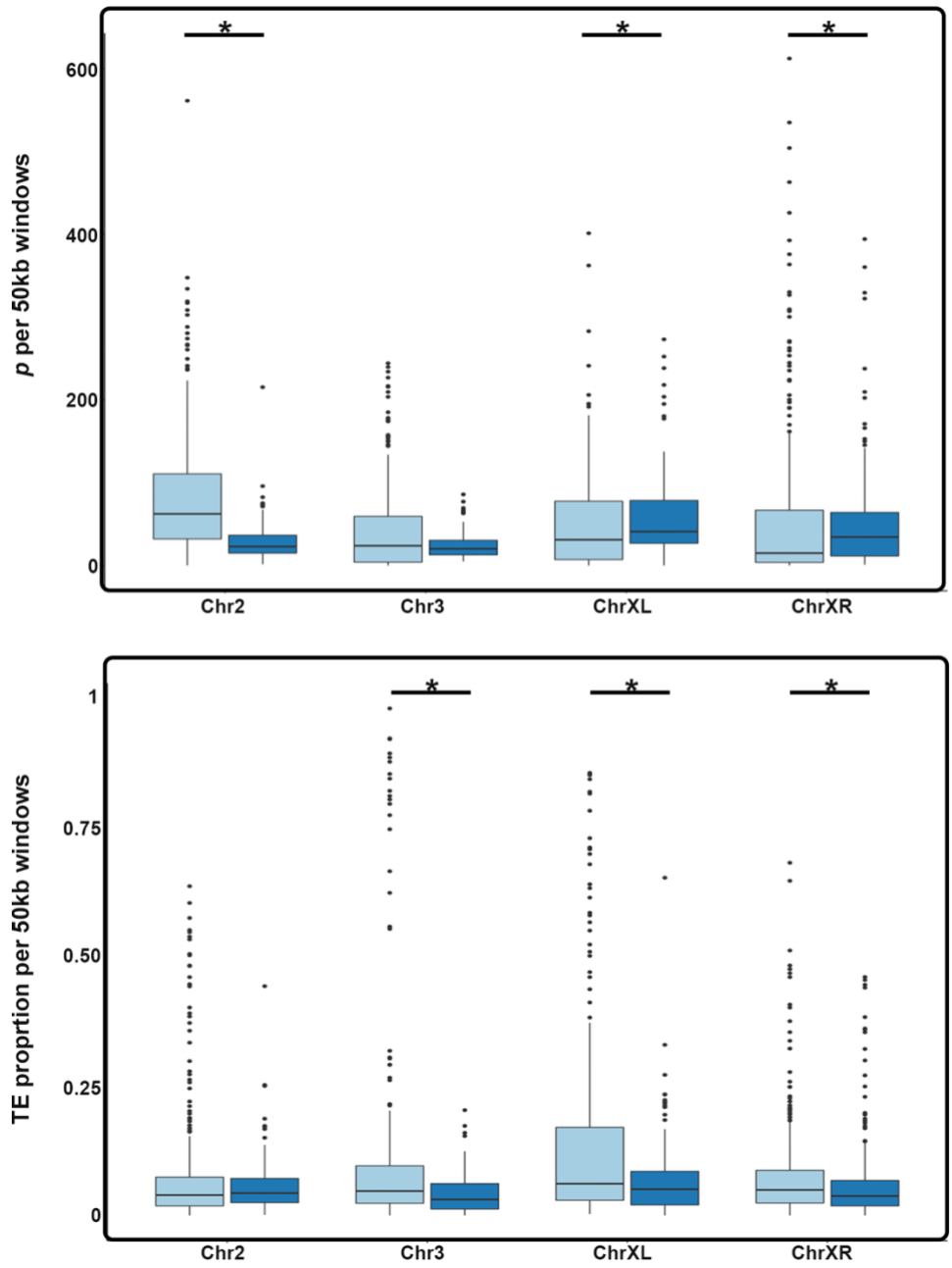


Figure S11. Recombination rate (top) and TE content (bottom) estimates comparing inverted (dark blue) and co-linear (light blue) regions for *D. persimilis*. * Recombination rate p-values (Wilcoxon test): Chr2 $p < 2.2e\text{-}16$, ChrXL $p = 0.001705$, ChrXR $p < 0.0003285$. * TE proportion p-values (Wilcoxon test): Chr3 $p = 1.295e\text{-}05$, ChrXL $p = 0.0007596$, ChrXR $p = 0.001592$.

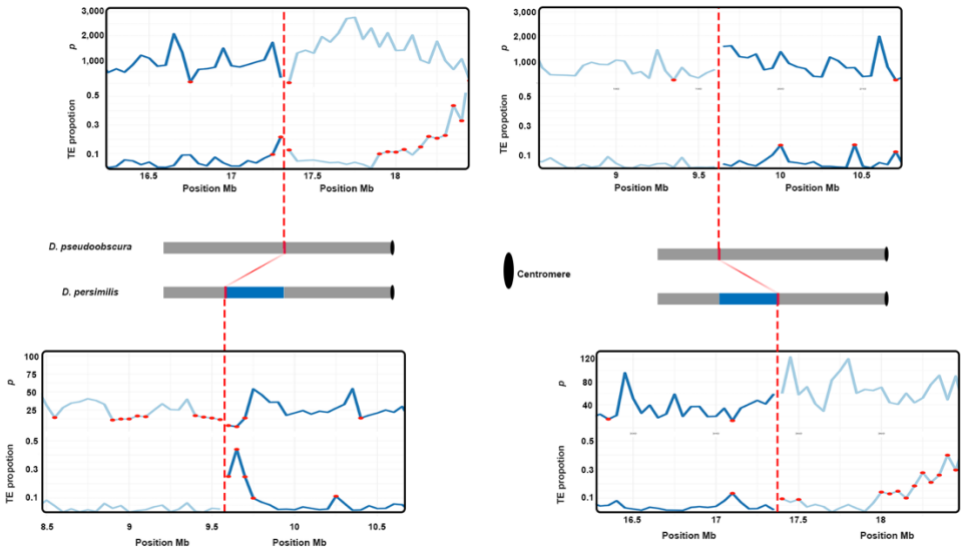


Figure S12. Recombination rates (top of each plot) and TE proportion (bottom of each plot) at the proximal *D. pseudoobscura* and distal *D. persimilis* (left) and distal *D. pseudoobscura* and proximal *D. persimilis* (right) for chromosome 2. Red dots represent windows with the lowest 20% values of recombination rate (top plot) and the top 20% values for TE proportion (bottom plot). Light blue depicts co-linear regions and dark blue inverted regions. Red dashed line depicts inversion breakpoints.

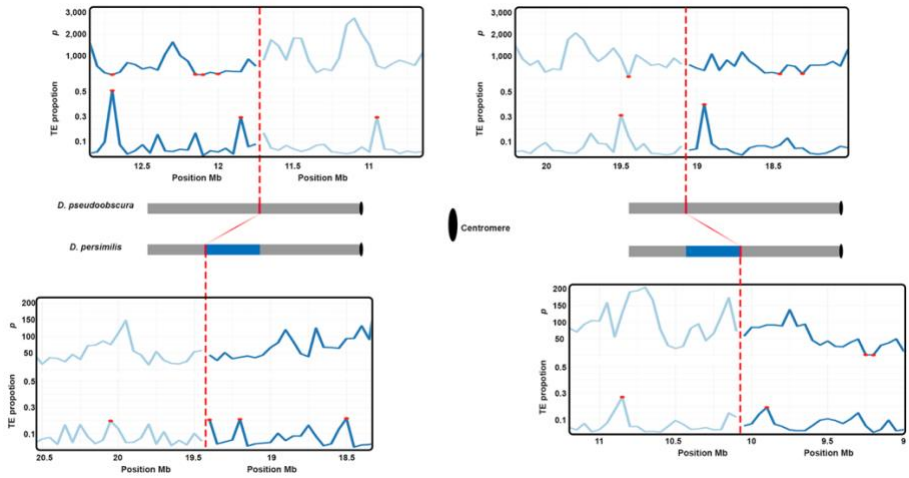


Figure S13. Recombination rates (top of each plot) and TE proportion (bottom of each plot) at the proximal *D. pseudoobscura* and distal *D. persimilis* (left) and distal *D. pseudoobscura* and proximal *D. persimilis* (right) for chromosome XL. Red dots represent windows with the lowest 20% values of recombination rate (top plot) and the top 20% values for TE proportion (bottom plot). Light blue depicts co-linear regions and dark blue inverted regions. Red dashed line depicts inversion breakpoints.

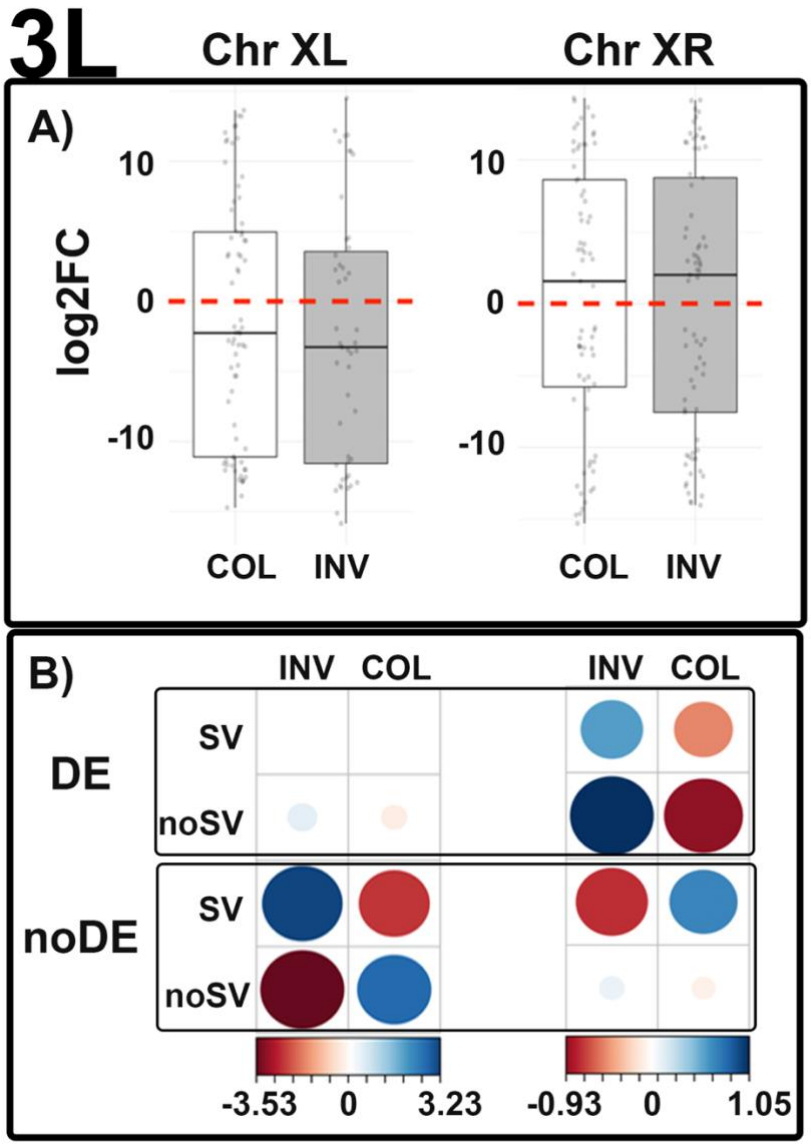


Figure S14. Gene expression and its association with SVs for chromosomes with inversion differences between *D. pseudoobscura* and *D. persimilis* for the 3L developmental stage. A) Log2 fold change values for differentially expressed genes comparing the co-linear and inverted regions; > 0 higher expression in *D. pseudoobscura*; < 0 higher expression in *D. persimilis*. B) Correspondence analysis showing the association of genes differentially expressed (DE) or not (noDE) with the presence of absence of SVs in the 3L stage; circle sizes depict number of genes, and color depicts correlation values (contribution of the overall Chi-square statistic).

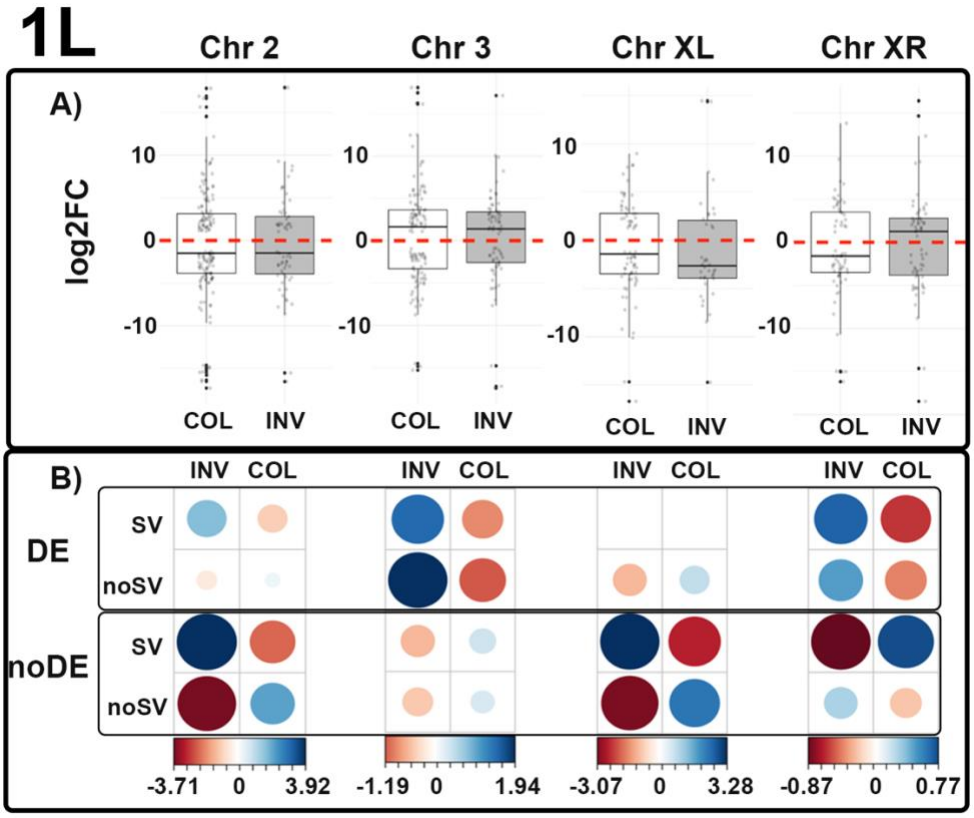


Figure S15. Gene expression and its association with SVs for chromosomes with inversion differences between *D. pseudoobscura* and *D. persimilis* for the 1L developmental stage. A) Log2 fold change values for differentially expressed genes comparing the co-linear and inverted regions; > 0 higher expression in *D. pseudoobscura*; < 0 higher expression in *D. persimilis*. B) Correspondence analysis showing the association of genes differentially expressed (DE) or not (noDE) with the presence of absence of SVs in the 1L stage; circle sizes depict number of genes, and color depicts correlation values (contribution of the overall Chi-square statistic).

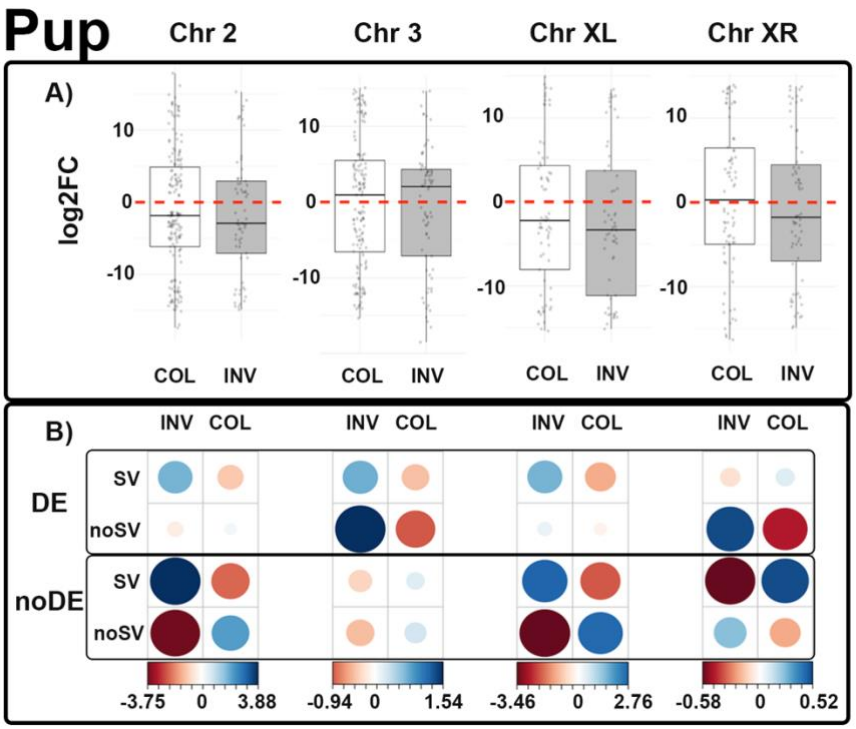


Figure S16. Gene expression and its association with SVs for chromosomes with inversion differences between *D. pseudoobscura* and *D. persimilis* for the Pup developmental stage. A) Log2 fold change values for differentially expressed genes comparing the co-linear and inverted regions; > 0 higher expression in *D. pseudoobscura*; < 0 higher expression in *D. persimilis*. B) Correspondence analysis showing the association of genes differentially expressed (DE) or not (noDE) with the presence of absence of SVs in the Pup stage; circle sizes depict number of genes, and color depicts correlation values (contribution of the overall Chi-square statistic).

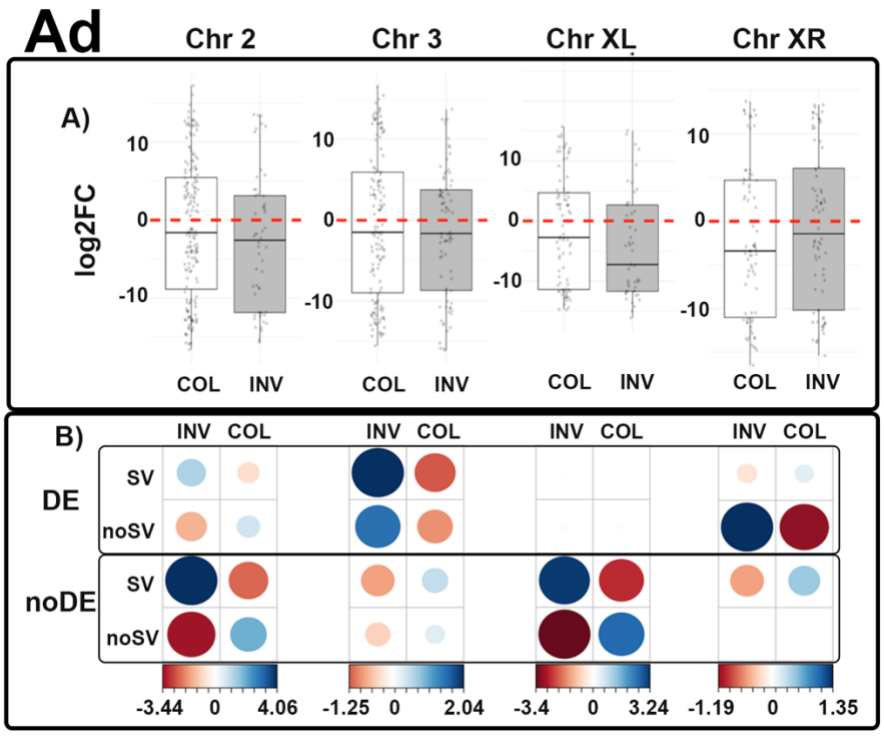


Figure S17. Gene expression and its association with SVs for chromosomes with inversion differences between *D. pseudoobscura* and *D. persimilis* for the Ad developmental stage. A) Log2 fold change values for differentially expressed genes comparing the co-linear and inverted regions; > 0 higher expression in *D. pseudoobscura*; < 0 higher expression in *D. persimilis*. B) Correspondence analysis showing the association of genes differentially expressed (DE) or not (noDE) with the presence of absence of SVs in the 3L stage; circle sizes depict number of genes, and color depicts correlation values (contribution of the overall Chi-square statistic).

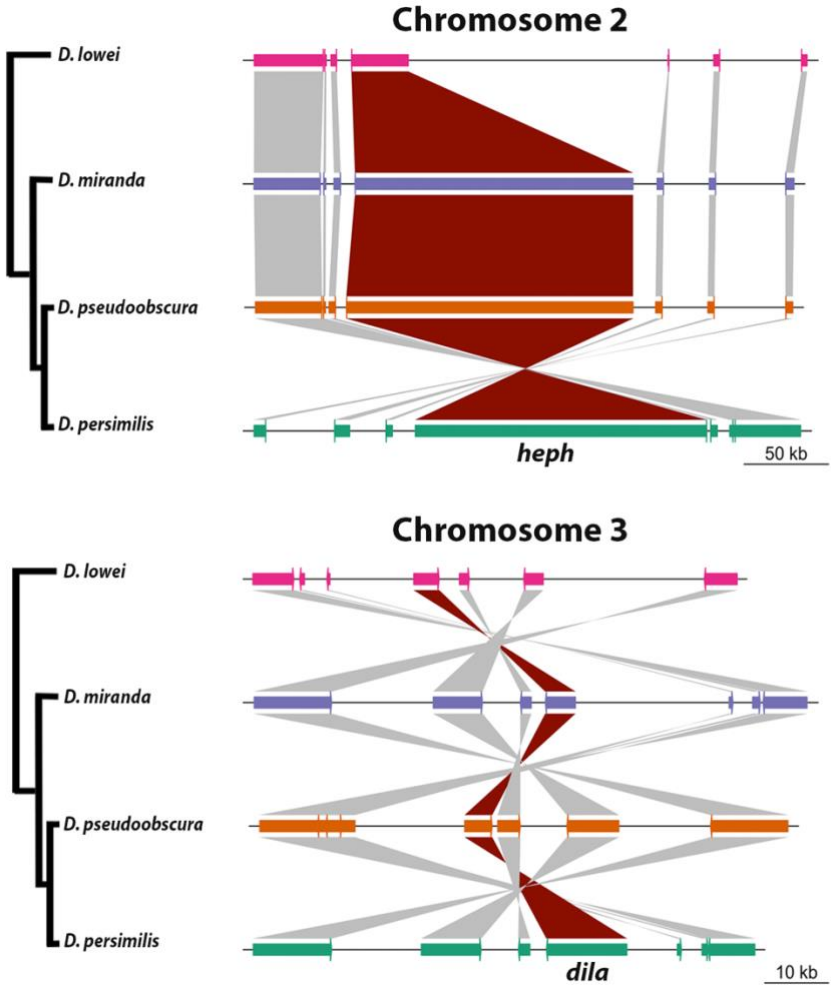


Figure S18. Genomic context (up) of the *heph* and *dila* (darked ribbon) genes showing conservation on the order of neighbor genes both up and downstream across the *D. pseudoobscura* subgroup.

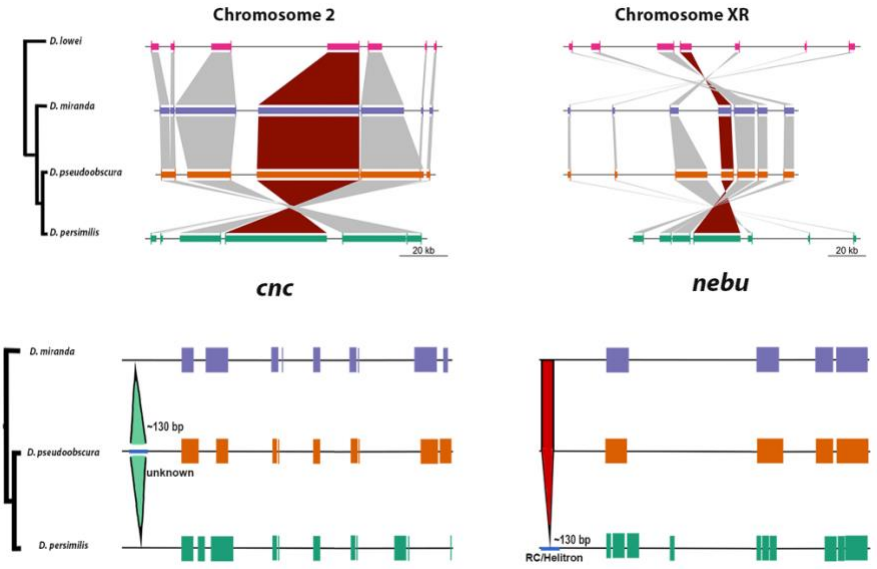


Figure S19. Genomic context (up) of the *cnc* and *nebu* (darked ribbon) genes showing conservation on the order of neighbor genes both up and downstream across the *D. pseudoobscura* subgroup. The bottom figures show the *cnc* and *nebu* gene models and the position of an INS occurring in *D. pseudoobscura* and a DEL occurring in *D. persimilis*, both in the 10kb-upstream region.

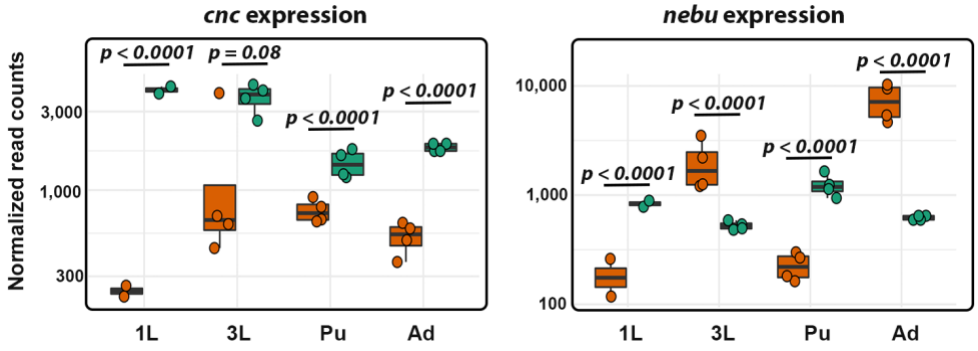


Figure S20. Boxplots showing the DESeq2 normalized read counts in log scale for *cnc* and *nebu* over four developmental stages: 1L: first instar larvae, 3L: third instar larvae, Pup: Pupae, Ad: Adult, between *D. pseudoobscura* (orange) and *D. persimilis* (green).

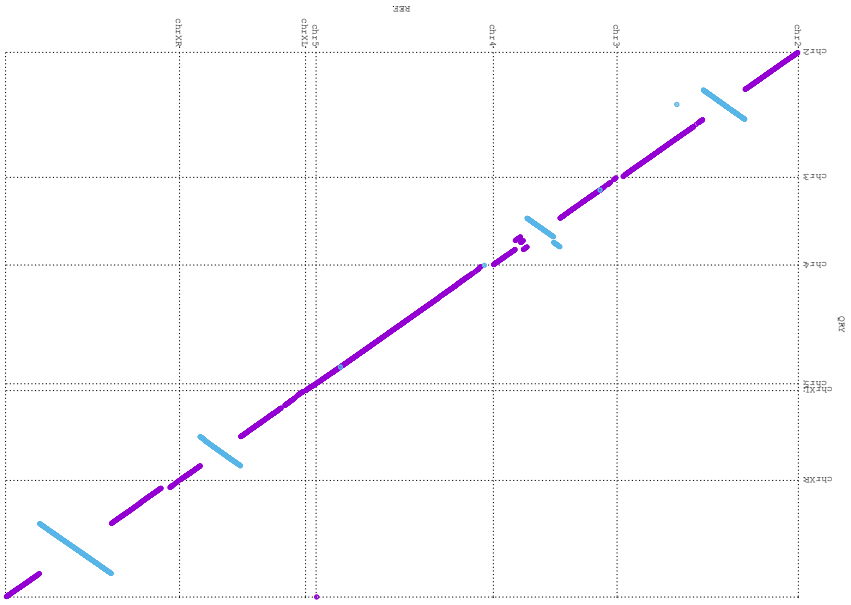


Figure S21. Dot plot for the final assemblies of *D. pseudoobscura* and *D. persimilis*. Purple dots represent co-linear blocks and light blue dots represent inversions in the corresponding chromosomes. REF(x-axis): *D. pseudoobscura*; QRY(y-axis): *D. persimilis*.

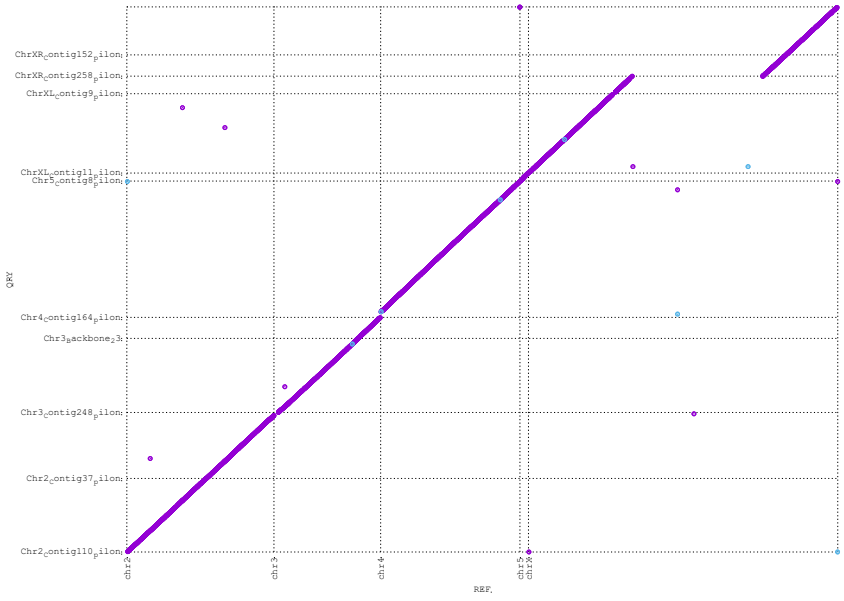


Figure S22. Dot plot comparing our *D. pseudoobscura* assembly versus the *D. pseudoobscura* -MV225, assembly from (Liao et al., 2021). Purple dots represent co-linear blocks and light blue dots represent inversions in the corresponding chromosomes. REF(x-axis): *D. pseudoobscura* – MV225; QRY(y-axis): *D. pseudoobscura* – this study.

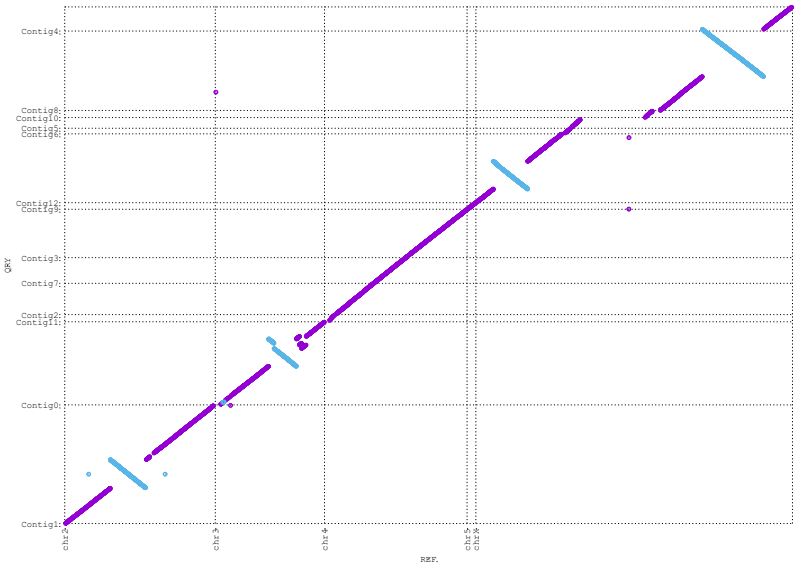


Figure S23. Dot plot comparing our *D. persimilis* assembly versus the *D. pseudoobscura* -MV225, assembly from (Liao et al., 2021). Purple dots represent co-linear blocks and light blue dots represent inversions in the corresponding chromosomes. REF(x-axis): *D. pseudoobscura* – MV225; QRY(y-axis): *D. persimilis* – this study.

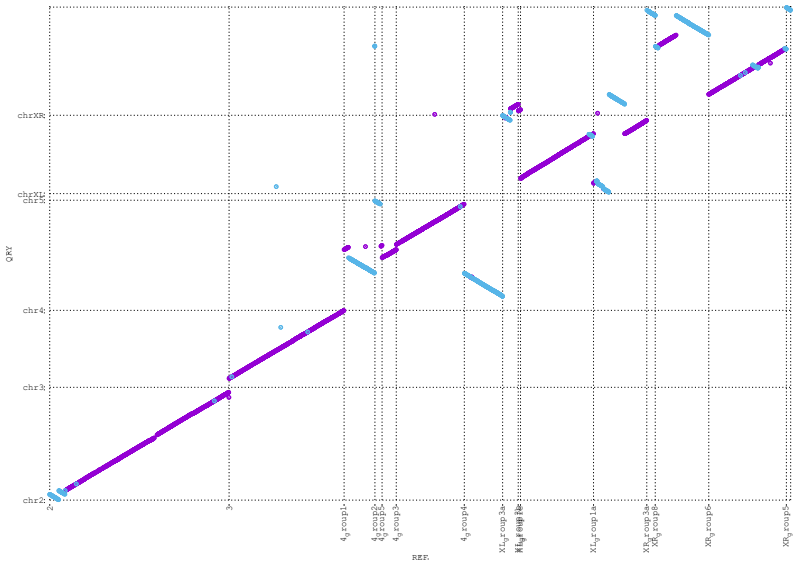


Figure S24. Dot plot comparing our *D. pseudoobscura* assembly versus the *D. pseudoobscura* -MV225, assembly from FlyBase (Thurmond et al., 2019). Purple dots represent co-linear blocks and light blue dots represent inversions in the corresponding chromosomes. REF(x-axis): *D. pseudoobscura* – FlyBase; QRY(y-axis): *D. pseudoobscura* – this study.

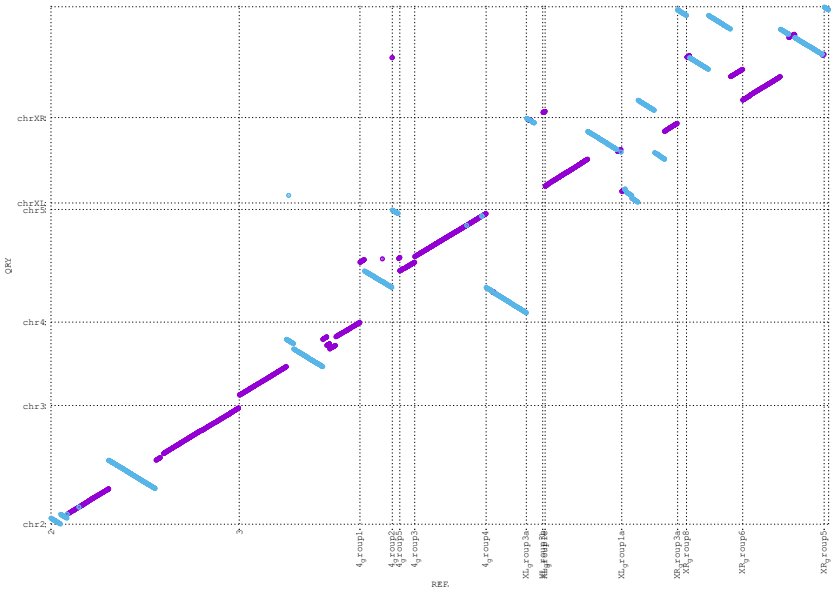


Figure S25. Dot plot comparing our *D. persimilis* assembly versus the *D. pseudoobscura* -MV225, assembly from FlyBase. Purple dots represent co-linear blocks and light blue dots

represent inversions in the corresponding chromosomes. REF(x-axis): *D. pseudoobscura* – FlyBase; QRY(y-axis): *D. persimilis* – this study.

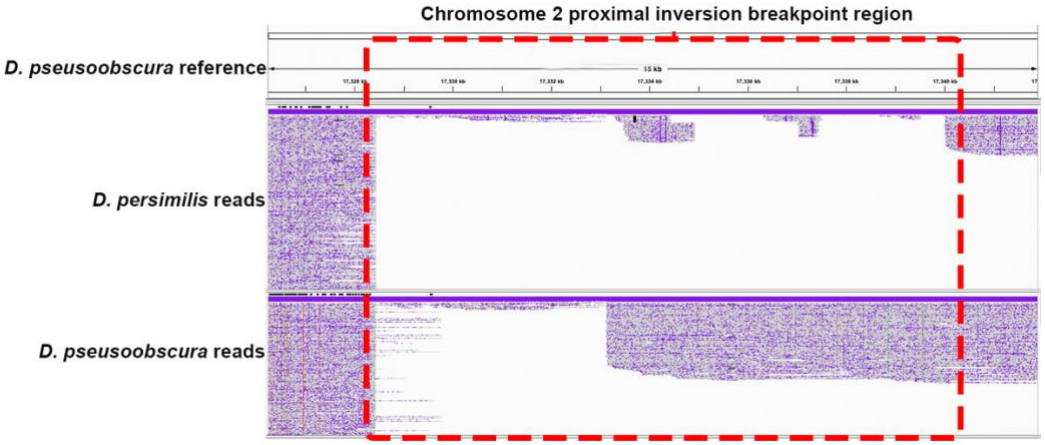


Figure S26. Chromosome 2 proximal inversion breakpoint of *D. pseudoobscura.* Dashed red square indicates the location of the inversion breakpoint region identified in this study.

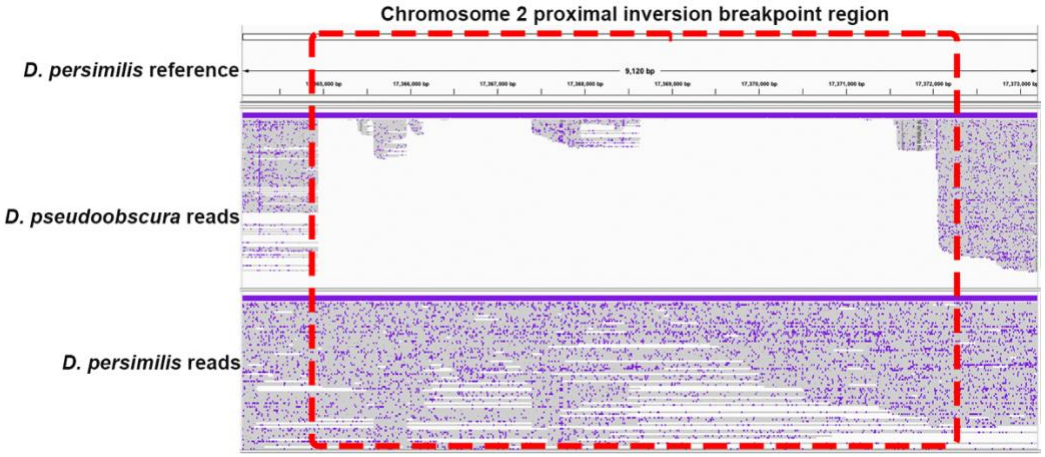


Figure S27. Chromosome 2 distal inversion breakpoint of *D. pseudoobscura.* Dashed red square indicates the location of the inversion breakpoint region identified in this study.

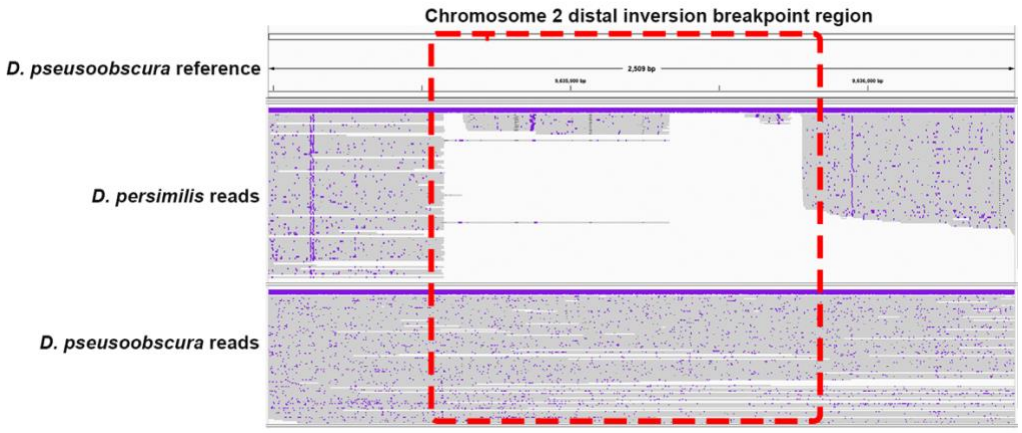


Figure S28. Chromosome 2 proximal inversion breakpoint of *D. pseudoobscura.* Dashed red square indicates the location of the inversion breakpoint region identified in this study.

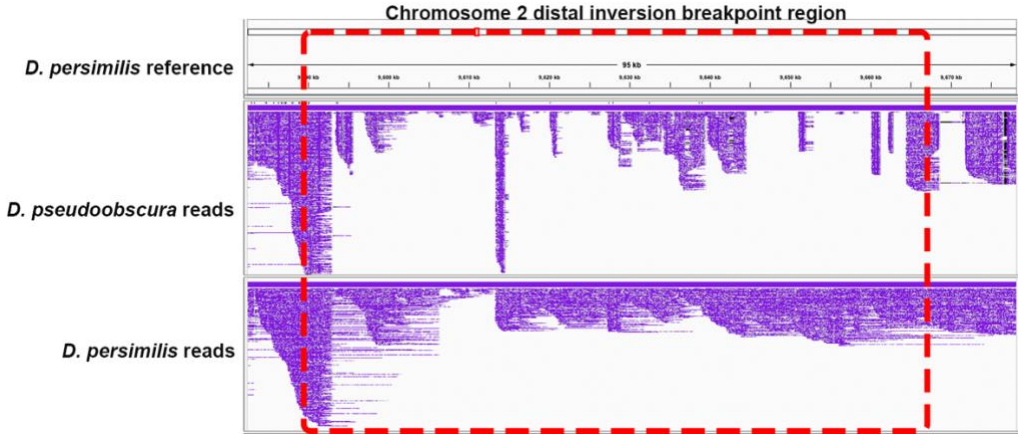


Figure S29. Chromosome 2 distal inversion breakpoint of *D. persimilis.* Dashed red square indicates the location of the inversion breakpoint region identified in this study.

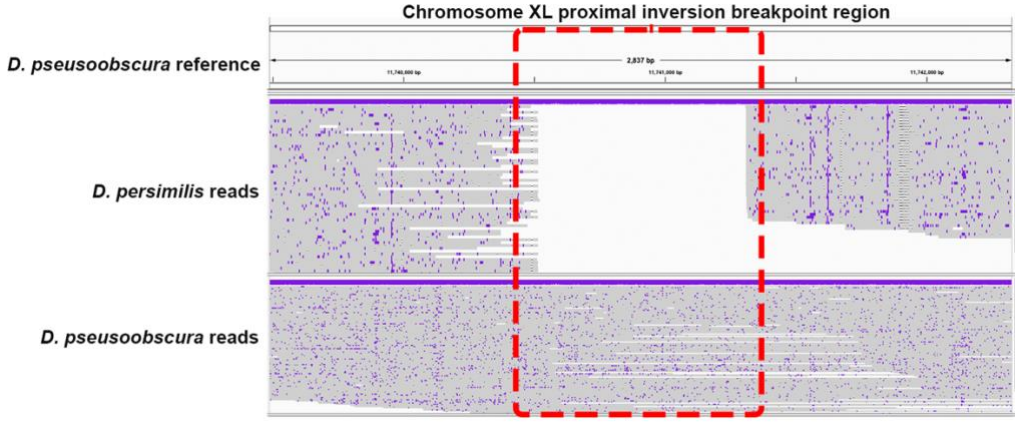


Figure S30. Chromosome XL proximal inversion breakpoint of *D. pseudoobscura.* Dashed red square indicates the location of the inversion breakpoint region identified in this study.

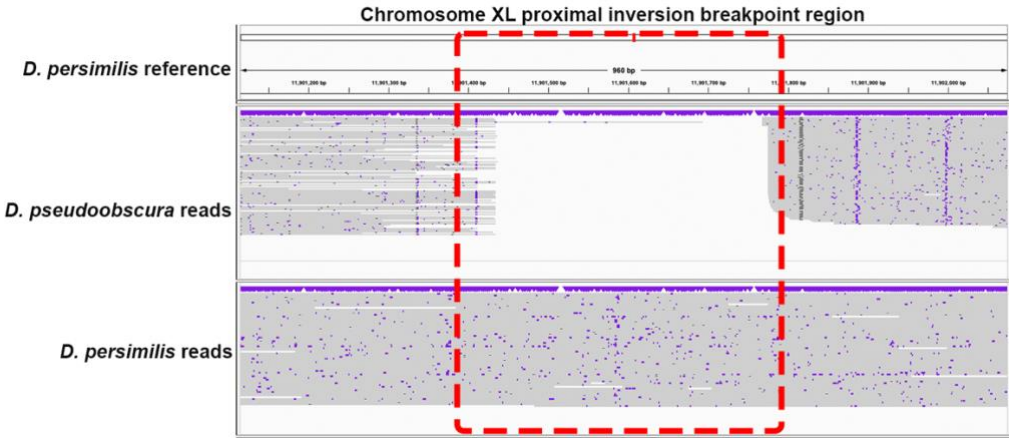


Figure S31. Chromosome XL proximal inversion breakpoint of *D. persimilis.* Dashed red square indicates the location of the inversion breakpoint region identified in this study.

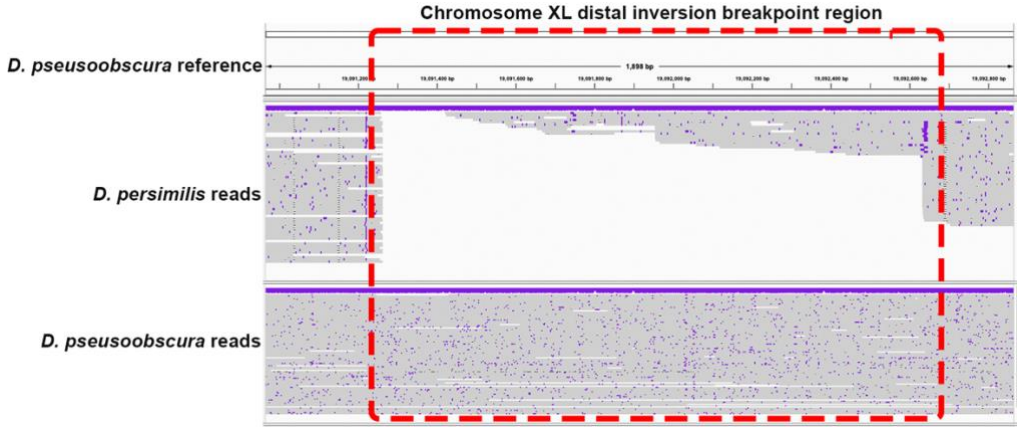


Figure S32. Chromosome XL distal inversion breakpoint of *D. pseudoobscura.* Dashed red square indicates the location of the inversion breakpoint region identified in this study.

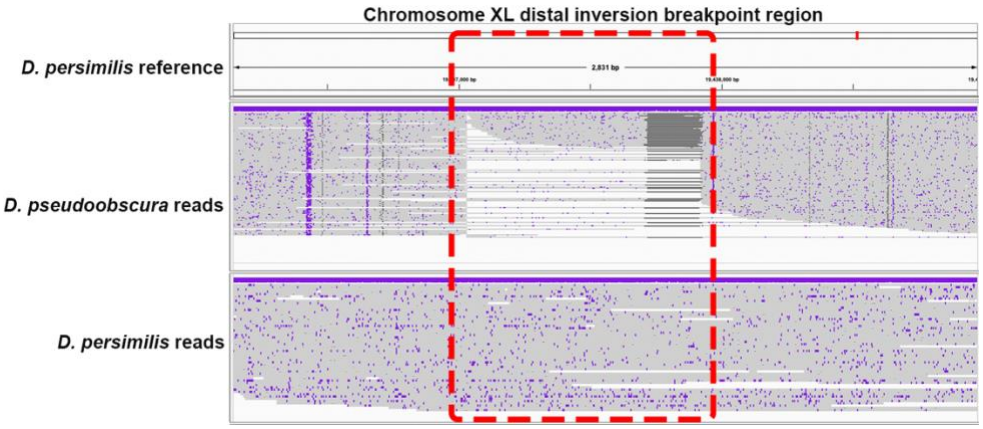


Figure S33. Chromosome XL distal inversion breakpoint of *D. persimilis.* Dashed red square indicates the location of the inversion breakpoint region identified in this study.

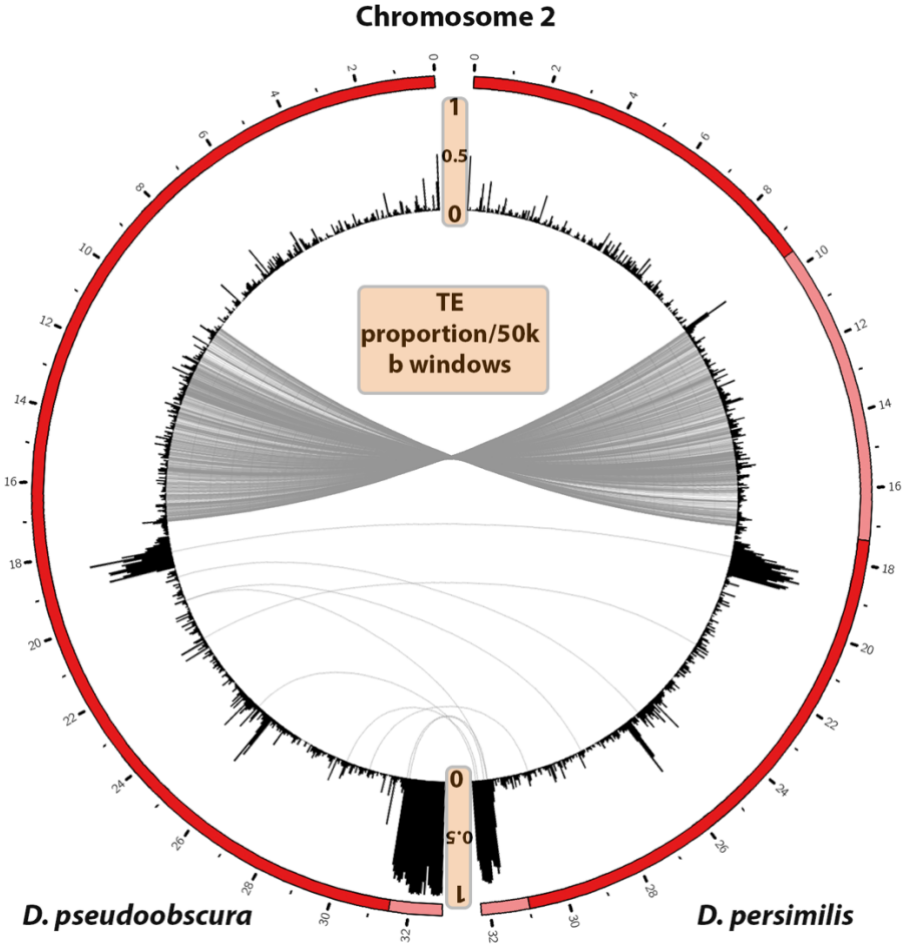


Figure S34. TE proportion for each 50 kb windows in chromosome 2 for *D. pseudoobscura* and *D. persimilis*.

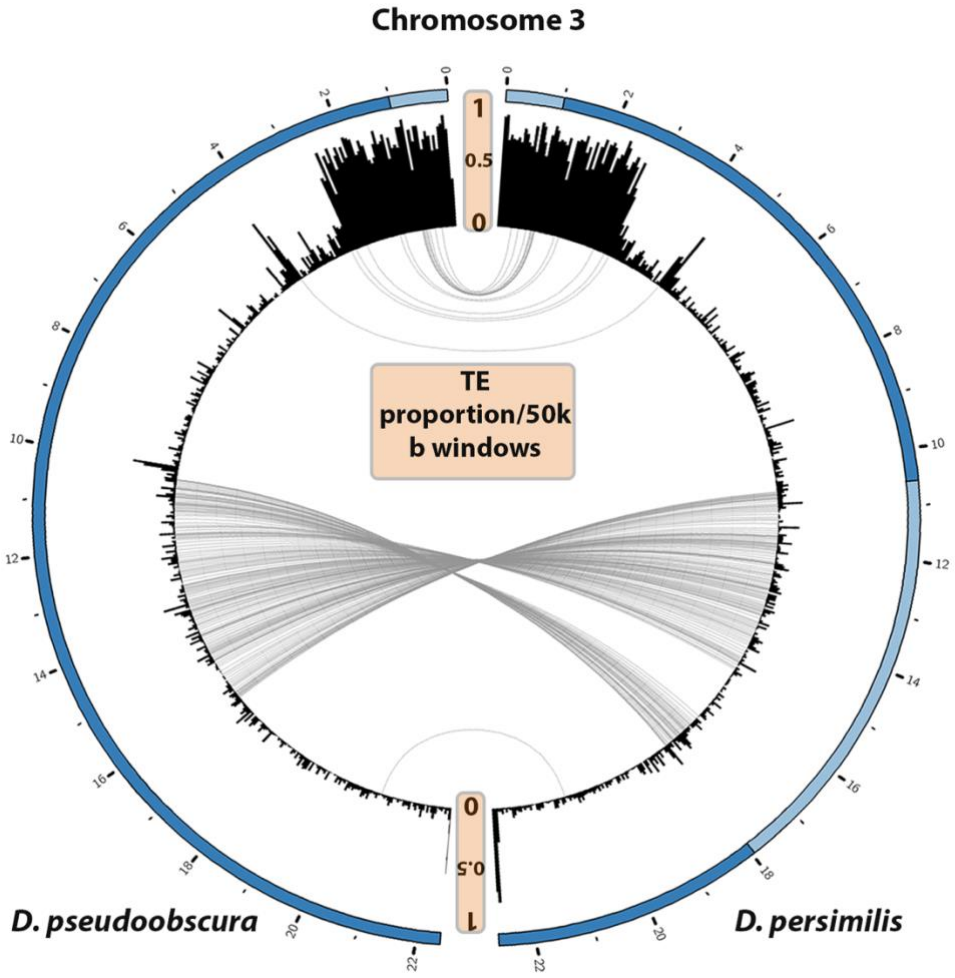


Figure S35. TE proportion for each 50 kb windows in chromosome 3 for *D. pseudoobscura* and *D. persimilis*.

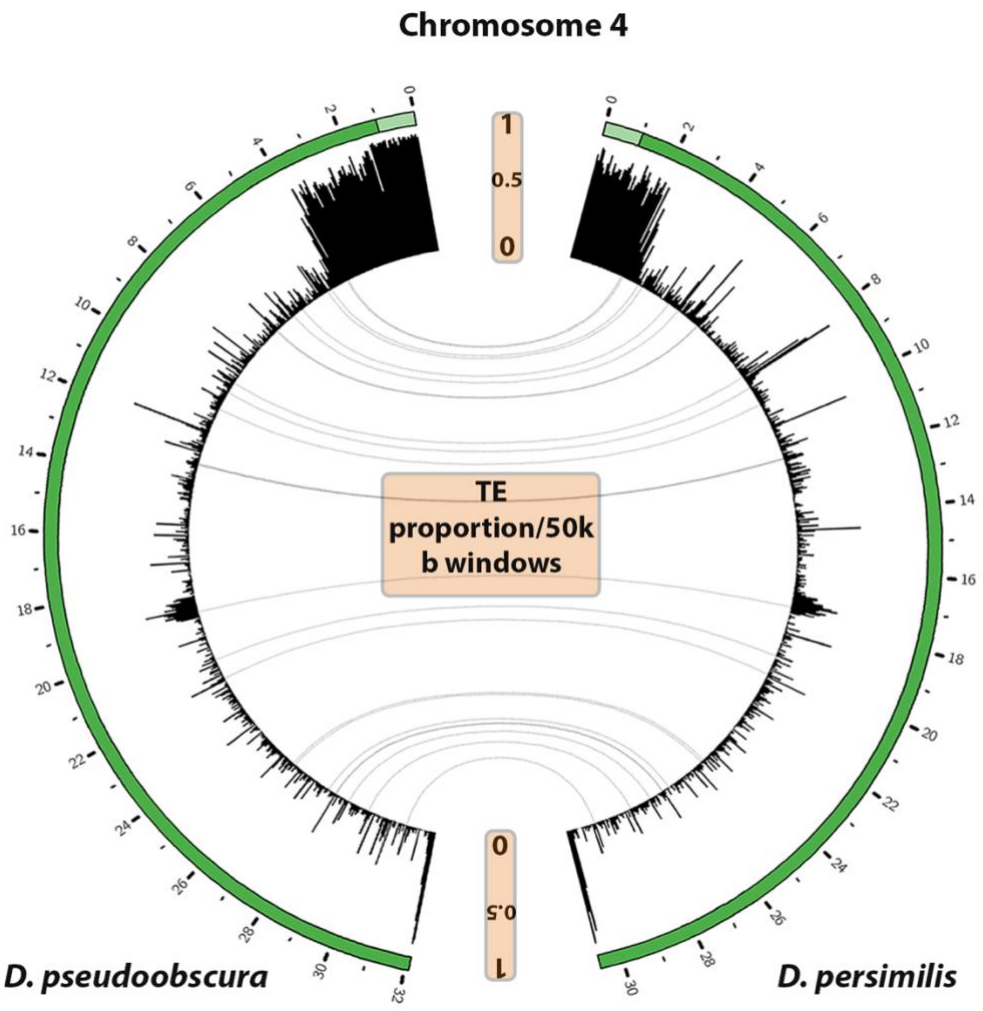


Figure S36. TE proportion for each 50 kb windows in chromosome 4 for *D. pseudoobscura* and *D. persimilis*.

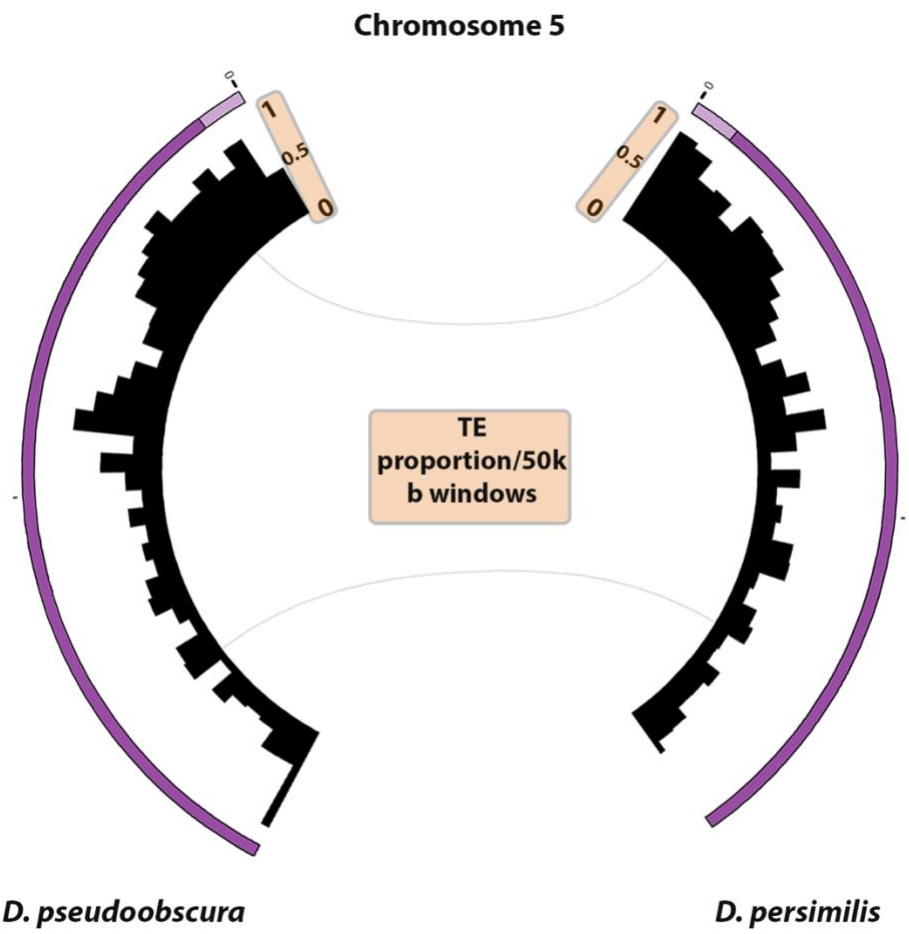


Figure S37. TE proportion for each 50 kb windows in chromosome 5 for *D. pseudoobscura* and *D. persimilis*.

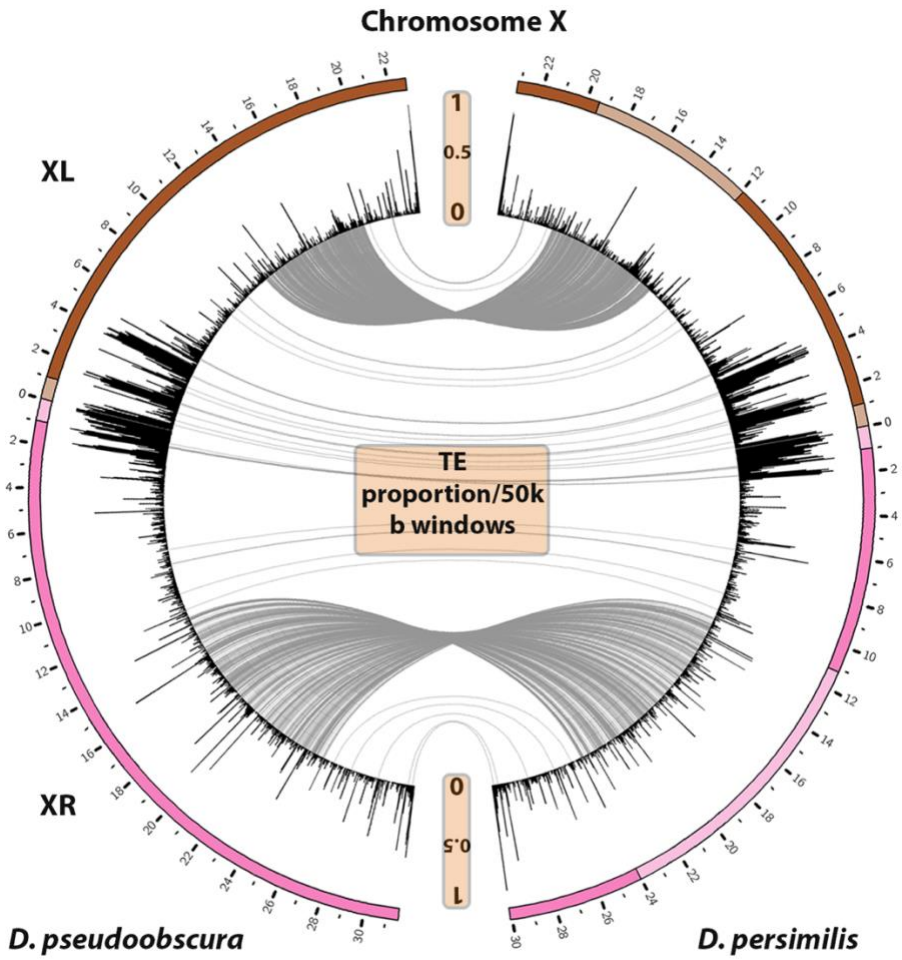


Figure S38. TE proportion for each 50 kb windows in chromosome X for *D. pseudoobscura* and *D. persimilis*.

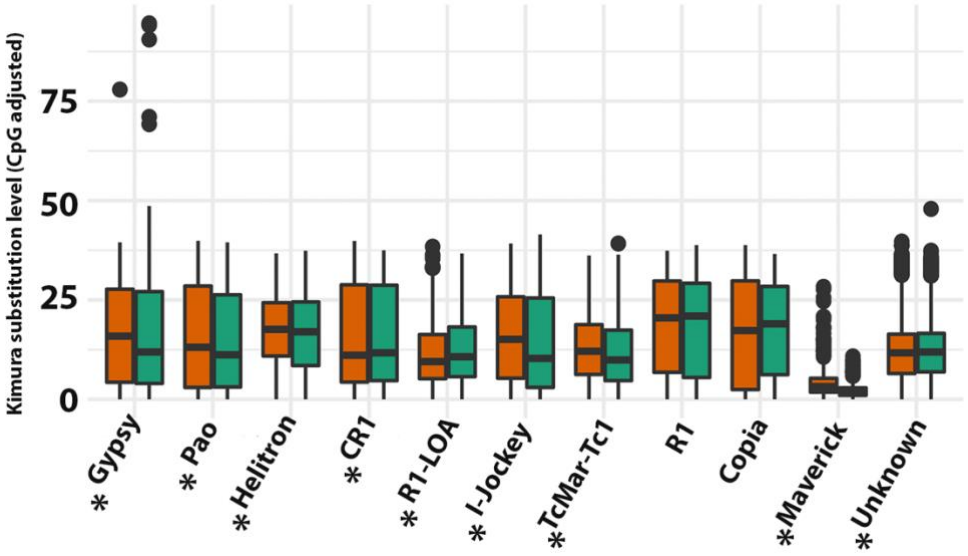


Figure S39. Genome-wide kimura substitution levels of the ten most abundant TE families in both *D. pseudoobscura* (orange) and *D. persimilis* (green). * Significant difference; $p < 0.05$.

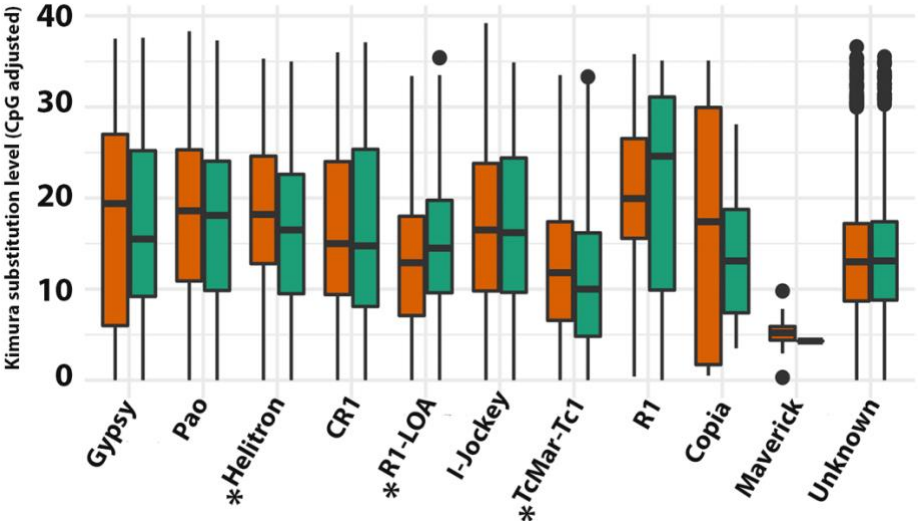


Figure S40. Kimura substitution levels of the ten most abundant TE families in both *D. pseudoobscura* (orange) and *D. persimilis* (green) for chromosome 2. * Significant difference; $p < 0.05$.

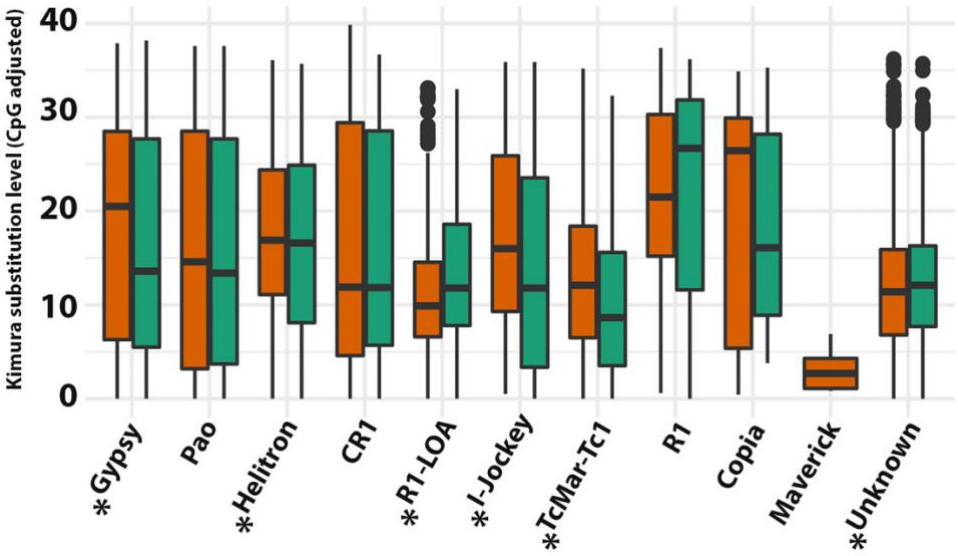


Figure S41. Kimura substitution levels of the ten most abundant TE families in both *D. pseudoobscura* (orange) and *D. persimilis* (green) for chromosome 3. * Significant difference; $p < 0.05$.

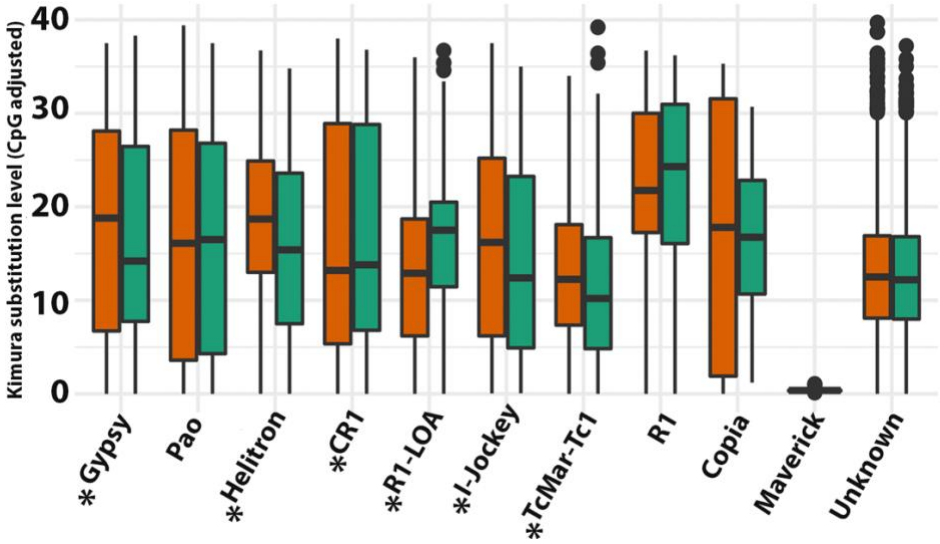


Figure S42. Kimura substitution levels of the ten most abundant TE families in both *D. pseudoobscura* (orange) and *D. persimilis* (green) for chromosome 4. * Significant difference; $p < 0.05$.

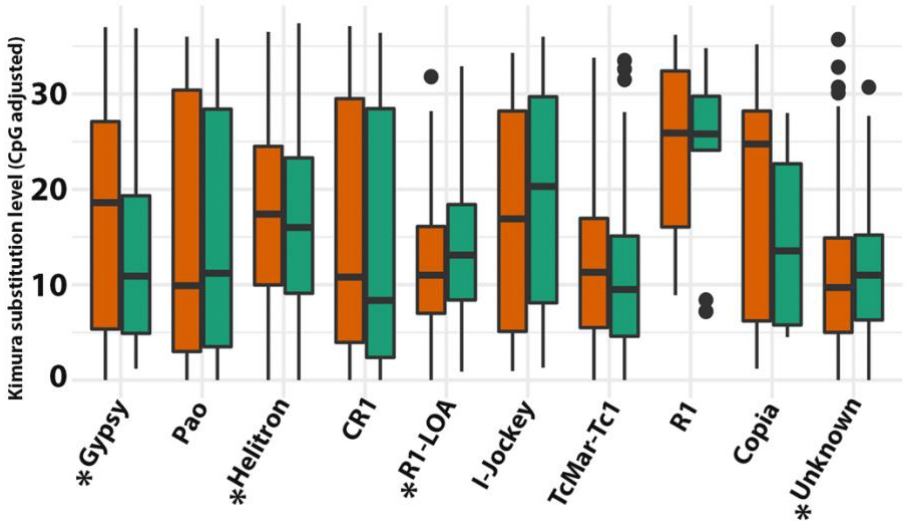


Figure S43. Kimura substitution levels of the ten most abundant TE families in both *D. pseudoobscura* (orange) and *D. persimilis* (green) for chromosome 5. * Significant difference; $p < 0.05$.

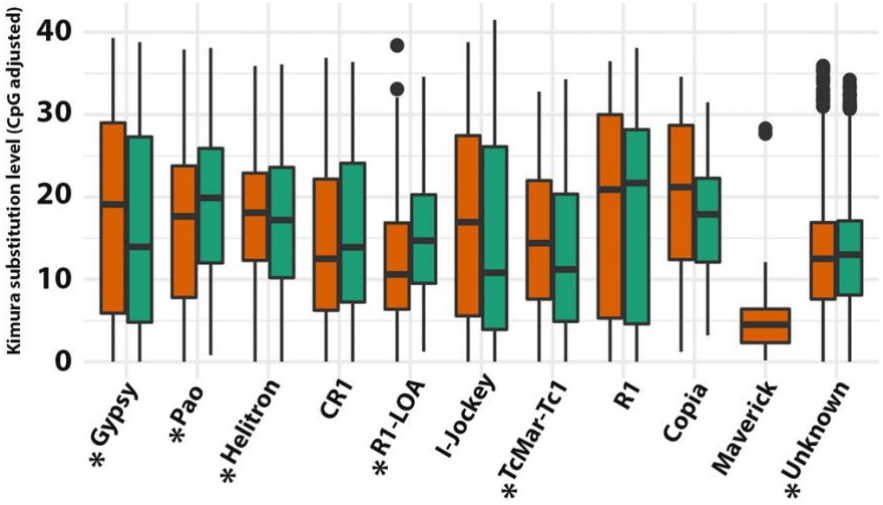


Figure S44. Kimura substitution levels of the ten most abundant TE families in both *D. pseudoobscura* (orange) and *D. persimilis* (green) for chromosome XL. * Significant difference; $p < 0.05$.

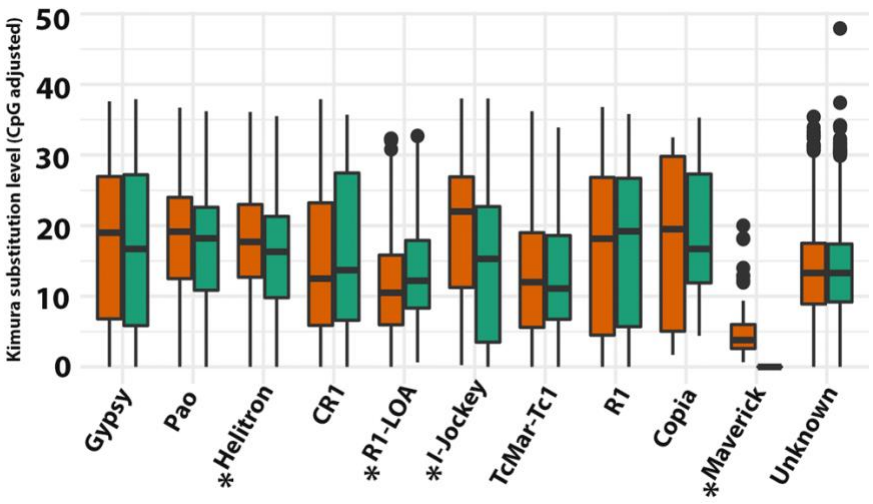


Figure S45. Kimura substitution levels of the ten most abundant TE families in both *D. pseudoobscura* (orange) and *D. persimilis* (green) for chromosome XR. * Significant difference; $p < 0.05$.

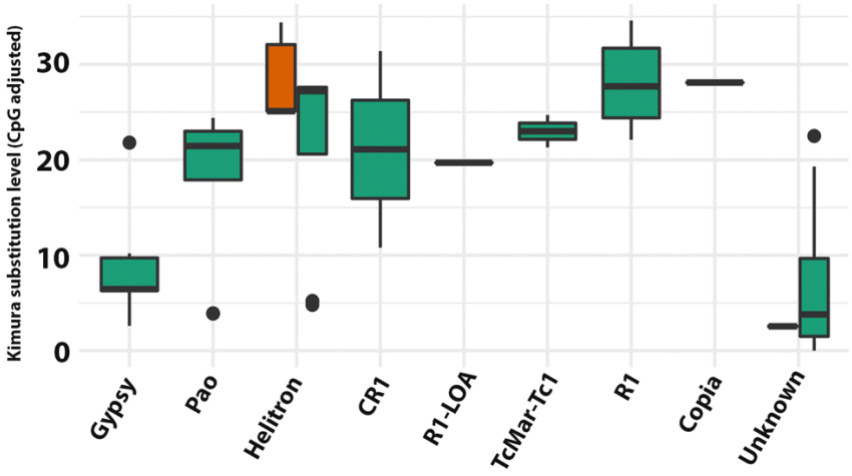


Figure S46. Kimura substitution levels of the TE families present at the proximal and distal inversion breakpoints in *D. pseudoobscura* (orange) and *D. persimilis* (green), respectively for chromosome 2. Gyspy, Pao, CR1, R1-LOA, TcMar-Tc1, R1 and Copia are present only in *D. persimilis*.

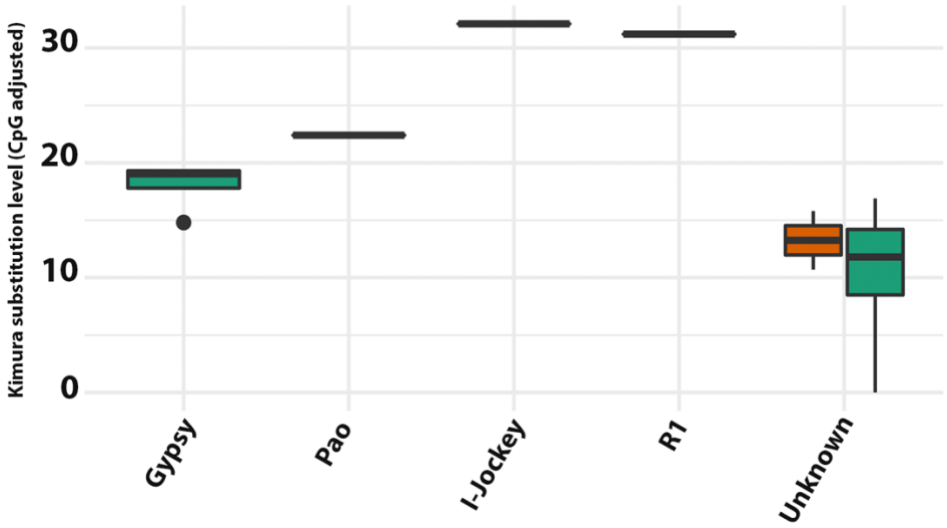


Figure S47. Kimura substitution levels of the TE families present at the proximal and distal inversion breakpoints in *D. persimilis* (green) and *D. pseudoobscura* (orange), respectively for chromosome 2. Gyspy, Pao, I -Jockey and R1 are present only in *D. persimilis.*

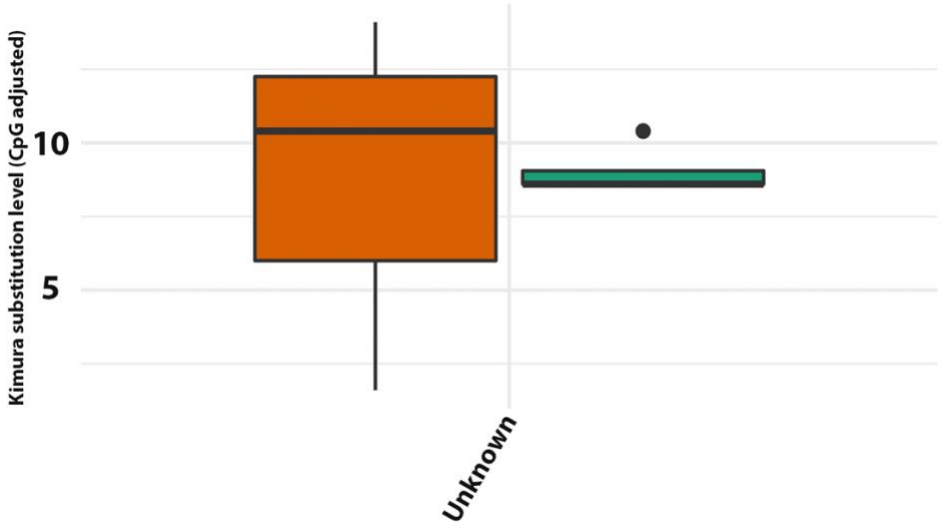


Figure S48. Kimura substitution levels of the TE families present at the proximal and distal inversion breakpoints in *D. pseudoobscura* (orange) and *D. persimilis* (green), respectively for chromosome XL.

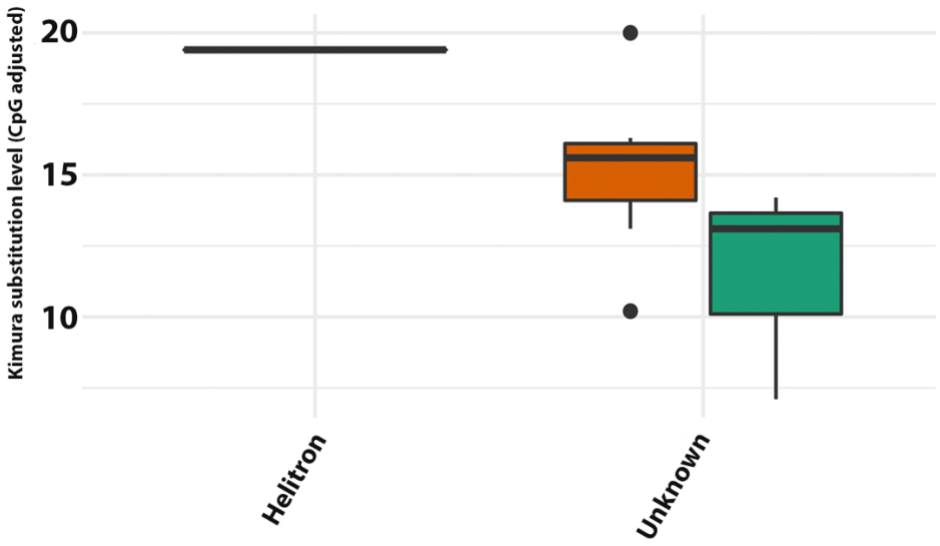


Figure S49. Kimura substitution levels of the TE families present at the proximal and distal inversion breakpoints in *D. persimilis* (green) and *D. pseudoobscura* (orange), respectively for chromosome XL. Helitron is present only in *D. pseudoobscura.*

Supplementary Tables

Table S1. Differentially expressed genes located inside the inversions of chromosomes 2, 3, and X.

Dpse	Dper	Gene name	Expression	Chromosome
gene-215575	**gene-107339**	cnc	1L	chr2
gene-230458	**gene-94538**	heph	1L	chr2
gene-80899	**gene-250095**	Ten-m	1L	chrXR
gene-73554	**gene-256505**	CG7638/emei	1L	chrXR
gene-77232	**gene-253161**	Spn	1L, Ad	chrXR
gene-234029	**gene-91618**	Task6	3L	chr2
gene-176344	**gene-41817**	Su(var)2-10	1L, 3L, Pup, Ad	chr3
gene-52708	**gene-239812**	CG14997	3L	chrXR
gene-72853	**gene-257164**	CG7560	3L, Pup	chrXR
gene-175133	**gene-43174**	St3	1L, Pup, Ad	chr3
gene-229787	**gene-95159**	Sgt1	1L, 3L, Pup, Ad	chr2
gene-230679	**gene-94412**	Map205	1L	chr2
gene-170813	**gene-51095**	dila	1L	chr3
gene-80899	**gene-250095**	Ten-m	1L	chrXR
gene-53767	**gene-238589**	cmpy	1L, Pup	chrXR
gene-75351	**gene-254856**	CG10960/nebu	Ad	chrXR

Appendix B: Supplementary Materials for Chapter2

Supplementary Figures

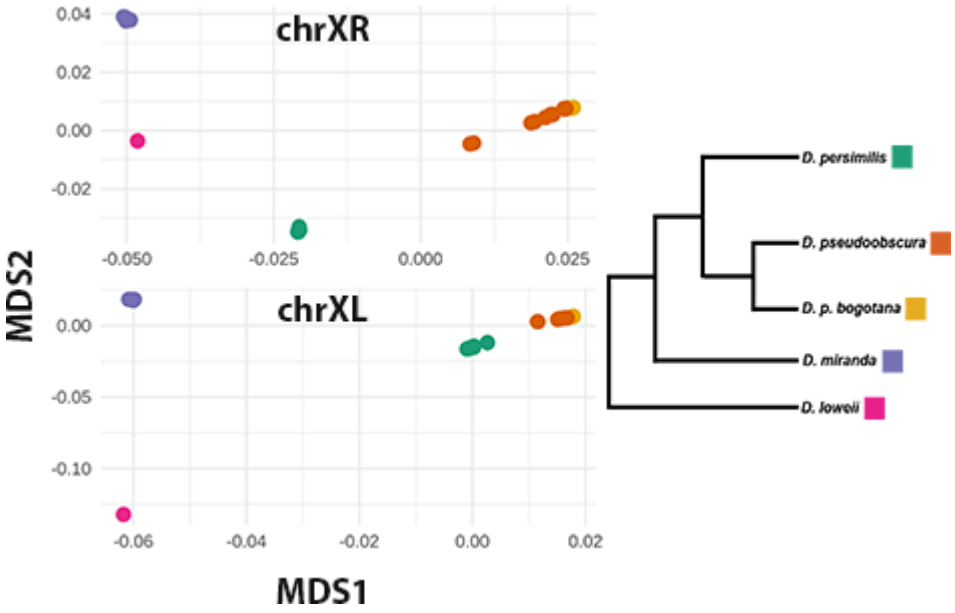


Figure S1. Multidimensional scaling plots for chromosome X divided by the two arms. Plots show the IBS relationships between *D. persimilis*, *D. pseudoobscura*, *D. p. bogotana*, *D. miranda* and *D. loweii* (cladogram).

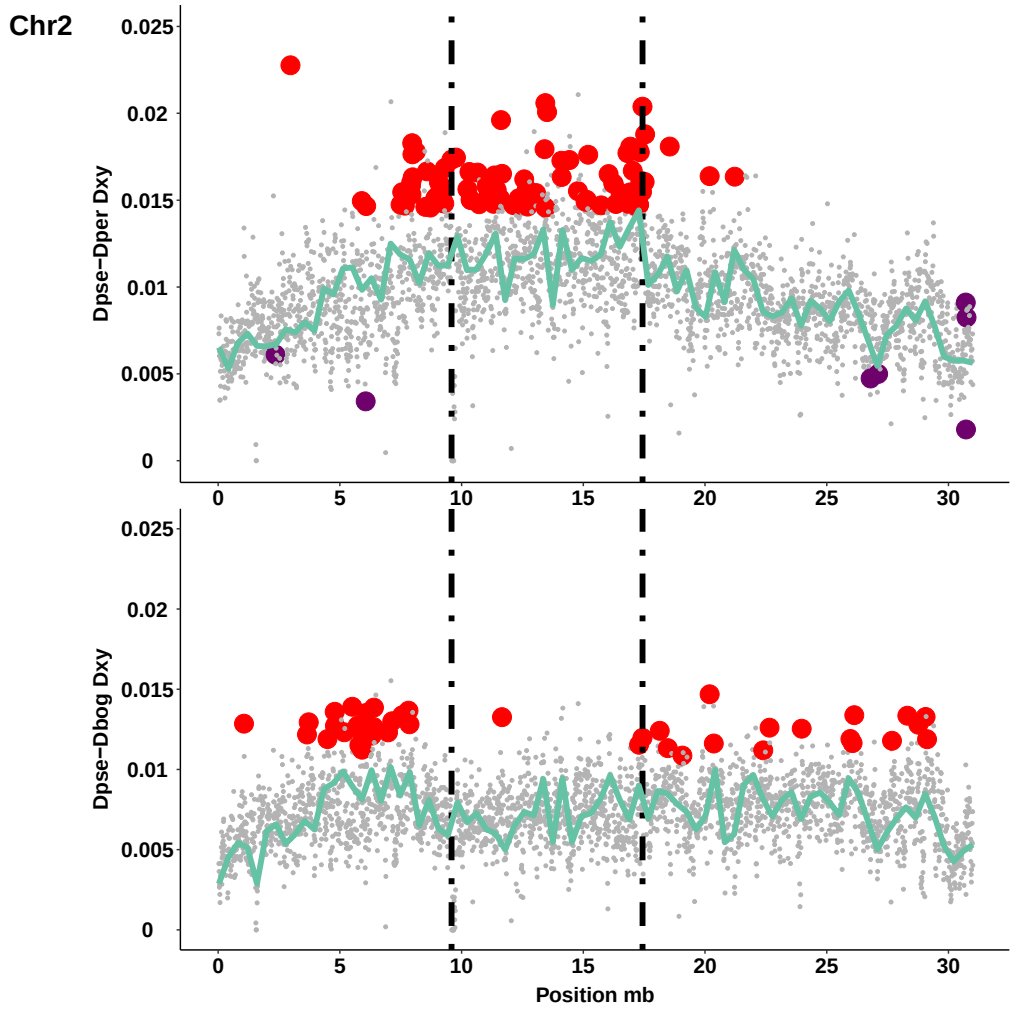


Figure S2. Dxy landscape for chromosome 2 for two species comparisons: *D. pseudoobscura* vs *D. persimilis* (Dpse-Dper; top), and *D. pseudoobscura* vs *D. p. bogotana* (Dpse-Dbog; bottom). Each dot represents a 10 kb non-overlapping window. Red (divergent) and purple (introgressed) dots show outlier regions based on kNN analyses. Horizontal dashed lines mark the inversion breakpoints occurring between *D. pseudoobscura* and *D. persimilis.*

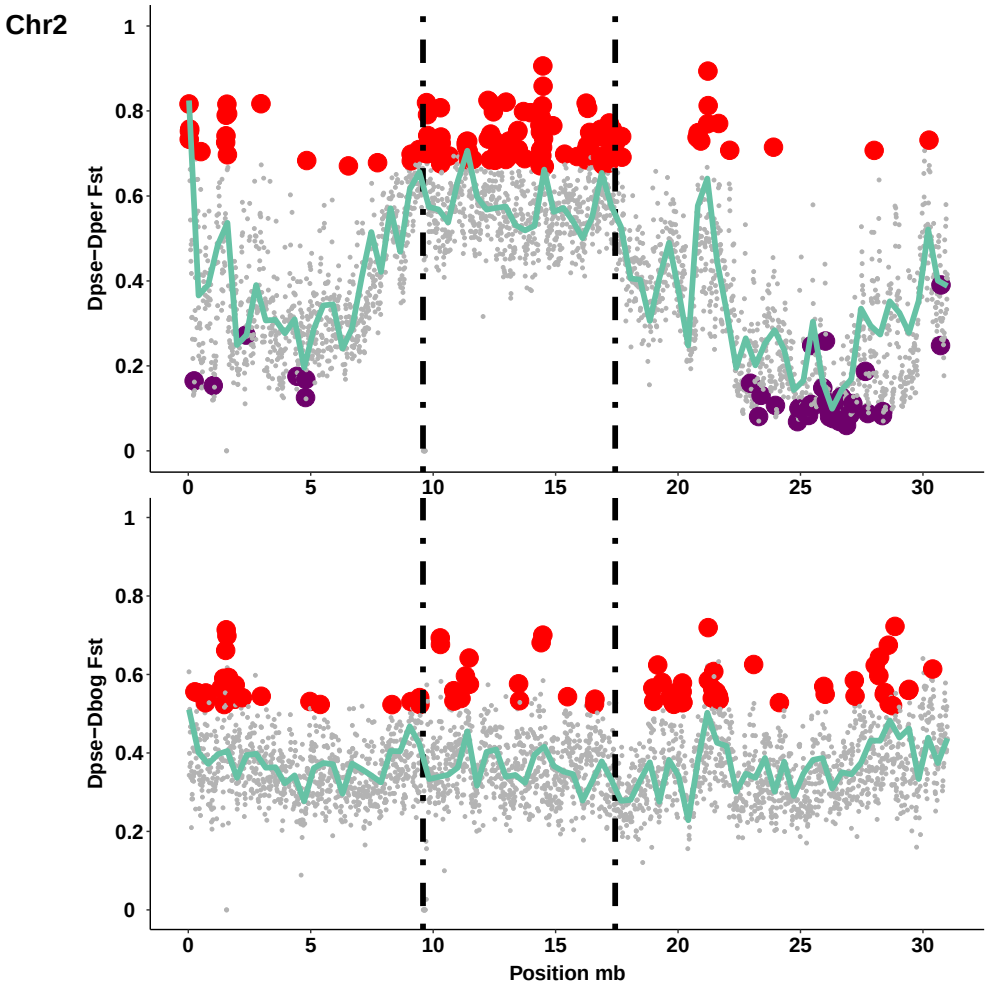


Figure S3. Fst landscape for chromosome 2 for two species comparisons: *D. pseudoobscura* vs *D. persimilis* (Dpse-Dper; top), and *D. pseudoobscura* vs *D. p. bogotana* (Dpse-Dbog; bottom). Each dot represents a 10 kb non-overlapping window. Red (divergent) and purple (introgressed) dots show outlier regions based on kNN analyses. Horizontal dashed lines mark the inversion breakpoints occurring between *D. pseudoobscura* and *D. persimilis.*

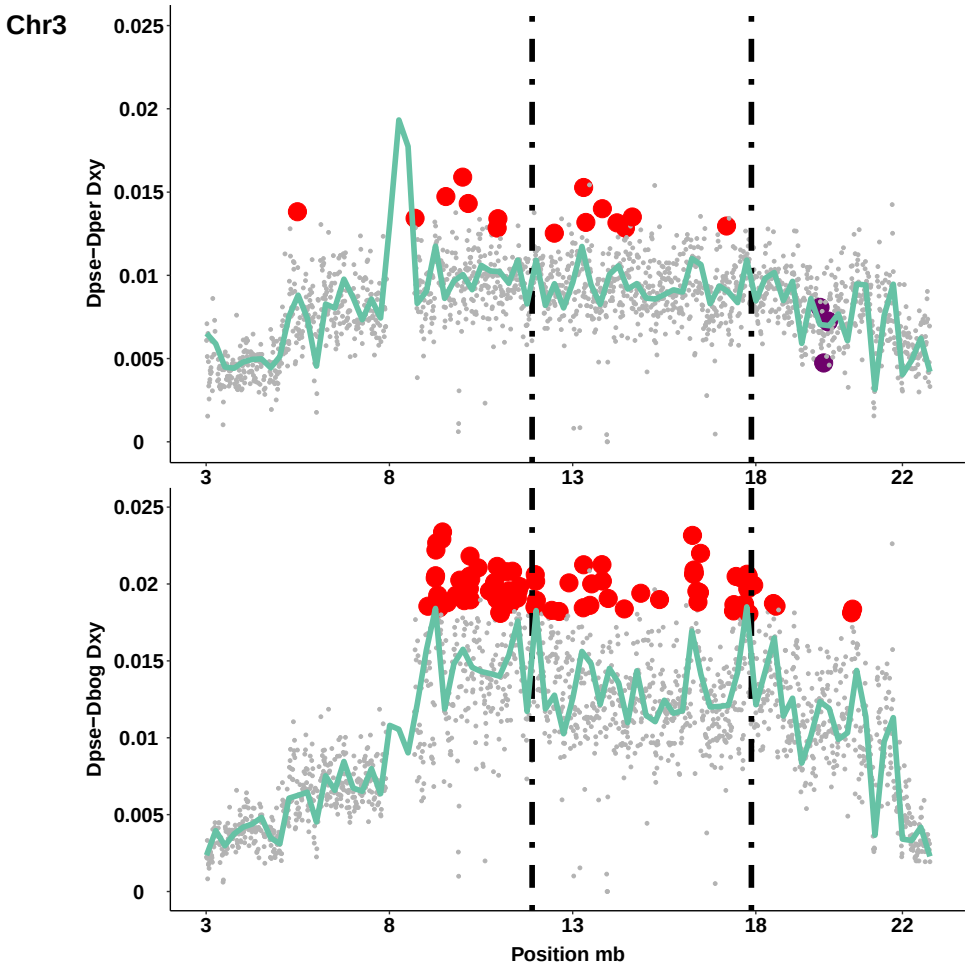


Figure S4. Dxy landscape for chromosome 3 for two species comparisons: *D. pseudoobscura* vs *D. persimilis* (Dpse-Dper; top), and *D. pseudoobscura* vs *D. p. bogotana* (Dpse-Dbog; bottom). Each dot represents a 10 kb non-overlapping window. Red (divergent) and purple (introgressed) dots show outlier regions based on kNN analyses. Horizontal dashed lines mark the inversion breakpoints occurring between *D. pseudoobscura* and *D. persimilis.*

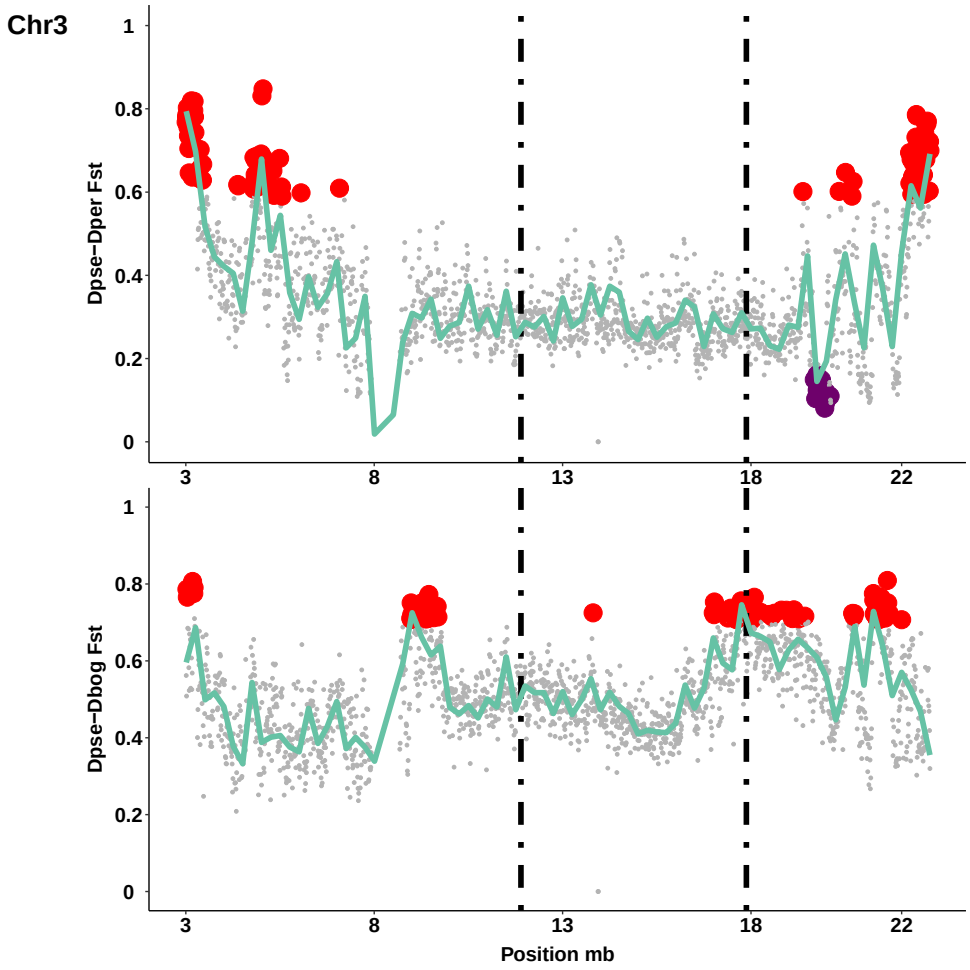


Figure S5. Fst landscape for chromosome 3 for two species comparisons: *D. pseudoobscura* vs *D. persimilis* (Dpse-Dper; top), and *D. pseudoobscura* vs *D. p. bogotana* (Dpse-Dbog; bottom). Each dot represents a 10 kb non-overlapping window. Red (divergent) and purple (introgressed) dots show outlier regions based on kNN analyses. Horizontal dashed lines mark the inversion breakpoints occurring between *D. pseudoobscura* and *D. persimilis.*

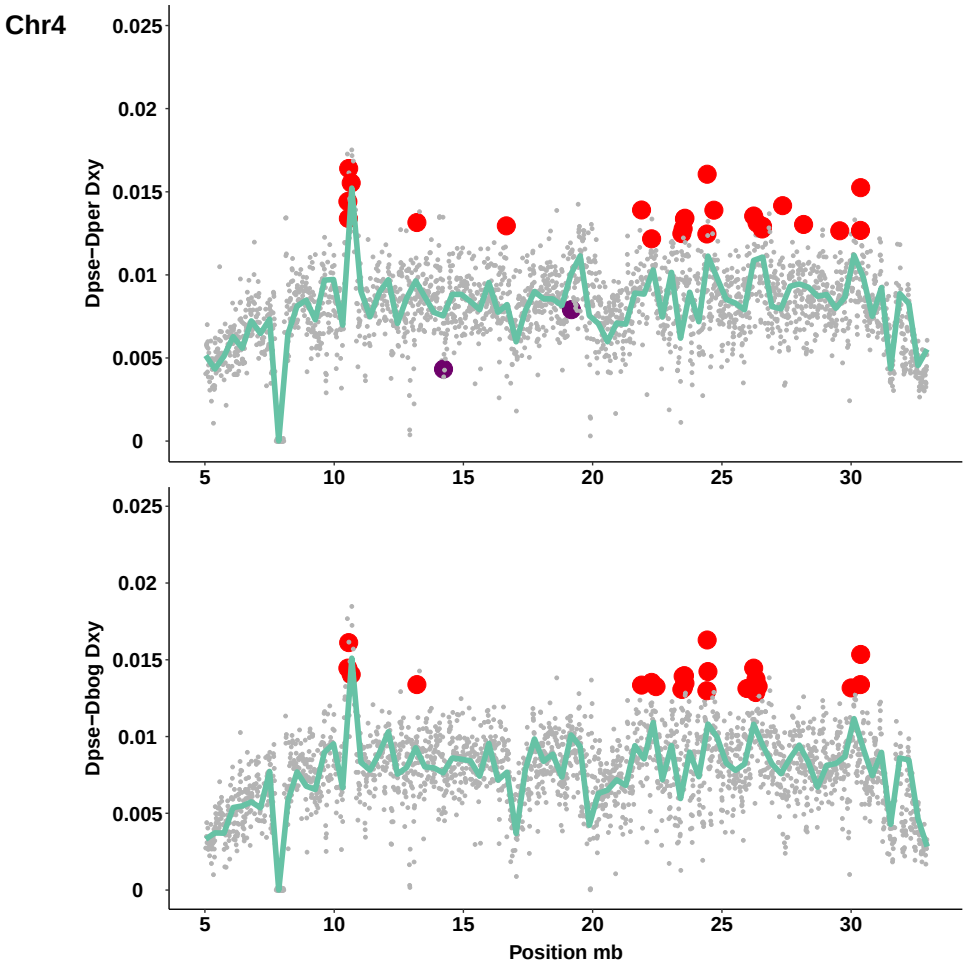


Figure S6. Dxy landscape for chromosome 4 for two species comparisons: *D. pseudoobscura* vs *D. persimilis* (Dpse-Dper; top), and *D. pseudoobscura* vs *D. p. bogotana* (Dpse-Dbog; bottom). Each dot represents a 10 kb non-overlapping window. Red (divergent) and purple (introgressed) dots show outlier regions based on kNN analyses.

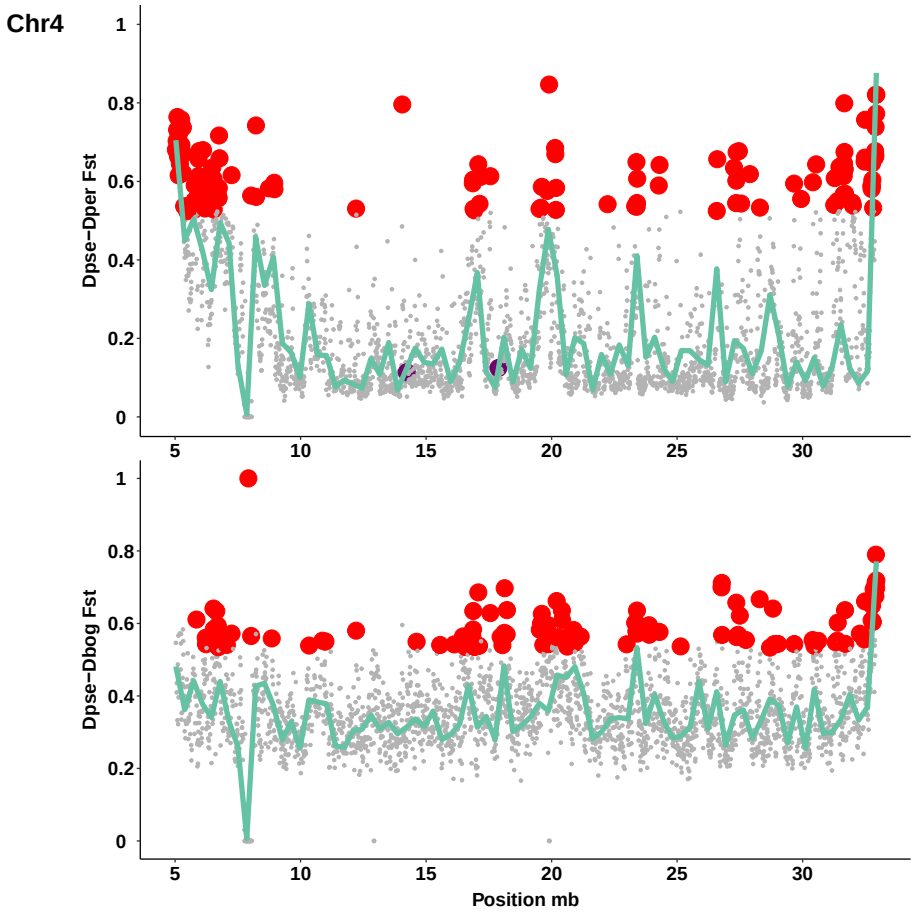


Figure S7. Fst landscape for chromosome 4 for two species comparisons: *D. pseudoobscura* vs *D. persimilis* (Dpse-Dper; top), and *D. pseudoobscura* vs *D. p. bogotana* (Dpse-Dbog; bottom). Each dot represents a 10 kb non-overlapping window. Red (divergent) and purple (introgressed) dots show outlier regions based on kNN analyses.

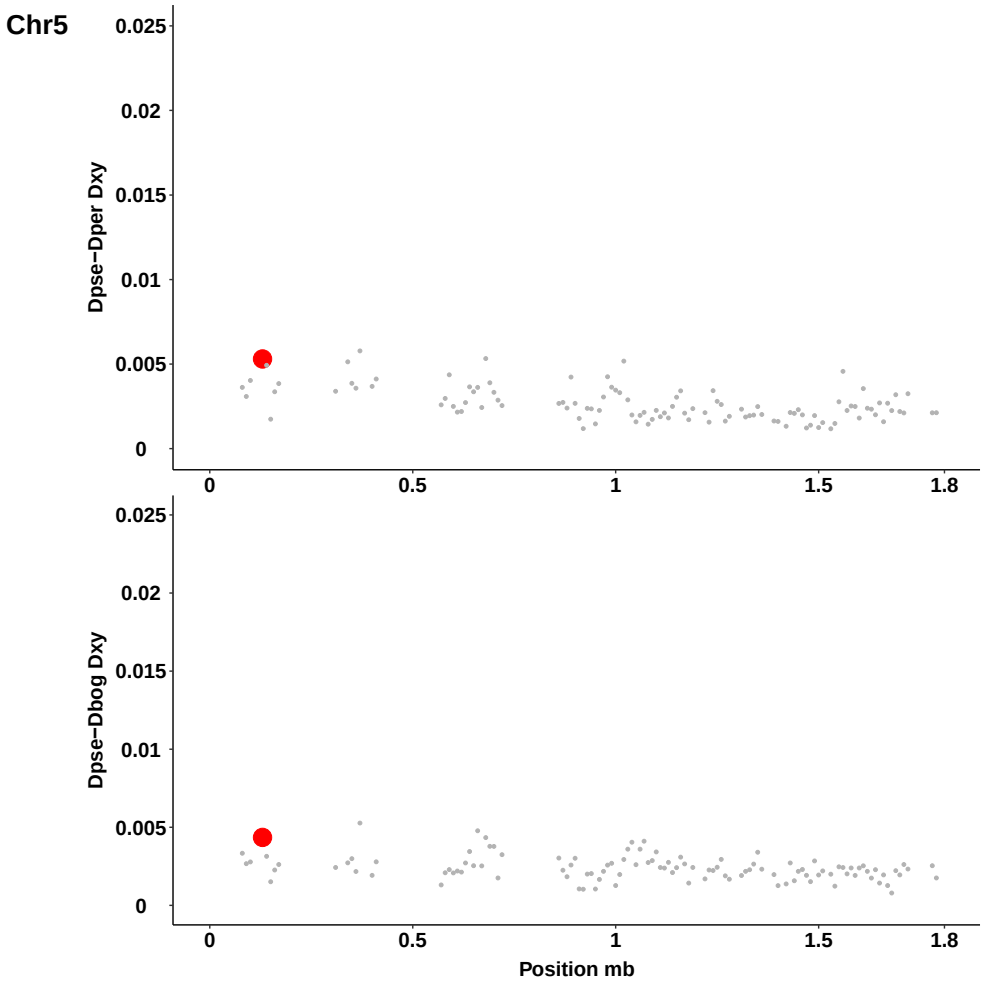


Figure S8. Dxy landscape for chromosome 5 for two species comparisons: *D. pseudoobscura* vs *D. persimilis* (Dpse-Dper; top), and *D. pseudoobscura* vs *D. p. bogotana* (Dpse-Dbog; bottom). Each dot represents a 10 kb non-overlapping window. Red (divergent) and purple (introgressed) dots show outlier regions based on kNN analyses.

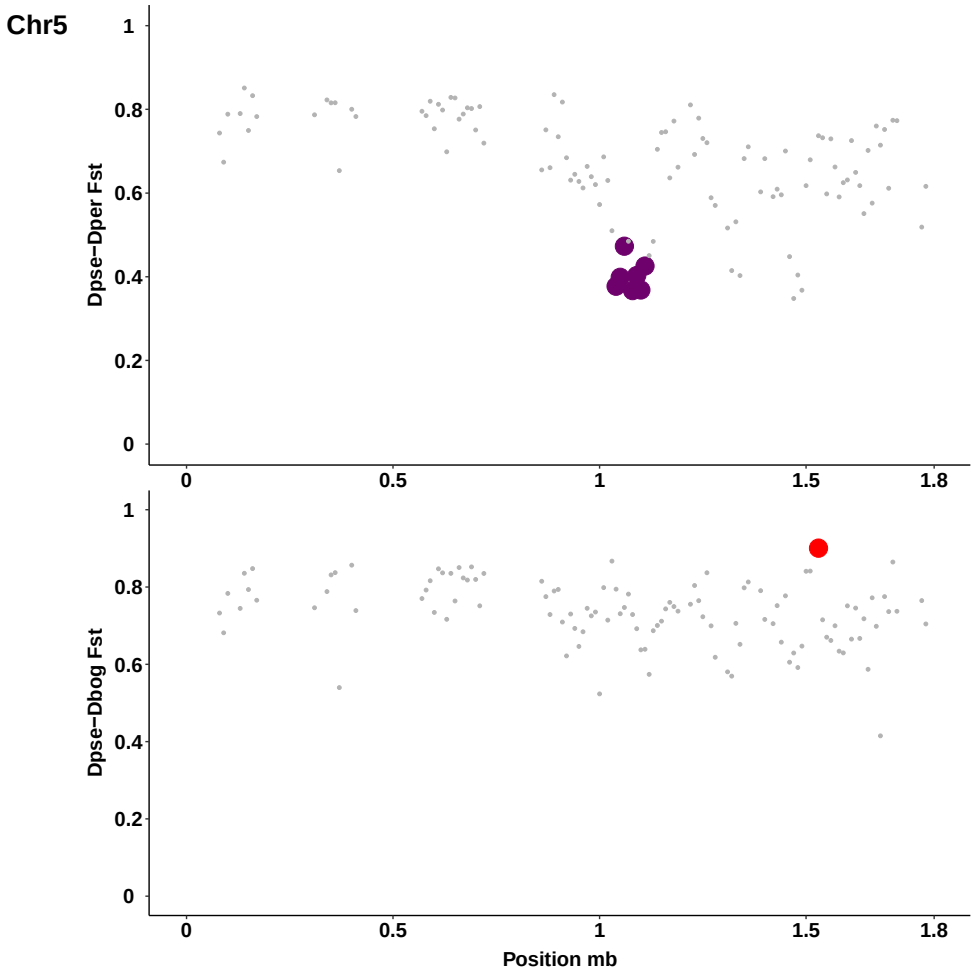


Figure S9. Fst landscape for chromosome 5 for two species comparisons: *D. pseudoobscura* vs *D. persimilis* (Dpse-Dper; top), and *D. pseudoobscura* vs *D. p. bogotana* (Dpse-Dbog; bottom). Each dot represents a 10 kb non-overlapping window. Red (divergent) and purple (introgressed) dots show outlier regions based on kNN analyses.

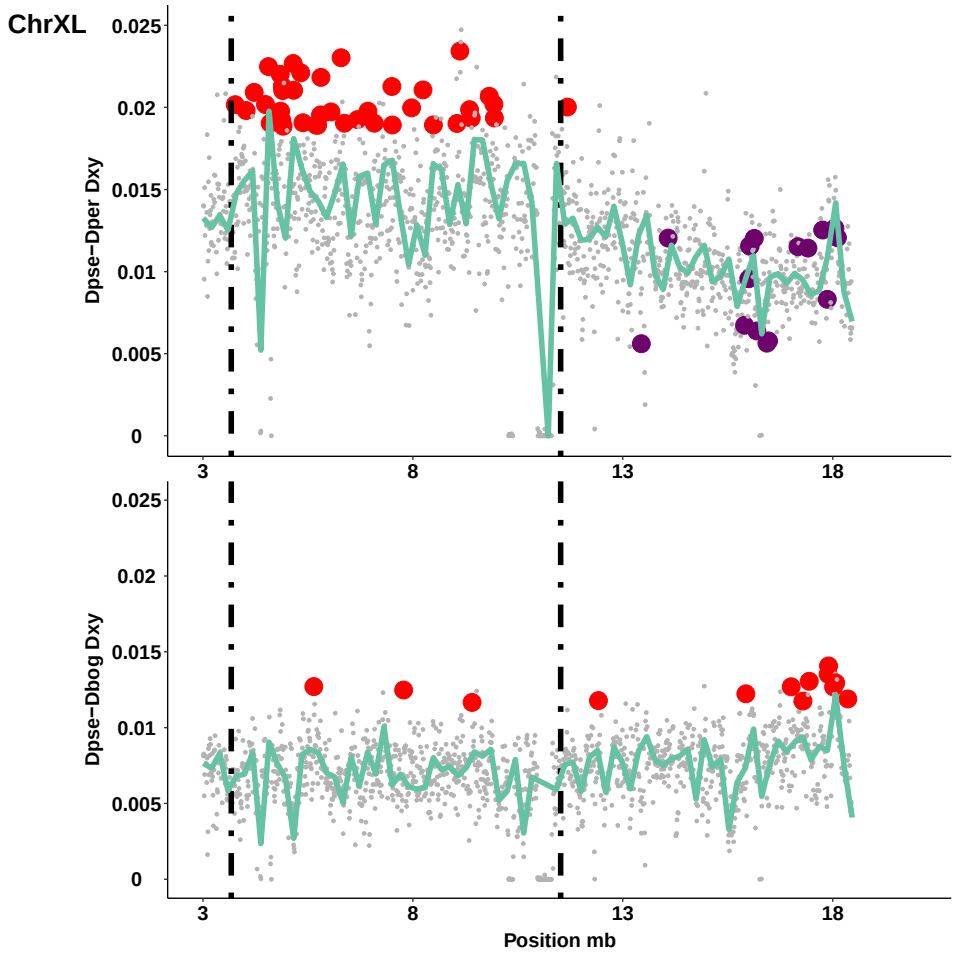


Figure S10. Dxy landscape for chromosome XL for two species comparisons: *D. pseudoobscura* vs *D. persimilis* (Dpse-Dper; top), and *D. pseudoobscura* vs *D. p. bogotana* (Dpse-Dbog; bottom). Each dot represents a 10 kb non-overlapping window. Red (divergent) and purple (introgressed) dots show outlier regions based on kNN analyses.

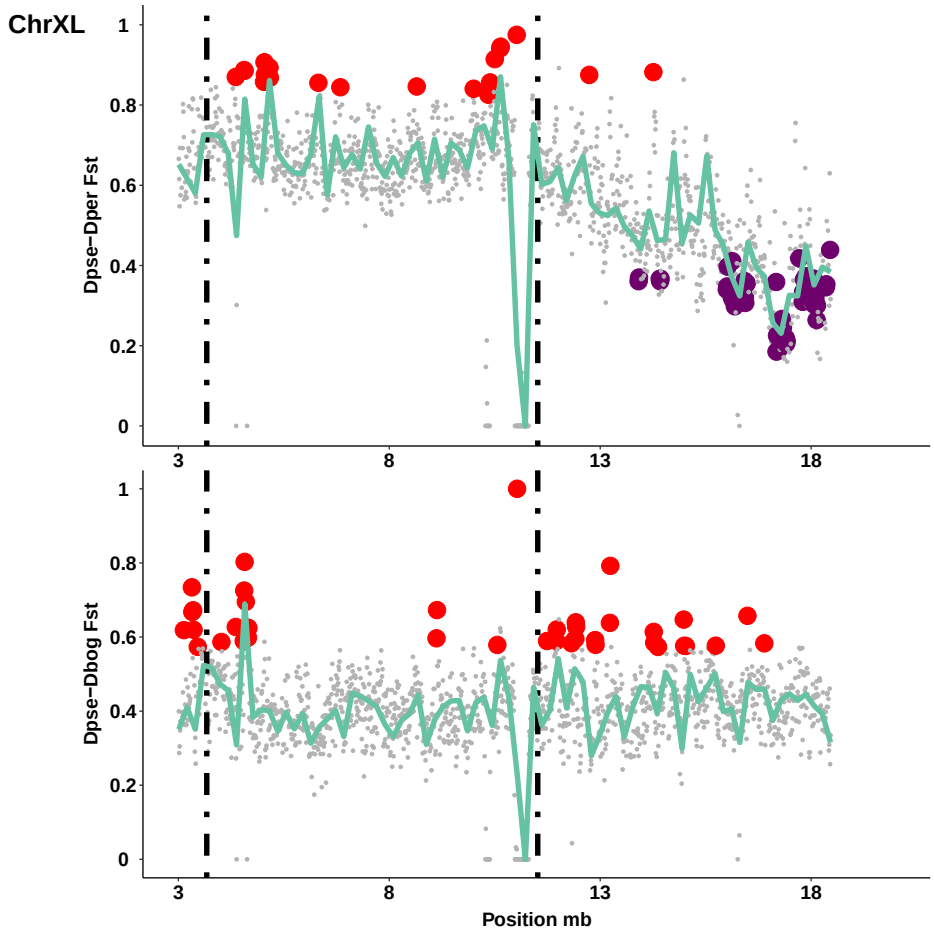


Figure S11. Fst landscape for chromosome XL for two species comparisons: *D. pseudoobscura* vs *D. persimilis* (Dpse-Dper; top), and *D. pseudoobscura* vs *D. p. bogotana* (Dpse-Dbog; bottom). Each dot represents a 10 kb non-overlapping window. Red (divergent) and purple (introgressed) dots show outlier regions based on kNN analyses.

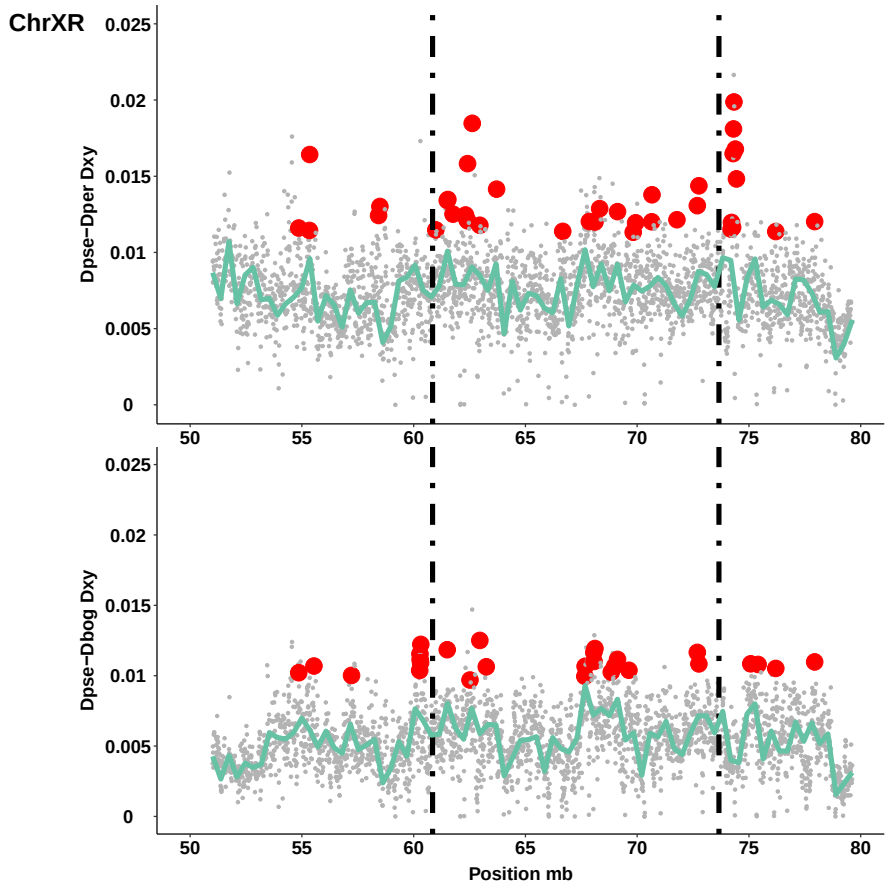


Figure S12. Dxy landscape for chromosome XR for two species comparisons: *D. pseudoobscura* vs *D. persimilis* (Dpse-Dper; top), and *D. pseudoobscura* vs *D. p. bogotana* (Dpse-Dbog; bottom). Each dot represents a 10 kb non-overlapping window. Red (divergent) and purple (introgressed) dots show outlier regions based on kNN analyses.

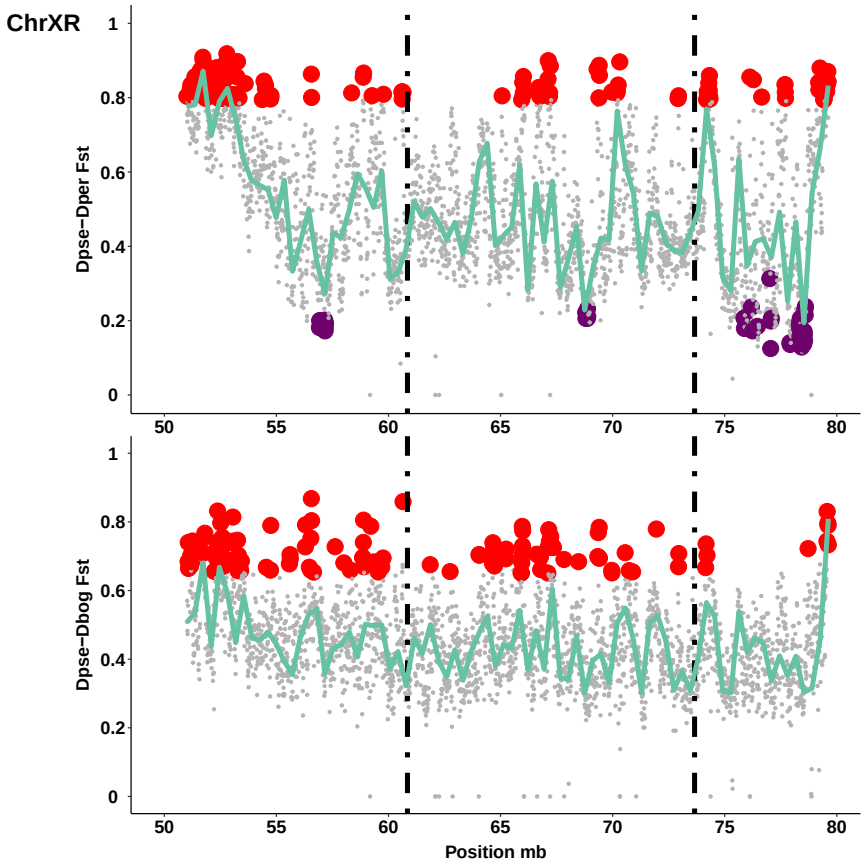


Figure S13. Fst landscape for chromosome XR for two species comparisons: *D. pseudoobscura* vs *D. persimilis* (Dpse-Dper; top), and *D. pseudoobscura* vs *D. p. bogotana* (Dpse-Dbog; bottom). Each dot represents a 10 kb non-overlapping window. Red (divergent) and purple (introgressed) dots show outlier regions based on kNN analyses.

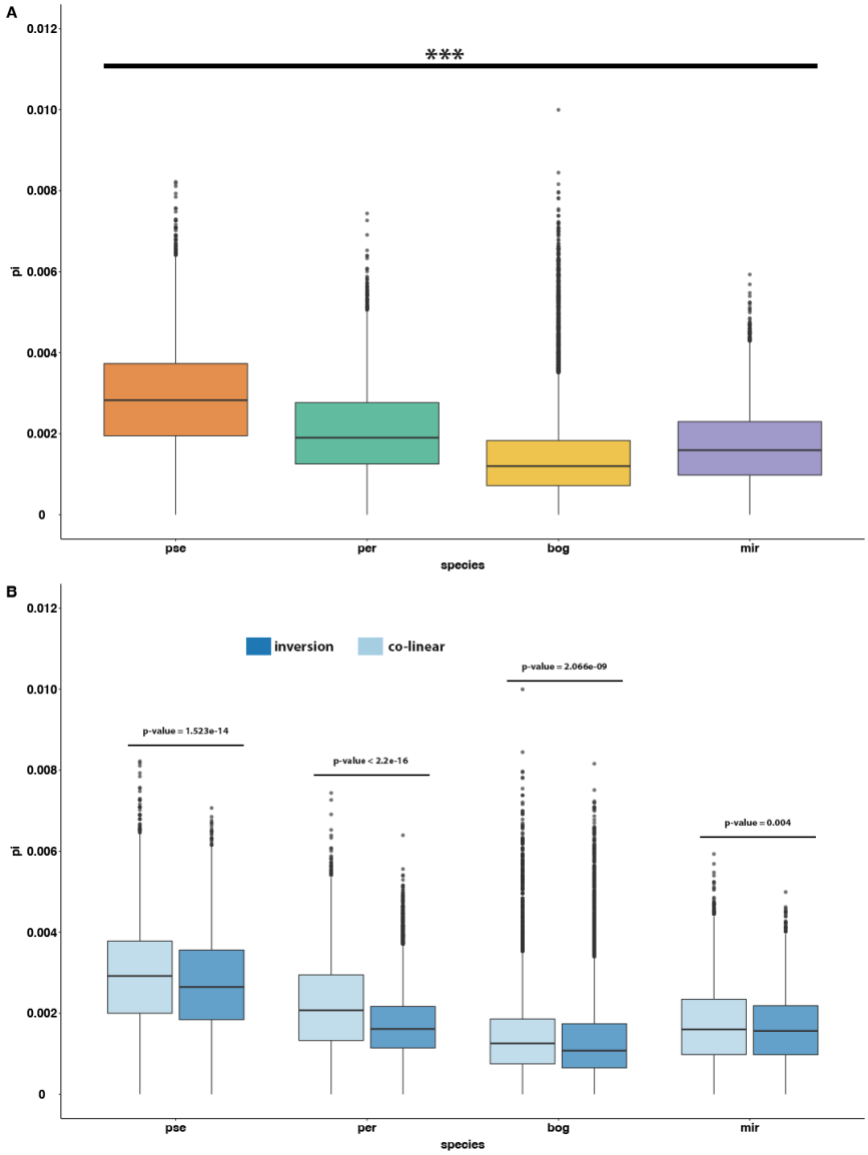


Figure S14. Nucleotide diversity (pi) distribution for the species of the pseudoobscura subgroup: pse-*D. pseudoobscura*; per-*D. persimilis*; bog-*D. p. bogotana*; mir-*D. miranda*. Genome-wide distribution of nucleotide diversity for each species (A) and nucleotide diversity within-species comparisons between co-linear and inversion differences occurring between *D. pseudoobscura* and *D. persimilis* (B). ***Wilcox test: p-value < 2.2e-16.

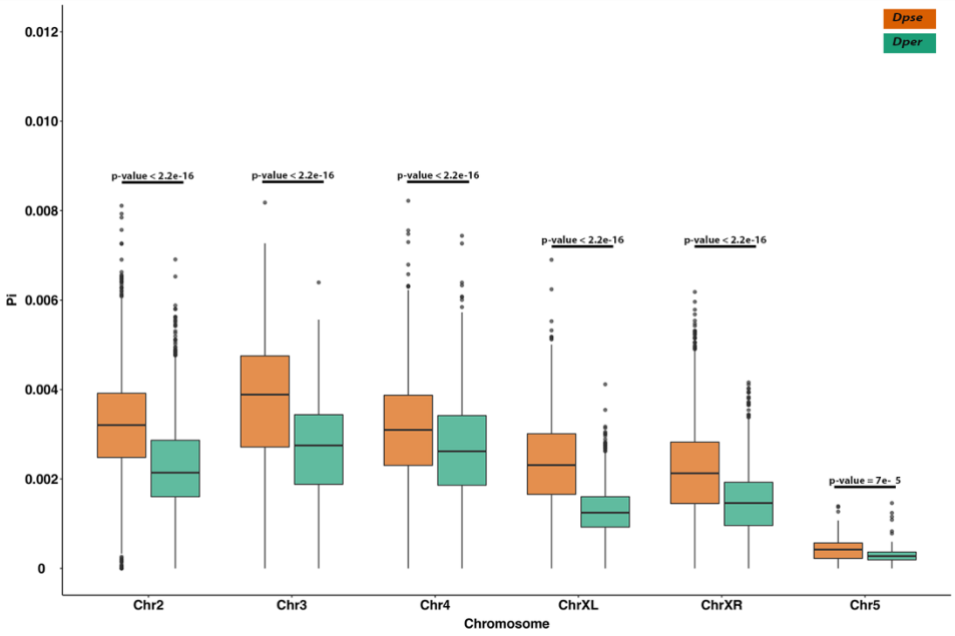


Figure S15. Nucleotide diversity (Pi) distribution across chromosomes between *D. pseudoobscura* (Dpse) and *D. persimilis* (Dper).

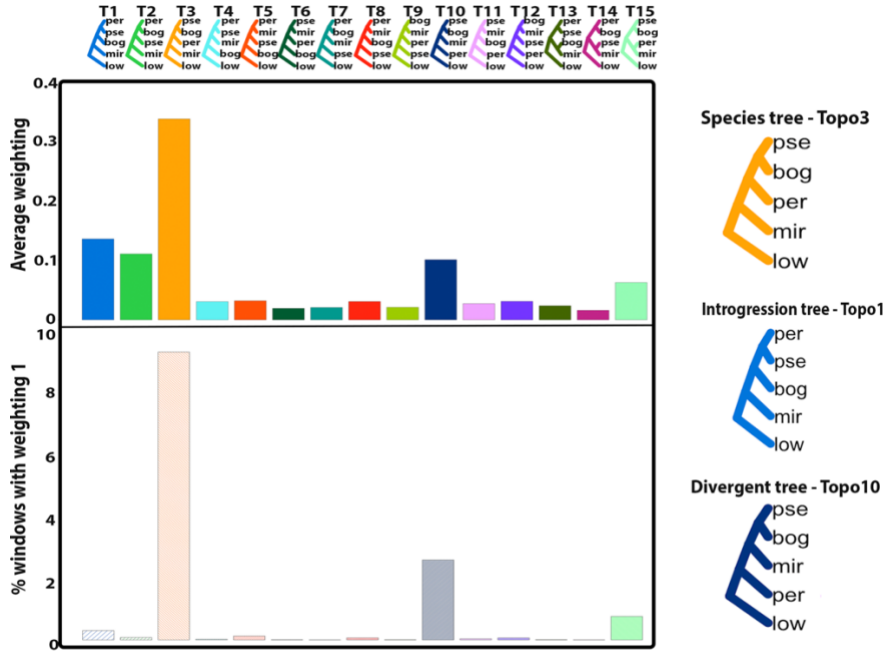


Figure S16. Genome-wide topology weighting analysis for the pseudoobscura group. Dendrograms show all possible topologies for four species. Average weighting for each of the topologies is shown in the top barplot and the percentage of windows (50 SNPs) with weighting = 1 us shown at the bottom, which indicates that all populations for each of the species are

clustered as a monophyletic group. Three of the most represented topologies are shown in the right.

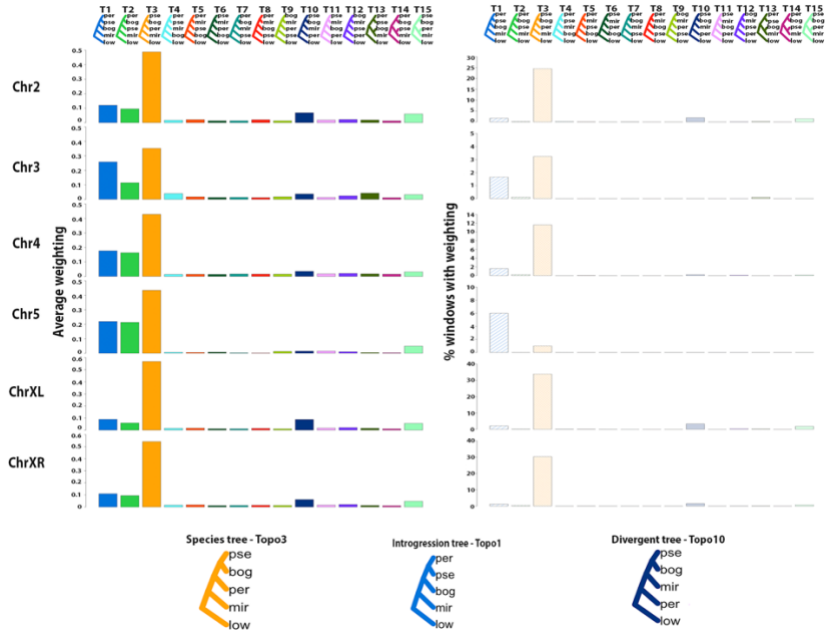


Figure S17. Topology weighting analysis for genes only of the pseudoobscura group by chromosome. Dendrograms show all possible topologies for four species. Average weighting for each of the topologies is shown in the left barplots and the percentage of genes with weighting = 1 us shown at the right, which indicates that all populations for each of the species are clustered as a monophyletic group. Three of the most represented topologies are shown at the bottom.

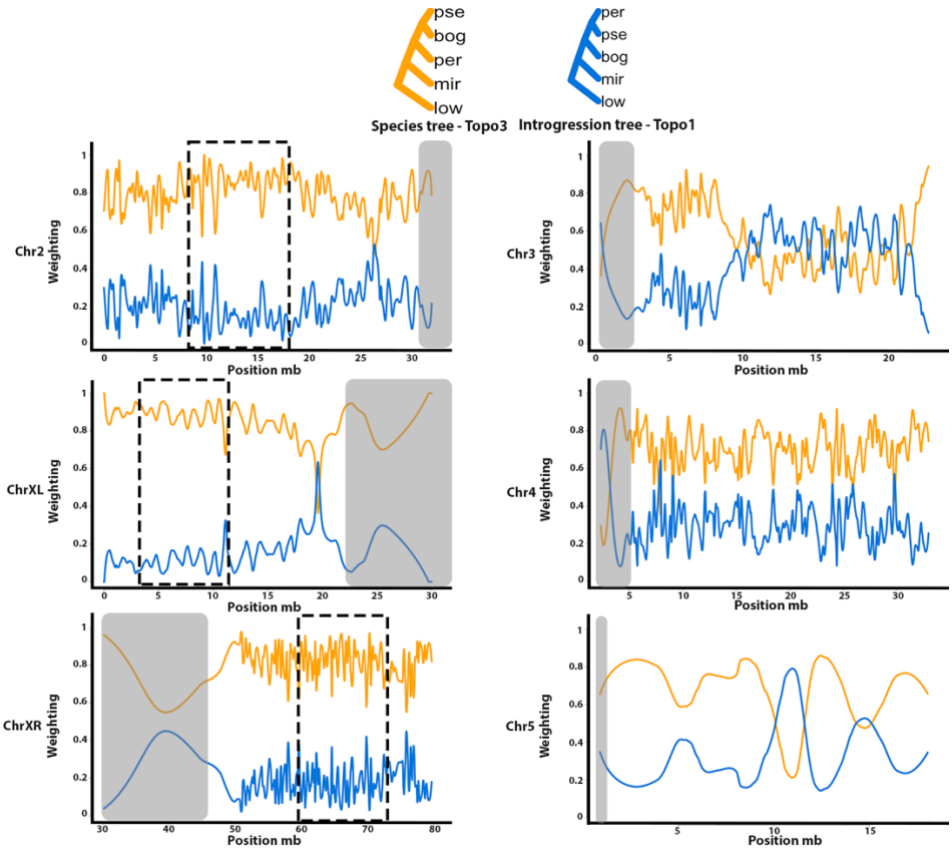


Figure S18. Gene topology weighting landscape of the two most represented topologies for all chromosomes. Colors on the line plots correspond to the topologies shown at the top of the figure. Dashed squared highlight the regions with inversion differences between *D. pseudoobscura* and *D. persimilis*. Grey areas highlight centromeric regions.

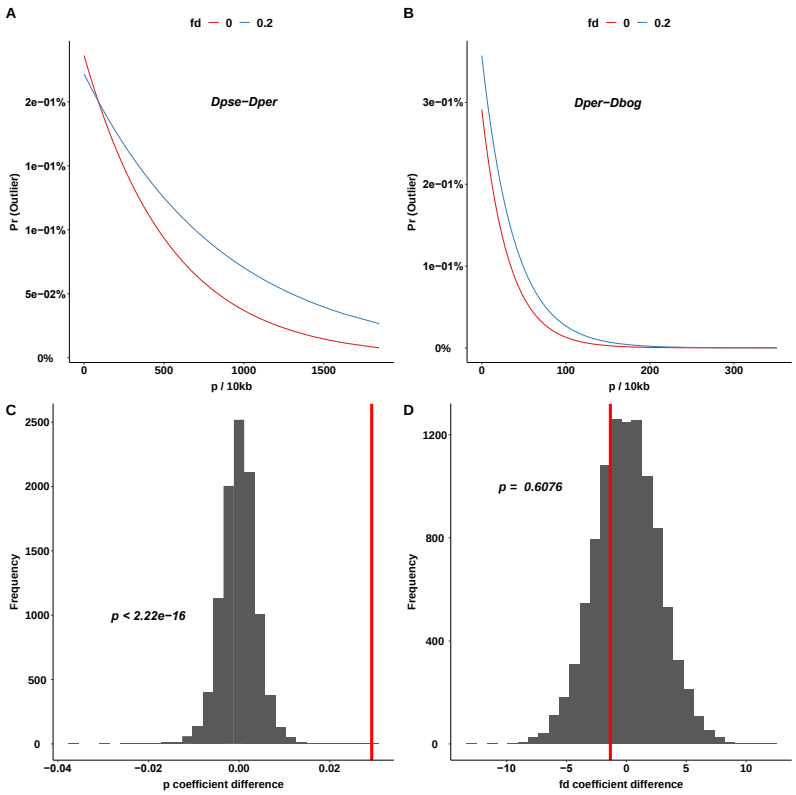


Figure S19. The tendency of Dxy-kNN outliers with evidence of selection to fall in regions of low recombination and admixture. Line plots show the probability of being an outlier as a function of the population recombination rate *p* at two *fd* levels: *fd = 0* and *fd* = 0.2. A shows the comparison between *D. pseudoobscura* and *D. persimilis* and B shows the null comparison between *D. persimilis* and *D. p. bogotana*. Barplots show the null distribution of *p* (C) and *fd* (D,) coefficient difference values between the two species comparisons with 10,000 permutations. Two-sided p-values were calculated by calculating the fraction of null distribution values greater than the observed difference value and multiplying by two.

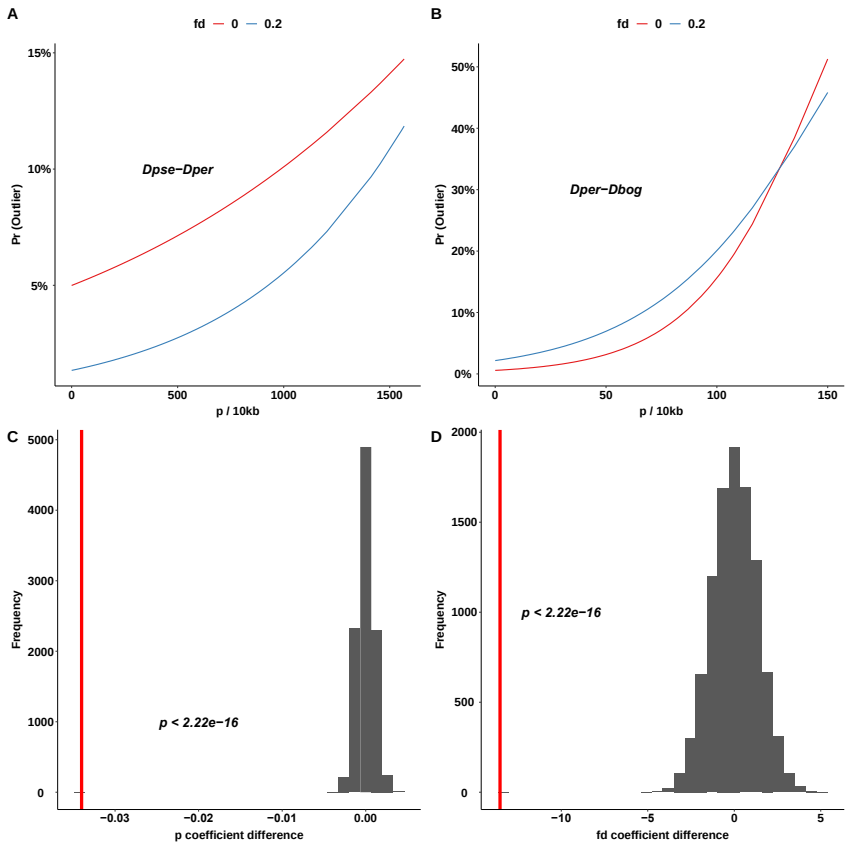


Figure S20. The tendency of Dxy (top %5) outliers with evidence of selection to fall in regions of low recombination and admixture in chromosome 2. Line plots show the probability of being an outlier as a function of the population recombination rate *p* at two *fd* levels: *fd = 0* and *fd* = 0.2. A shows the comparison between *D. pseudoobscura* and *D. persimilis* and B shows the null comparison between *D. persimilis* and *D. p. bogotana*. Barplots show the null distribution of *p* (C) and *fd* (D,) coefficient difference values between the two species comparisons with 10,000 permutations. Two-sided p-values were calculated by calculating the fraction of null distribution values greater than the observed difference value and multiplying by two.

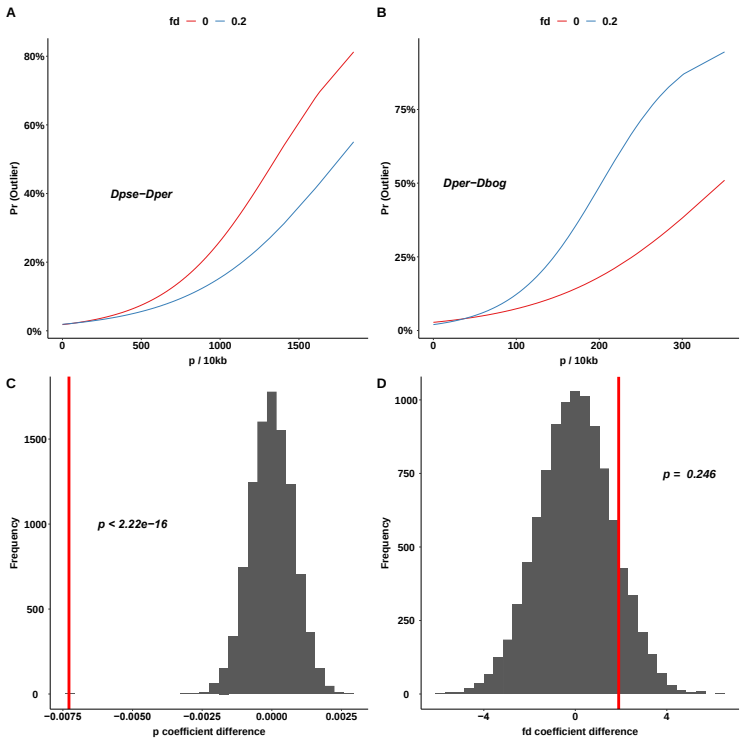


Figure S21. The tendency of Dxy (top %5) outliers with evidence of selection to fall in regions of low recombination and admixture in chromosome 4. Line plots show the probability of being an outlier as a function of the population recombination rate *p* at two *fd* levels: *fd = 0* and *fd* = 0.2. A shows the comparison between *D. pseudoobscura* and *D. persimilis* and B shows the null comparison between *D. persimilis* and *D. p. bogotana*. Barplots show the null distribution of *p* (C) and *fd* (D,) coefficient difference values between the two species comparisons with 10,000 permutations. Two-sided p-values were calculated by calculating the fraction of null distribution values greater than the observed difference value and multiplying by two.

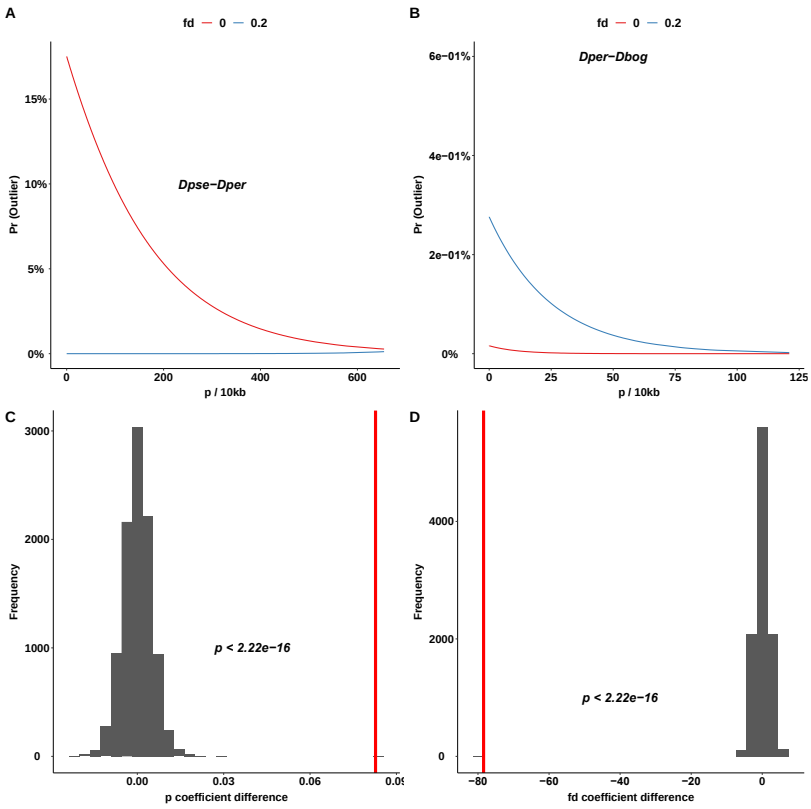


Figure S22. The tendency of Dxy (top %5) outliers with evidence of selection to fall in regions of low recombination and admixture in chromosome XL. Line plots show the probability of being an outlier as a function of the population recombination rate *p* at two *fd* levels: *fd = 0* and *fd* = 0.2. A shows the comparison between *D. pseudoobscura* and *D. persimilis* and B shows the null comparison between *D. persimilis* and *D. p. bogotana*. Barplots show the null distribution of *p* (C) and *fd* (D,) coefficient difference values between the two species comparisons with 10,000 permutations. Two-sided p-values were calculated by calculating the fraction of null distribution values greater than the observed difference value and multiplying by two.

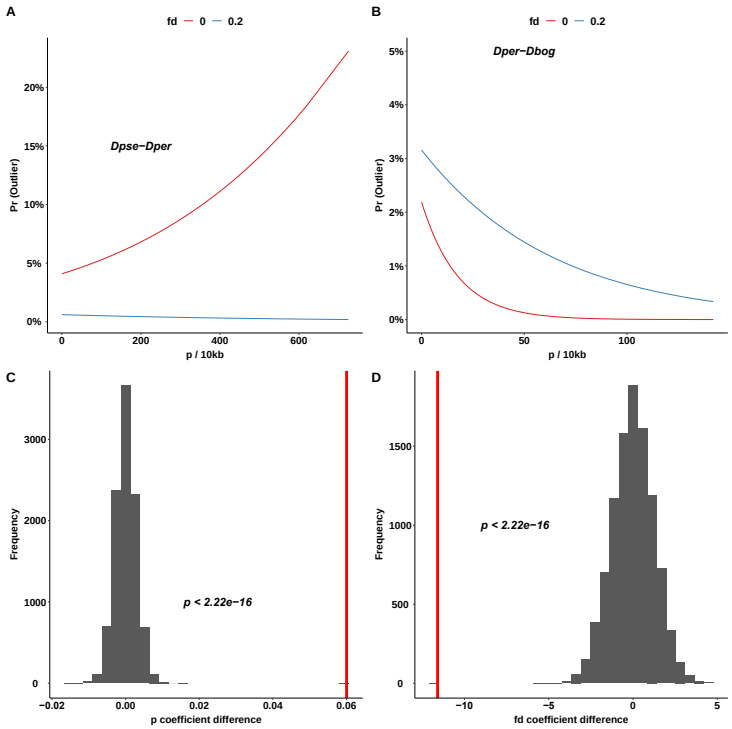


Figure S23. The tendency of Dxy (top %5) outliers with evidence of selection to fall in regions of low recombination and admixture in chromosome XR. Line plots show the probability of being an outlier as a function of the population recombination rate *p* at two *fd* levels: *fd = 0* and *fd* = 0.2. A shows the comparison between *D. pseudoobscura* and *D. persimilis* and B shows the null comparison between *D. persimilis* and *D. p. bogotana*. Barplots show the null distribution of *p* (C) and *fd* (D,) coefficient difference values between the two species comparisons with 10,000 permutations. Two-sided p-values were calculated by calculating the fraction of null distribution values greater than the observed difference value and multiplying by two.

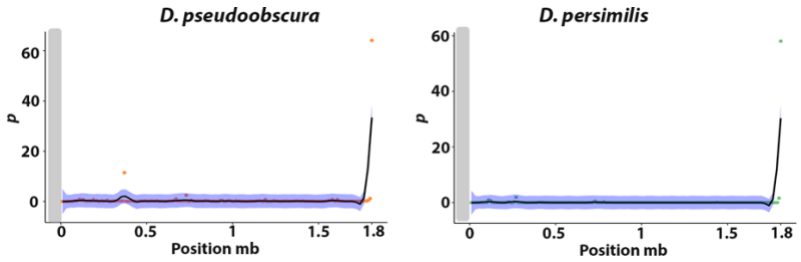


Figure S24. Population recombination rate (*p*) variation across the fifth chromosome of *D. pseudoobscura* and *D. persimilis*. Each dot represents a 10 kb sliding windows, and solid lines represent the locally weighted average (loess span = 0.8). Centromeres are denoted by grey bars.

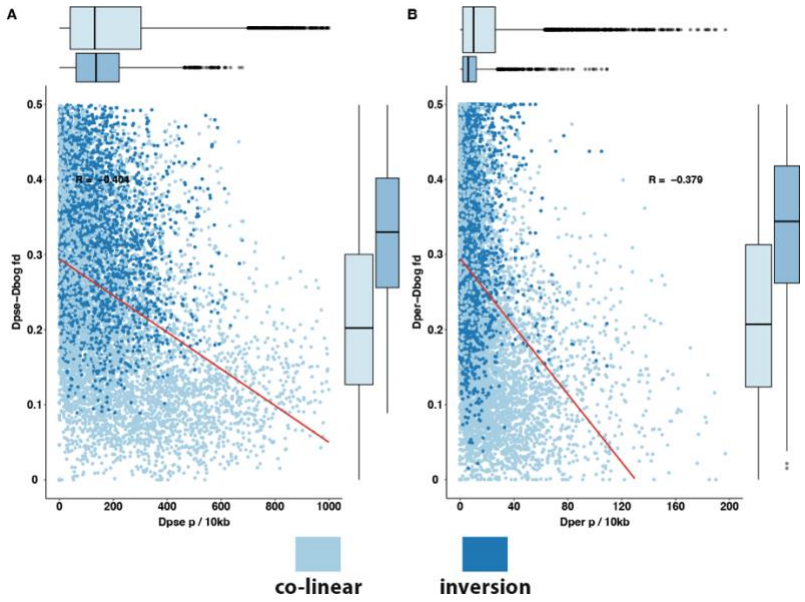


Figure S25. The correlation of recombination rate and admixture proportion. A shows the correlation of *D. pseudoobscura* recombination rates estimates and the *fd* calculations between *D. pseudoobscura* and *D. p. bogotana*. B shows the correlation of *D. persimilis* recombination rates estimates and the *fd* calculations between *D. persimilis* and *D. p. bogotana*.

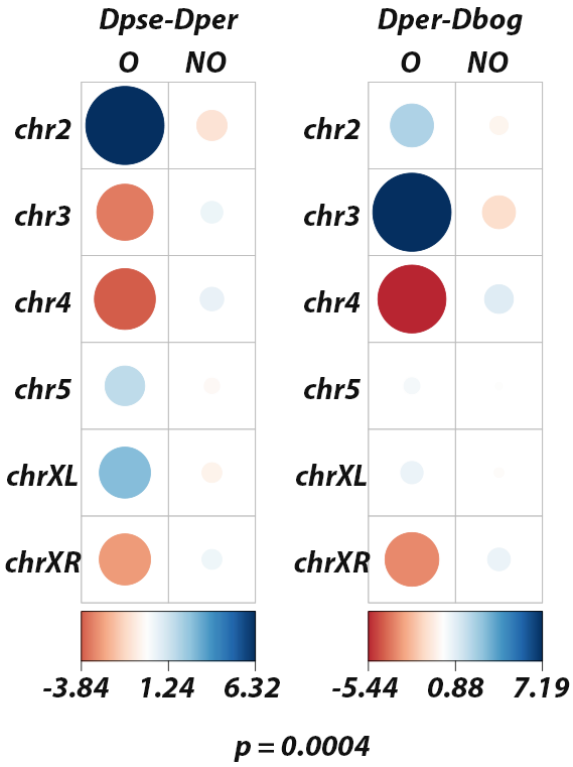


Figure S26. Test of independence comparing the association of kNN outliers to chromosomes for two pairwise comparisons. Circle size depicts the contribution of each row/column association and color depicts the correlation between columns and rows.

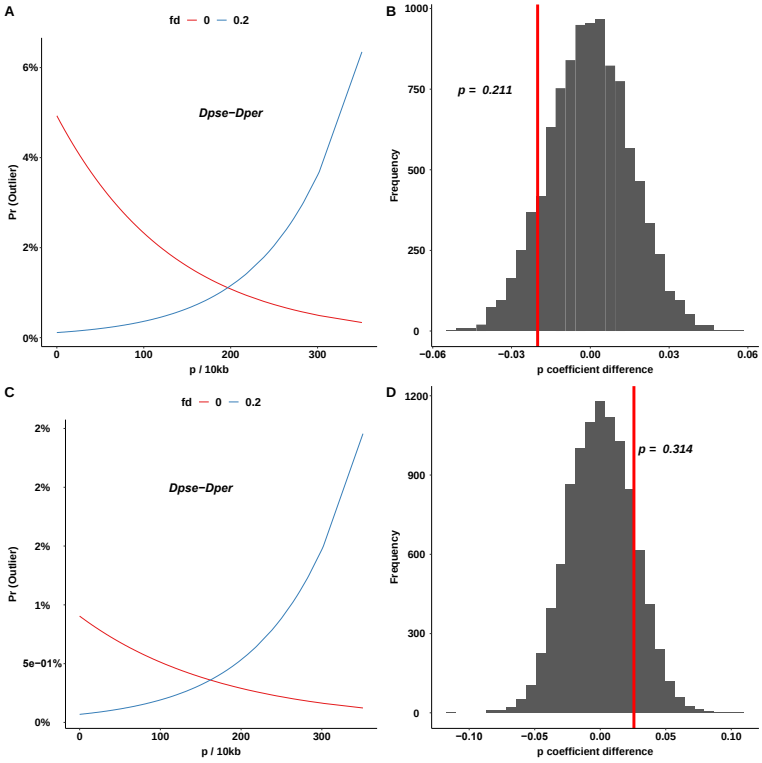


Figure S27. The tendency of Dxy outliers (top %5) (A,B) with evidence of selection (C,D) to fall in regions of low recombination and admixture using *D. persimilis* recombination rates. Line plots (A,C) show the probability of being an outlier as a function of the population recombination rate *p* at two *fd* levels: *fd = 0* and *fd* = 0.2 for *D. pseudoobscura* and *D. persimilis* comparisons. Barplots (B,D) show the null distribution of *p*, coefficient difference values between the two species comparisons with 10,000 permutations. Two-sided p-values were calculated by calculating the fraction of null distribution values greater than the observed difference value and multiplying by two.

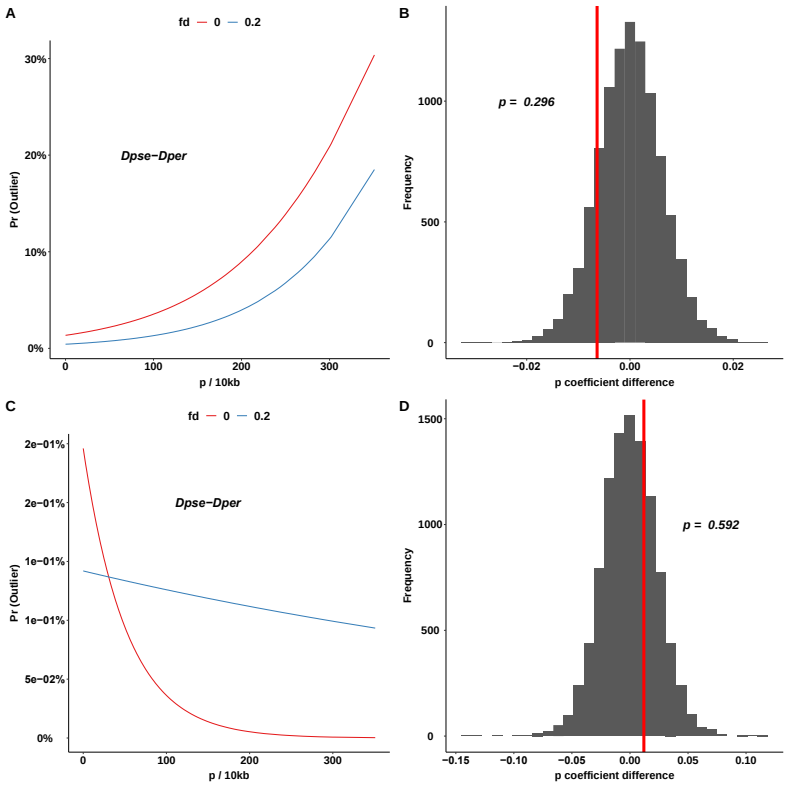


Figure S28. The tendency of Dxy kNN outliers (top %5) (A,B) with evidence of selection (C,D) to fall in regions of low recombination and admixture using *D. persimilis* recombination rates. Line plots (A,C) show the probability of being an outlier as a function of the population recombination rate *p* at two *fd* levels: *fd = 0* and *fd* = 0.2 for *D. pseudoobscura* and *D. persimilis* comparisons. Barplots (B,D) show the null distribution of *p*, coefficient difference values between the two species comparisons with 10,000 permutations. Two-sided p-values were calculated by calculating the fraction of null distribution values greater than the observed difference value and multiplying by two.

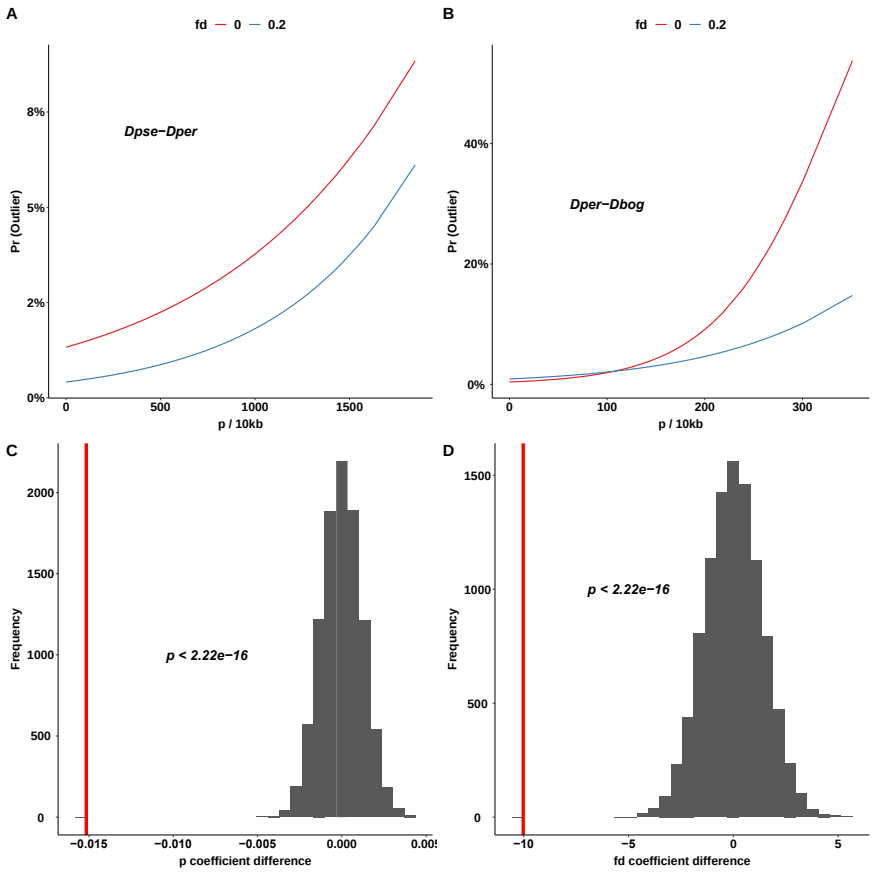


Figure S29. The tendency of Dxy kNN outliers (top %5) to fall in regions of low recombination and admixture using *D. persimilis* recombination rates. Line plots (A,B) show the probability of being an outlier as a function of the population recombination rate *p* at two *fd* levels: *fd = 0* and *fd* = 0.2 for *D. pseudoobscura* and *D. persimilis* comparisons. Barplots (C,D) show the null distribution of *p*, coefficient difference values between the two species comparisons with 10,000 permutations. Two-sided p-values were calculated by calculating the fraction of null distribution values greater than the observed difference value and multiplying by two.

Supplementary Tables

Table S1. *D. pseudoobscura* inversion breakpoints coordinates and size of previously described inversion differences between *D. pseudoobscura* and *D. persimilis*.

Chromosome	***Start***	***End***	***Size***
Chr2	9,775,358	17,555,677	7,780,319
Chr3	12,447,575	18,595,376	6,147,801

ChrXL	3,726,680	11,142,385	7,415,705
ChrXR	52,143,828	65,162,122	13,018,294

Table S2. *D. persimilis* inversion breakpoints coordinates and size of previously described inversion differences between *D. pseudoobscura* and *D. persimilis*.

Chromosome	***Start***	***End***	***Size***
Chr2	9,587,227	17,434,661	7,847,434
Chr3	11,894,872	17,877,305	5,982,433
ChrXL	3,674,883	11,520,718	7,845,835
ChrXR	60,847,215	73,658,943	12,811,728

Table S3. *D. pseudoobscura* inversion breakpoints coordinates and size of new inversions between *D. pseudoobscura* and *D. persimilis*.

Chromosome	***Start***	***End***	***Size***
Chr2	27,377,506	27,387,965	10,459
Chr2	31,917,619	31,932,450	14,831
Chr3	1,563,515	1,580,601	17,086
Chr3	1,987,648	2,188,797	201,149
Chr3	4,757,321	4,769,591	12,270
Chr4	189,980	205,698	15,718
Chr4	1,797,591	1,801,364	3,773
Chr4	1,900,914	2,532,779	631,865
Chr4	3,588,564	3,608,379	19,815
Chr4	29,588,798	29,599,456	10,658
ChrXL	19,888,081	19,893,079	4,998
ChrXL	22,578,820	22,584,585	5,765
ChrXL	23,720,401	23,775,392	54,991
ChrXL	24,382,797	24,556,531	173,734
ChrXL	24,711,162	24,740,125	28,963
ChrXL	34,552,102	35,276,191	724,089
ChrXL	35,424,154	35,508,312	84,158
ChrXL	35,866,095	35,873,521	7,426
ChrXL	38,159,752	38,171,683	11,931
ChrXL	38,835,035	38,855,615	20,580
ChrXR	44,534,745	44,553,718	18,973

Table S4. *D. persimilis* inversion breakpoints coordinates and size of new inversions between *D. pseudoobscura* and *D. persimilis*.

Chromosome	*Start*	*End*	*Size*
Chr2	27,272,507	27,282,980	10,473
Chr2	31,982,092	32,009,655	27,563
Chr3	282,801	292,680	9,879
Chr3	806,792	1,007,452	200,660
Chr3	3,757,439	3,763,192	5,753
Chr4	1,120,193	1,167,004	46,811
Chr4	2,149,654	2,162,039	12,385
Chr4	2,274,474	2,950,151	675,677
Chr4	4,118,057	4,128,094	10,037
Chr4	29,085,816	29,098,736	12,920
ChrXL	21,279,947	21,300,941	20,994
ChrXL	24,177,780	24,198,637	20,857
ChrXL	25,800,699	25,827,243	26,544
ChrXL	26,822,969	26,972,781	149,812
ChrXL	27,343,445	27,388,482	45,037
ChrXL	36,760,556	38,787,747	2,027,191
ChrXL	38,916,288	39,084,194	167,906
ChrXL	41,025,571	41,208,742	183,171
ChrXL	46,322,301	46,334,982	12,681
ChrXL	47,373,301	47,394,562	21,261
ChrXR	53,620,685	53,631,692	11,007

Table S5. ABBA-BABA test results per chromosome.

chromosome	*Sp 1*	*Sp 2.*	*Sp 3*	*D-statistic*	*Z-score*	*p-value*	*ABBA*	*BABA*
Chr2	bog	pse	per	0.0453	8.7209	2.70E-18	5600.1	5114.17
Chr3	per	pse	bog	0.516	23.8559	0	12525.4	3998.53
Chr4	bog	pse	per	0.0485	8.989	2.10E-19	6048.57	5488.58
Chr5	bog	pse	per	0.4174	3.8036	0.0001	64.566	26.5368
ChrXL	bog	pse	per	0.0354	3.0598	0.0022	1517.03	1413.04
ChrXR	bog	pse	per	0.053	5.9361	2.90E-09	3346.76	3009.78

Table S6. Estimated regression parameters for the generalized linear model of kNN Dxy outliers between *D. pseudoobscura* and *D. persimilis*. Predictors of interest: p, fd and the interaction of both predictors. GD: Gene density; GC: GC content. Numbers in bold highlight coefficients of interest and significant p-values.

Dpse-Dper	*Estimate*	*Std. Error*	*z value*	**P-*value***
Intercept	-0.877	1.213	-0.723	0.4698
p	**0.001**	0.0005	1.834	**0.0666**
fd	**-5.8467**	2.7211	-2.149	**0.0316**
GD	0.1462	0.0698	2.094	0.0363
GC	-8.1252	2.7728	-2.93	0.0033
p:fd	**0.002**	0.0051	0.403	0.687

Table S7. Estimated regression parameters for the generalized linear model of kNN Dxy outliers between *D. persimilis* and *D. p. bogotana*. Predictors of interest: p, fd and the interaction of both predictors. GD: Gene density; GC: GC content. Numbers in bold highlight coefficients of interest and significant p-values.

Dper-Dbog	*Estimate*	*Std. Error*	*z value*	**P-*value***
Intercept	-1.7155	1.1453	-1.498	0.1341
p	**0.0162**	0.0042	3.779	**0.0001**
fd	**4.1772**	0.8515	4.905	**9.30E-07**
GD	0.1462	0.0651	2.244	0.0248
GC	-9.0201	2.6838	-3.361	0.0007
p:fd	**-0.039**	0.0288	-1.353	0.1761

Table S8. Estimated regression parameters for the generalized linear model of Dxy top %5 outliers with selection between *D. pseudoobscura* and *D. persimilis*. Predictors of interest: p, fd and the interaction of both predictors. GD: Gene density; GC: GC content. Numbers in bold highlight coefficients of interest and significant p-values.

Dpse-Dper	*Estimate*	*Std. Error*	*z value*	**P-*value***
Intercept	-3.2235	1.9133	-1.685	0.092
p	**-0.0013**	0.0011	-1.186	0.2354
fd	**-12.9822**	5.0963	-2.547	**0.0109**
GD	-0.048	0.1203	-0.399	0.6897
GC	-2.931	4.3098	-0.68	0.4965
p:fd	**0.0131**	0.0088	1.477	0.1396

Table S9. Estimated regression parameters for the generalized linear model of Dxy top %5 outliers with selection between *D. persimilis* and *D. p. bogotana*. Predictors of interest: p, fd and the interaction of both predictors. GD: Gene density; GC: GC content. Numbers in bold highlight coefficients of interest and significant p-values.

Dper-Dbog	*Estimate*	*Std. Error*	*z value*	**P-*value***
Intercept	-10.7049	2.8555	-3.749	0.0001
p	**-0.0314**	0.0464	-0.676	0.4987

fd	**8.9979**	2.6023	3.458	**0.0005**
GD	0.2404	0.1167	2.059	0.0394
GC	3.3268	6.1121	0.544	0.5862
p:fd	**0.1379**	0.1256	1.098	0.2721

Table S10. Estimated regression parameters for the generalized linear model of kNN Dxy outliers with selection between *D. pseudoobscura* and *D. persimilis*. Predictors of interest: p, fd and the interaction of both predictors. GD: Gene density; GC: GC content. Numbers in bold highlight coefficients of interest and significant p-values.

Dpse-Dper	***Estimate***	***Std. Error***	***z value***	**P-*value***
Intercept	-1.5071	2.9717	-0.507	0.612
p	**-0.0018**	0.0022	-0.838	0.402
fd	**-0.3137**	5.7791	-0.054	0.957
GD	-0.0401	0.2064	-0.194	0.846
GC	-10.02	6.8754	-1.457	0.145
p:fd	**0.0035**	0.0157	0.224	0.823

Table S11. Estimated regression parameters for the generalized linear model of kNN Dxy outliers with selection between *D. persimilis* and *D. p. bogotana*. Predictors of interest: p, fd and the interaction of both predictors. GD: Gene density; GC: GC content. Numbers in bold highlight coefficients of interest and significant p-values.

Dper-Dbog	***Estimate***	***Std. Error***	***z value***	**P-*value***
Intercept	-3.9407	2.4861	-1.585	0.113
p	-0.0311	0.0359	-0.868	0.3855
fd	1.0209	2.034	0.502	0.6157
GD	0.2787	0.1165	2.392	0.0167
GC	-5.1102	5.7364	-0.891	0.373
p:fd	0.026	0.1591	0.164	0.8701

Table S12. Estimated regression parameters for the generalized linear model of Dxy top %5 outliers between *D. pseudoobscura* and *D. persimilis*, using *D. persimilis* p estimates. Predictors of interest: p, fd and the interaction of both predictors. GD: Gene density; GC: GC content. Numbers in bold highlight coefficients of interest and significant p-values.

Dpse-Dper	***Estimate***	***Std. Error***	***z value***	**P-*value***
Intercept	-3.98927	0.9184	-4.344	1.40E-05
p	**-0.0077**	0.0045	-1.697	**0.0896**
fd	**-19.1075**	2.6868	-7.111	**1.15E-12**
GD	0.1102	0.0473	2.329	0.0198
GC	1.9561	2.0433	0.957	0.3384

p:fd	**0.0971**	0.026	3.736	**0.0001**

Table S13. Estimated regression parameters for the generalized linear model of Dxy top %5 outliers between *D. persimilis* and *D. p. bogotana*, using *D. persimilis* p estimates. Predictors of interest: p, fd and the interaction of both predictors. GD: Gene density; GC: GC content. Numbers in bold highlight coefficients of interest and significant p-values.

Dper-Dbog	***Estimate***	***Std. Error***	***z value***	**P-*value***
Intercept	-7.5719	1.2589	-6.015	1.80E-09
p	**0.0122**	0.0074	1.638	0.101
fd	**12.8261**	1.1418	11.233	**< 2e-16**
GD	0.0816	0.0623	1.31	0.19
GC	-2.5717	2.7507	-0.935	0.35
p:fd	**0.0155**	0.0306	0.508	0.612

Table S14. Estimated regression parameters for the generalized linear model of Dxy top %5 outliers with selection between *D. pseudoobscura* and *D. persimilis*, using *D. persimilis* p estimates. Predictors of interest: p, fd and the interaction of both predictors. GD: Gene density; GC: GC content. Numbers in bold highlight coefficients of interest and significant p-values.

Dpse-Dper	***Estimate***	***Std. Error***	***z value***	**P-*value***
Intercept	-3.3843	1.9152	-1.767	0.0772
p	**-0.0057**	0.0087	-0.656	0.512
fd	**-12.929**	4.5407	-2.847	**0.0044**
GD	-0.0344	0.1198	-0.287	0.774
GC	-2.8259	4.3365	-0.652	0.5146
p:fd	**0.0798**	0.0416	1.919	**0.0549**

Table S15. Estimated regression parameters for the generalized linear model of Dxy top %5 outliers with selection between *D. persimilis* and *D. p. bogotana*, using *D. persimilis* p estimates. Predictors of interest: p, fd and the interaction of both predictors. GD: Gene density; GC: GC content. Numbers in bold highlight coefficients of interest and significant p-values.

Dper-Dbog	***Estimate***	***Std. Error***	***z value***	**P-*value***
Intercept	-10.704	2.8555	-3.749	0.0001
p	**-0.0314**	0.0464	-0.676	0.4987
fd	**8.9979**	2.6023	3.458	**0.0005**
GD	0.2404	0.1167	2.059	0.039
GC	3.3268	6.1121	0.544	0.5862
p:fd	**0.1379**	0.1256	1.098	0.272

Table S16. Estimated regression parameters for the generalized linear model of kNN Dxy outliers between *D. pseudoobscura* and *D. persimilis*, using *D. persimilis* p estimates. Predictors of interest: p, fd and the interaction of both predictors. GD: Gene density; GC: GC content. Numbers in bold highlight coefficients of interest and significant p-values.

Dpse-Dper	***Estimate***	***Std. Error***	***z value***	**P-*value***
Intercept	-0.6042	1.1976	-0.505	0.6139
p	**0.0098**	0.0038	2.569	**0.0101**
fd	**-5.7673**	2.258	-2.554	**0.0106**
GD	0.1481	0.0694	2.135	0.0327
GC	-8.7076	2.7718	-3.141	0.0016
p:fd	**0.0071**	0.0265	0.269	0.788

Table S17. Estimated regression parameters for the generalized linear model of kNN Dxy outliers between *D. persimilis* and *D. p. bogotana*, using *D. persimilis* p estimates. Predictors of interest: p, fd and the interaction of both predictors. GD: Gene density; GC: GC content. Numbers in bold highlight coefficients of interest and significant p-values.

Dper-Dbog	***Estimate***	***Std. Error***	***z value***	**P-*value***
Intercept	-1.7155	1.1453	-1.498	0.1341
p	**0.0162**	0.0042	3.779	**0.0001**
fd	**4.1772**	0.8515	4.905	**9.32E-07**
GD	0.1462	0.0651	2.244	0.0248
GC	-9.0201	2.6838	-3.361	0.0007
p:fd	**-0.039**	0.0288	-1.353	0.1761

Table S18. Estimated regression parameters for the generalized linear model of kNN Dxy outliers with selection between *D. pseudoobscura* and *D. persimilis*, using *D. persimilis* p estimates. Predictors of interest: p, fd and the interaction of both predictors. GD: Gene density; GC: GC content. Numbers in bold highlight coefficients of interest and significant p-values.

Dpse-Dper	***Estimate***	***Std. Error***	***z value***	**P-*value***
Intercept	-1.3509	2.9884	-0.452	0.651
p	**-0.0191**	0.0189	-1.011	0.312
fd	**-2.7555**	4.5515	-0.605	0.545
GD	-0.027	0.205	-0.132	0.895
GC	-10.3195	6.9509	-1.485	0.138
p:fd	**0.0899**	0.061	1.473	0.141

Table S19. Estimated regression parameters for the generalized linear model of kNN Dxy outliers with selection between *D. persimilis* and *D. p. bogotana*, using *D. persimilis* p estimates Predictors of interest: p, fd and the interaction of both predictors. GD: Gene density; GC: GC content. Numbers in bold highlight coefficients of interest and significant p-values.

Dpse-Dbog	*Estimate*	*Std. Error*	*z value*	**P-*value***
Intercept	-3.9407	2.4861	-1.585	0.113
p	-0.0311	0.0359	-0.868	0.3855
fd	1.0209	2.034	0.502	**0.6157**
GD	0.2787	0.1165	2.392	0.0167
GC	-5.1102	5.7364	-0.891	0.373
p:fd	0.026	0.1591	0.164	**0.8701**

Appendix C: Supplementary Materials for Chapter3

Supplementary Figures

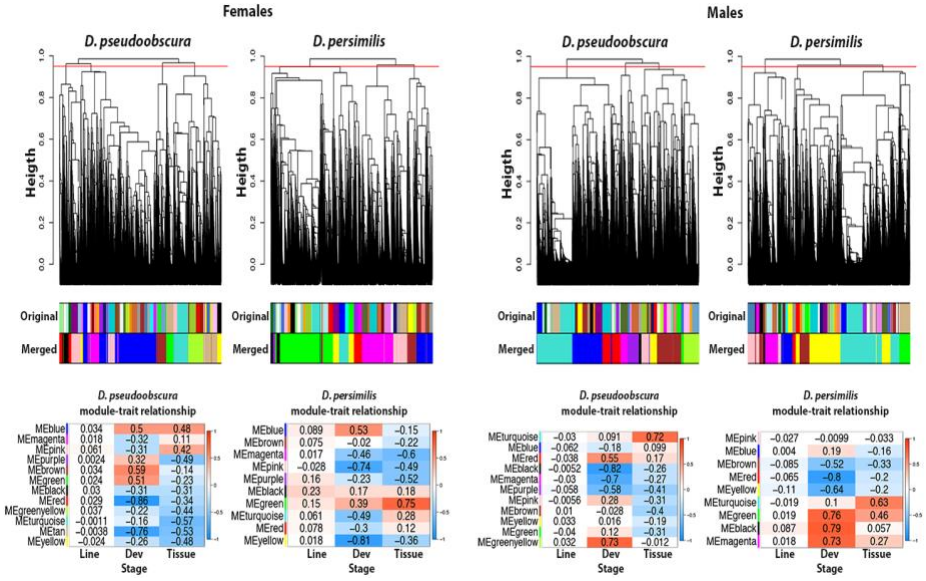


Figure S1. **Module assignments and module-trait relationships for each sex for each species.** Dendrograms and color bars show the module assignments before (original) and after the module merging step (merged) for each network. Bottom heatmaps show the module eigengene – condition relationships for each of the merged modules for each species. Dark red depicts strong positive correlation whereas dark blue depicts strong negative correlation. Line, Dev Stage and Tissue encodes the metadata associated with each RNAseq sample used for the network construction.

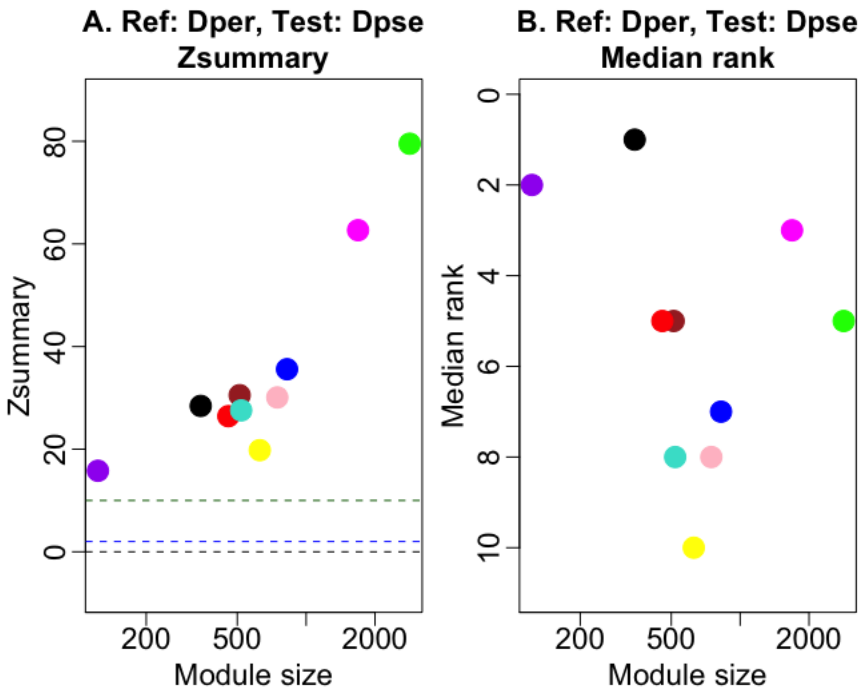


Figure S2. **Module preservation analysis for female networks using *D. persimilis* as a reference.** Module preservation Zsummary and Median rank summary statistics for females (top) and males (bottom) networks. Each colored dot represents a single module. Zsummary horizontal lines indicate preservation thresholds: black (Zsummary = 0); blue (Zsummary = 2); green (Zsummary =10).

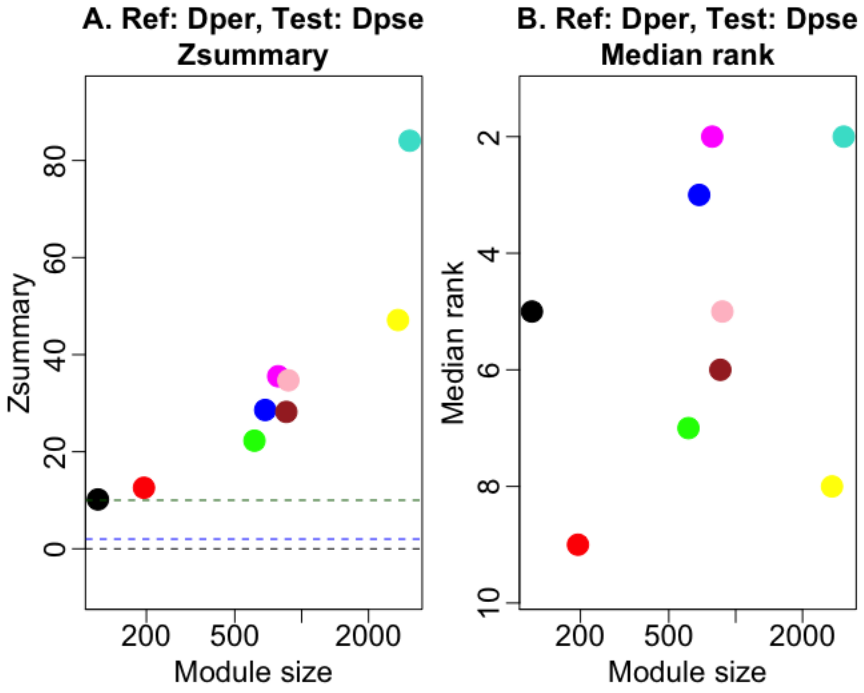


Figure S3. Module preservation analysis for male networks using *D. persimilis* as a reference. Module preservation Zsummary and Median rank summary statistics for females (top) and males (bottom) networks. Each colored dot represents a single module. Zsummary horizontal lines indicate preservation thresholds: black (Zsummary = 0); blue (Zsummary = 2); green (Zsummary =10)

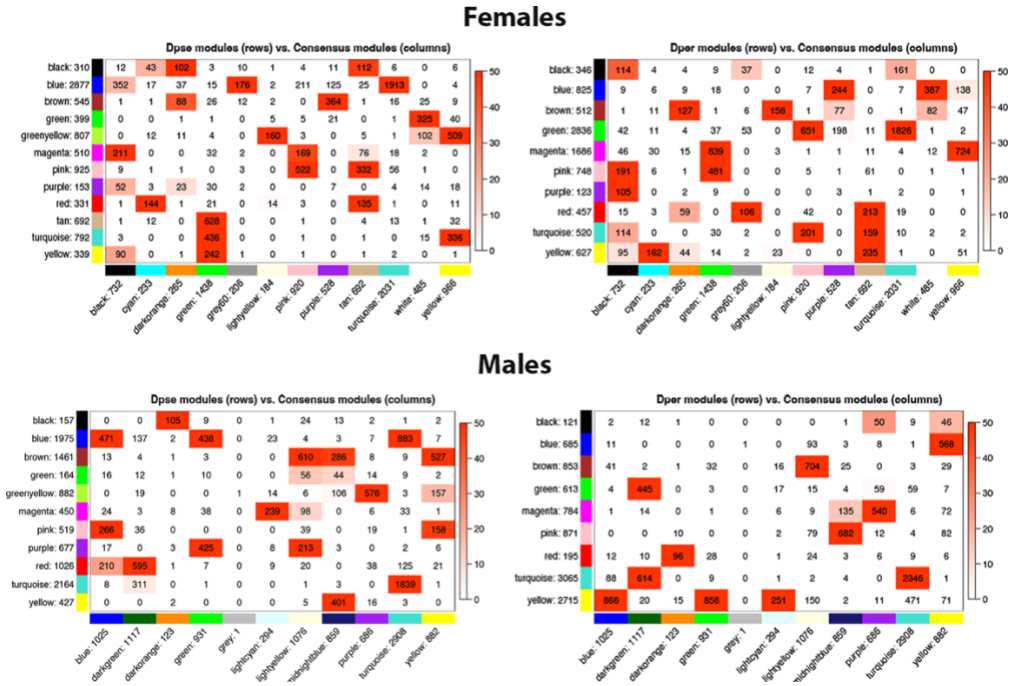


Figure S4. Module correspondence for both females (top) and males (bottom) between the consensus modules and *D. pseudoobscura* and *D. persimilis*. x-axis show *D. persimilis* modules and y-axis modules for *D. pseudoobscura*. Heatmaps represent the degree of overlap between modules across species. Color represent the -log(p-value): white (not significant; $p = 1$); dark red (extremely significant; $p < 0.0001$); Fisher exact test.

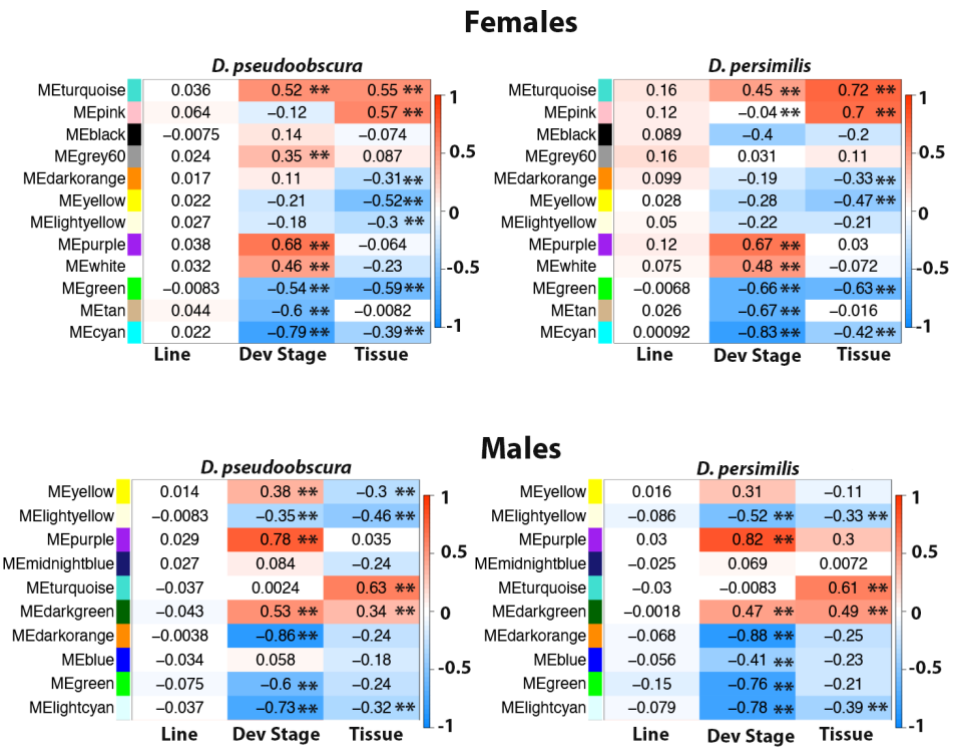


Figure S5. Module-trait relationships across consensus module eigengene for both males and females across species. Line, Dev Stage and Tissue encodes the metadata associated with each RNAseq sample used for the network construction. Numbers inside the boxes and colors indicate the strength of spearman correlation for each module and condition; (**) represent significant association.

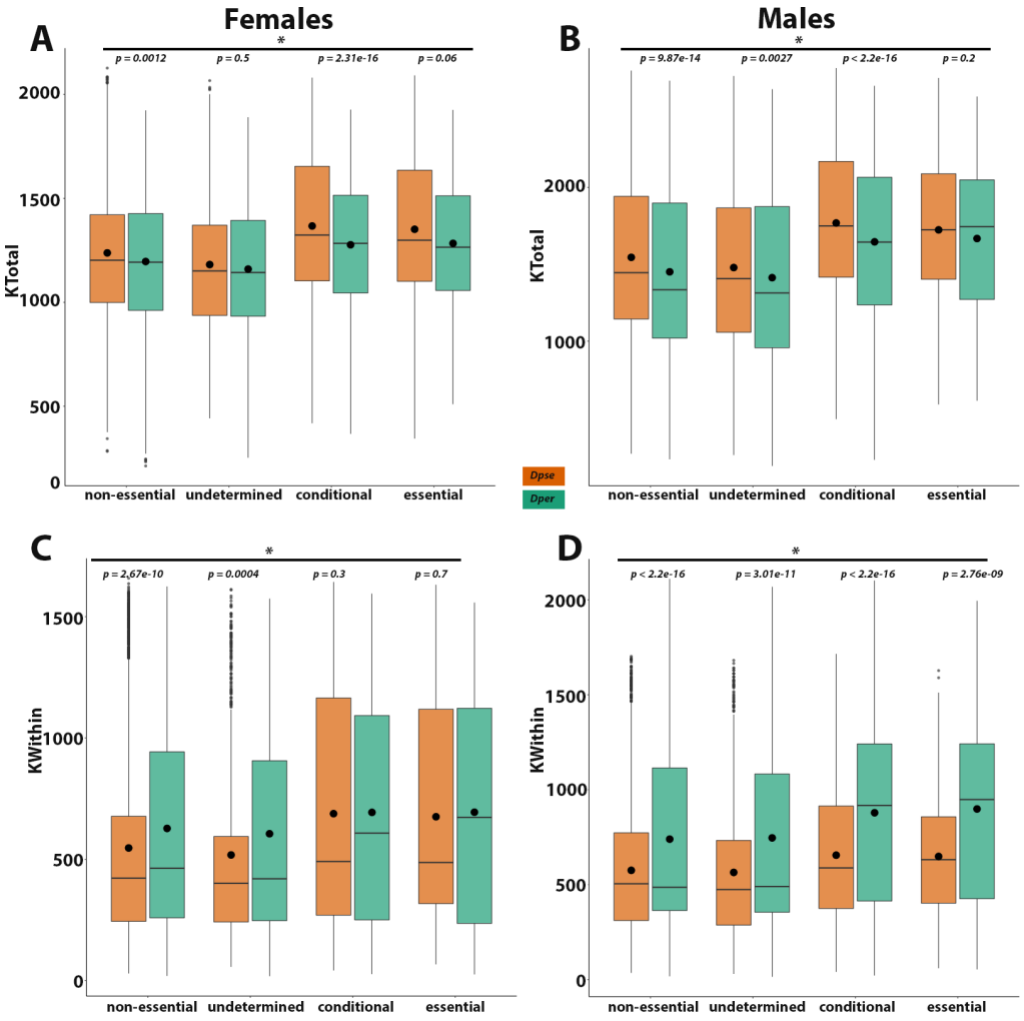


Figure S6. Gene connectivity comparisons across gene essentiality classes based on *D. melanogaster*. Panels A and B show KTotal, which is a connectivity measure based on the whole network connectivity. Panels C and D show KWithin, which is a intramodular connectivity measure. P-values (Wilcoxon rank sum test) are shown for each species comparison across edge classes. Top horizontal line with asterisk indicates significant differences across edge classes for each species ($p < 0.001$; Kruskal-Wallis rank sum test).

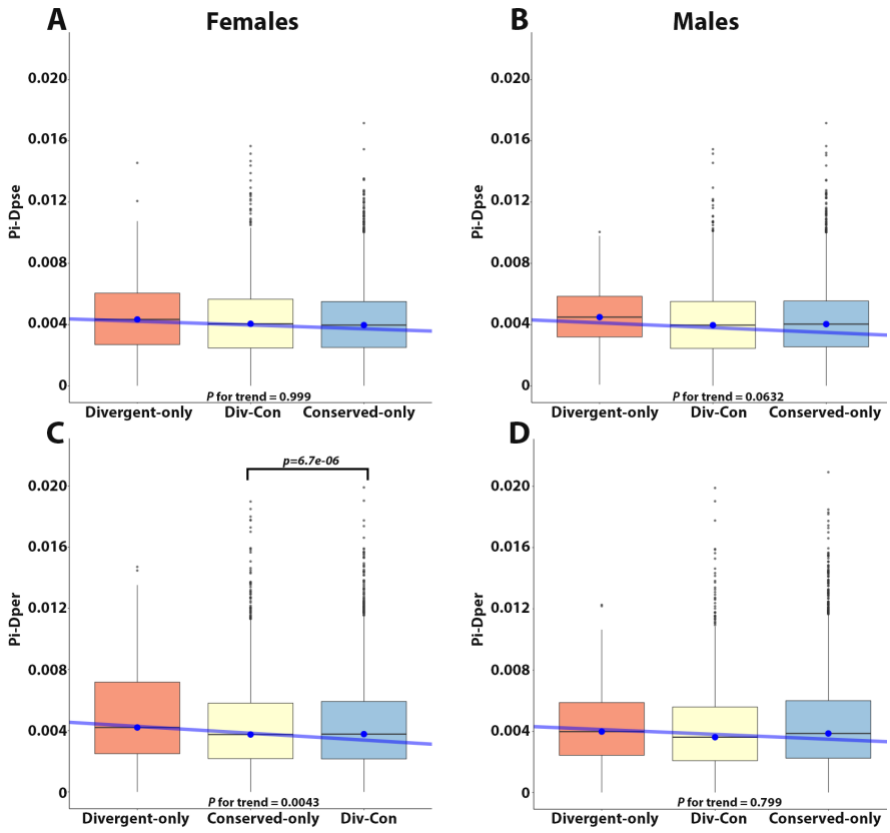


Figure S7. Nucleotide diversity (Pi) across genes grouped by edge classes. Panels A (Kruskal-Wllis rank sum test; p = 0.0842) and B (Kruskal-Wllis rank sum test; p = 0.1417) show *D. pseudoobscura* Pi differences across edge classes for males and females. Panels C (Kruskal-Wllis rank sum test; p = 0.0032) and D (Kruskal-Wllis rank sum test; p = 0.2824) show *D. persimilis* Pi differences across edge classes for males and females. Blue curves represent the linear regression of Dxy and gene classification. *P*=Wilcoxon rank-sum test for trend.

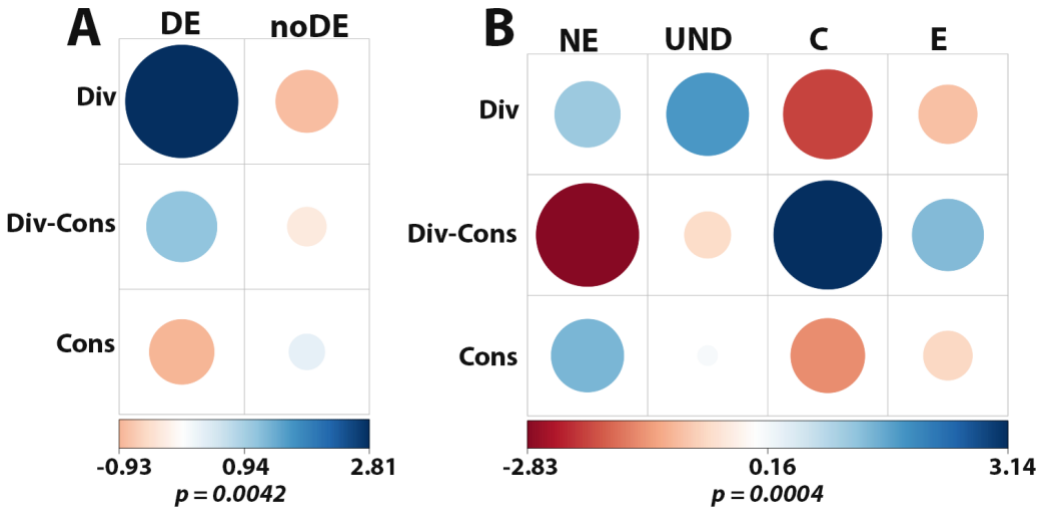


Figure S8. Test of independence comparing associations across differential expression status and essentiality status in females. Association between differential expression and edge classes is show in A. B shows association between gene essentiality and edge classes. Color of the circles indicates either positive association (blue) or negative association (red) between corresponding rows and column variables. The size of the circle is proportional to the amount of cell contribution. Chi-squared p-values are shown at the bottom of each contingency table. DE: differential expression; noDE: no evidence of differential expression; NE: non-essential genes; UND: not classified genes; C: conditional essentiality; E: essential. Div: genes with only divergent edges; Cons: genes with only conserved egdes; Div-Con: genes with both divergent and conserved edges.

Supplementary Tables

Table S1. Number of consensus clusters for each constructed network.

	D. pseudoobscura	*D. persimilis*
Male	38	35
Female	39	36
Male merged	11	9
Female merge	12	10

Table S2. Lines utilized for each species to generate developmental and tissue-specific RNA-seq data. Each column lists the species lines.

D. pseudoobscura	*D. persimilis*
Flagstaff16	111.35
Mather32	111.41

MSH9	111.51
MV-225	Mather40
TL	MSH42
MSH24	--

Table S3. Female module preservation statistics using *D. pseudoobscura* as reference.

Module	Zsummary	Median Rank
purple	11	11
black	14	11
yellow	18	9
red	19	7
magenta	25	7
brown	27	8
green	31	4
pink	36	8
greenyellow	46	4
tan	50	3
turquoise	52	1
gold	56	13
blue	64	10

Table S4. Male module preservation statistics using *D. pseudoobscura* as reference.

Module	Zsummary	Median Rank
black	15	4
blue	53	6
brown	35	8
gold	52	12
green	8.6	10
greenyellow	37	5
magenta	27	6
pink	18	9
purple	35	2
red	22	10
turquoise	75	1
yellow	26	3

Table S5. List of species-specific hub genes with evidence of selection and differential expression between *D. pseudoobscura* and *D. persimilis*. INV: inversion; COL: co-linear; permodule: *D. persimilis* module name; psemodule: *D. pseudoobscura* module name.

Gene ID	Gene name	Species hub	Chromosome	Location	permodule	psemodule
perhifi04038	*TfAP-2*	dpse	chrX_RagTag	INV	magenta	tan
perhifi06456	*beat-Iia*	dpse	chr2_RagTag	COL	magenta	turquoise
perhifi07370	*kat-60L1*	dpse	chr2_RagTag	COL	red	pink
perhifi07964	*CG11964*	dpse	chr2_RagTag	COL	turquoise	magenta
perhifi10293	*chico*	dpse	chr4_RagTag	COL	green	pink
perhifi12384	*Rab4*	dpse	chr3_RagTag	COL	turquoise	pink
perhifi14310	*CG8249*	dpse	chr3_RagTag	COL	magenta	greenyellow
perhifi14326	*fus*	dpse	chr3_RagTag	COL	magenta	purple
perhifi14330	*CG15117*	dpse	chr3_RagTag	COL	brown	brown
perhifi02592	*CG3918*	dper	chrX_RagTag	COL	green	blue
perhifi05670	*NA*	dper	chr2_RagTag	COL	brown	greenyellow
perhifi08892	*swm*	dper	chr4_RagTag	COL	black	blue
perhifi11821	*Cirl*	dper	chr3_RagTag	COL	pink	tan
perhifi02720	*CTPsyn*	dpse	chrX_RagTag	COL	green	blue
perhifi04109	*Aef1*	dpse	chrX_RagTag	INV	green	blue
perhifi04307	*CG5577*	dpse	chrX_RagTag	INV	blue	brown
perhifi08917	*Arpc1*	dpse	chr4_RagTag	COL	yellow	red
perhifi11779	*TBPH*	dpse	chr3_RagTag	COL	pink	magenta
perhifi02596	*dx*	dper	chrX_RagTag	COL	green	blue
perhifi03090	*Ccn*	dper	chrX_RagTag	COL	magenta	turquoise
perhifi05717	*CG9356*	dper	chr2_RagTag	COL	green	blue
perhifi10012	*SA1*	dper	chr4_RagTag	COL	green	blue
perhifi11450	*chinmo*	dper	chr4_RagTag	COL	yellow	tan
perhifi13148	*Cul1*	dper	chr3_RagTag	INV	green	pink
perhifi14326	*fus*	dper	chr3_RagTag	COL	magenta	purple
perhifi03584	*GAPsec*	dpse	chrX_RagTag	INV	turquoise	magenta

BIBLIOGRAPHY

Aeschbacher, S., Selby, J. P., Willis, J. H., & Coop, G. (2017). Population-genomic inference of the strength and timing of selection against gene flow. *Proc Natl Acad Sci U S A, 114*(27), 7061-7066. doi:10.1073/pnas.1616755114

Alachiotis, N., & Pavlidis, P. (2018). RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. *Commun Biol, 1*, 79. doi:10.1038/s42003-018-0085-8

Allen, S. L., Delaney, E. K., Kopp, A., & Chenoweth, S. F. (2017). Single-Molecule Sequencing of the Drosophila serrata Genome. *G3-Genes Genomes Genetics, 7*(3), 781-788. doi:10.1534/g3.116.037598

Alonge, M., Lebeigle, L., Kirsche, M., Aganezov, S., Wang, X., Lippman, Z. B., . . . Soyk, S. (2021). Automated assembly scaffolding elevates a new tomato system for high-throughput genome editing. *bioRxiv*.

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., . . . Lippman, Z. B. (2020). Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell, 182*(1), 145-161 e123. doi:10.1016/j.cell.2020.05.021

Alvarez-Ponce, D., Feyertag, F., & Chakraborty, S. (2017). Position Matters: Network Centrality Considerably Impacts Rates of Protein Evolution in the Human Protein-Protein Interaction Network. *Genome Biol Evol, 9*(6), 1742-1756. doi:10.1093/gbe/evx117

Anderson, W. W., Ayala, F. J., & Michod, R. E. (1977). Chromosomal and allozymic diagnosis of three species of Drosophila. Drosophila pseudoobscura, D. persimilis, and D. miranda. *J Hered, 68*(2), 71-74. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/874309

Aoki, K., Ogata, Y., & Shibata, D. (2007). Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant and Cell Physiology, 48*(3), 381-390. doi:10.1093/pcp/pcm013

Aviles-Pagan, E. E., Kang, A. S. W., & Orr-Weaver, T. L. (2020). Identification of New Regulators of the Oocyte-to-Embryo Transition in Drosophila. *G3-Genes Genomes Genetics, 10*(9), 2989-2998. doi:10.1534/g3.120.401415

Barbash, D. A., Roote, J., & Ashburner, M. (2000). The Drosophila melanogaster hybrid male rescue gene causes inviability in male and female species hybrids. *Genetics, 154*(4), 1747-1771. Retrieved from <Go to ISI>://WOS:000086491000027

Barbash, D. A., Siino, D. F., Tarone, A. M., & Roote, J. (2003). A rapidly evolving MYB-related protein causes species isolation in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America, 100*(9), 5302-5307. doi:10.1073/pnas.0836927100

Barrett, R. D. H., & Hoekstra, H. E. (2011). Molecular spandrels: tests of adaptation at the genetic level. *Nature Reviews Genetics, 12*(11), 767-780. doi:10.1038/nrg3015

Barton, N., & Bengtsson, B. O. (1986). The barrier to genetic exchange between hybridising populations. *Heredity, 57*, 357-376.

Bertranpetit, J., Pybus, M., Luisi, P., Alvarez-Ponce, D., Laayouni, H., & Fares, M. A. (2015). Recent Positive Selection Has Acted on Genes Encoding Proteins with More Interactions within the Whole Human Interactome. *Genome Biology and Evolution, 7*(4), 1141-1154. doi:10.1093/gbe/evv055

Booker, T. R., Yeaman, S., & Whitlock, M. C. (2020). Variation in recombination rate affects detection of outliers in genome scans under neutrality. *Mol Ecol, 29*(22), 4274-4279. doi:10.1111/mec.15501

Bracewell, R., Chatla, K., Nalley, M. J., & Bachtrog, D. (2019). Dynamic turnover of centromeres drives karyotype evolution in Drosophila. *Elife, 8*. doi:10.7554/eLife.49002

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

Burri, R., Nater, A., Kawakami, T., Mugal, C. F., Olason, P. I., Smeds, L., . . . Ellegren, H. (2015). Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. *Genome Res, 25*(11), 1656-1665. doi:10.1101/gr.196485.115

Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., . . . Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res, 18*(1), 188-196. doi:10.1101/gr.6743907

Chakraborty, M., Baldwin-Brown, J. G., Long, A. D., & Emerson, J. J. (2016). Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res, 44*(19), e147. doi:10.1093/nar/gkw654

Chakraborty, M., Emerson, J. J., Macdonald, S. J., & Long, A. D. (2019). Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nature Communications, 10*. doi:ARTN 4872

10.1038/s41467-019-12884-1

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

Chakraborty, S., & Alvarez-Ponce, D. (2016). Positive Selection and Centrality in the Yeast and Fly Protein-Protein Interaction Networks. *Biomed Res Int, 2016*, 4658506. doi:10.1155/2016/4658506

Chan, A. H., Jenkins, P. A., & Song, Y. S. (2012). Genome-wide fine-scale recombination rate variation in Drosophila melanogaster. *Plos Genetics, 8*(12), e1003090. doi:10.1371/journal.pgen.1003090

Charlesworth, B. (1998). Measures of divergence between populations and the effect of forces that reduce variability. *Mol. Biol. Evol., 15*(5), 538-543.

Charlesworth, B., & Barton, N. H. (2018). The Spread of an Inversion with Migration and Selection. *Genetics, 208*(1), 377-382. doi:DOI 10.1534/genetics.117.300426

Charlesworth, B., Morgan, M. T., & Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics, 134*(4), 1289-1303. doi:10.1093/genetics/134.4.1289

Chase, M. A., Ellegren, H., & Mugal, C. F. (2021). Positive selection plays a major role in shaping signatures of differentiation across the genomic landscape of two independent Ficedula flycatcher species pairs. *Evolution, 75*(9), 2179-2196. doi:10.1111/evo.14234

Chen, H., Patterson, N., & Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Research, 20*(3), 393-402. doi:10.1101/gr.100545.109

Chen, W. H., Lu, G. T., Chen, X., Zhao, X. M., & Bork, P. (2017). OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in

human cancer cell lines. *Nucleic Acids Research, 45*(D1), D940-D944. doi:10.1093/nar/gkw1013
Chen, W. H., Minguez, P., Lercher, M. J., & Bork, P. (2012). OGEE: an online gene essentiality database. *Nucleic Acids Research, 40*(D1), D901-D906. doi:10.1093/nar/gkr986
Chen, W. Y., Luan, X. J., Yan, Y. D., Wang, M., Zheng, Q. W., Chen, X., . . . Fang, J. (2020). CG8005 Mediates Transit-Amplifying Spermatogonial Divisions via Oxidative Stress in Drosophila Testes. *Oxidative Medicine and Cellular Longevity, 2020*. doi:Artn 2846727 10.1155/2020/2846727
Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods, 18*(2), 170-175. doi:10.1038/s41592-020-01056-5
Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., Kim, Y., . . . Consortium, G. (2017). The impact of structural variation on human gene expression. *Nature Genetics, 49*(5), 692-+. doi:10.1038/ng.3834
Choi, J. Y., & Lee, Y. C. G. (2020). Double-edged sword: The evolutionary consequences of the epigenetic silencing of transposable elements. *Plos Genetics, 16*(7), e1008872. doi:10.1371/journal.pgen.1008872
Choi, J. Y., Purugganan, M., & Stacy, E. A. (2020). Divergent Selection and Primary Gene Flow Shape Incipient Speciation of a Riparian Tree on Hawaii Island. *Mol Biol Evol, 37*(3), 695-710. doi:10.1093/molbev/msz259
Cooper, J. C., & Phadnis, N. (2016). A genomic approach to identify hybrid incompatibility genes. *Fly, 10*(3), 142-148. doi:10.1080/19336934.2016.1193657
Costanzo, M., Kuzmin, E., van Leeuwen, J., Mair, B., Moffat, J., Boone, C., & Andrews, B. (2019). Global Genetic Networks and the Genotype-to-Phenotype Relationship. *Cell, 177*(1), 85-100. doi:10.1016/j.cell.2019.01.033
Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., . . . Pathway Analysis Working Group of the International Cancer Genome, C. (2015). Pathway and network analysis of cancer genomes. *Nat Methods, 12*(7), 615-621. doi:10.1038/nmeth.3440
Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology, 23*(13), 3133-3157. doi:10.1111/mec.12796
Cutter, A. D., & Payseur, B. A. (2013). Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews Genetics, 14*(4), 262-274. doi:10.1038/nrg3425
Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., . . . Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience, 10*(2). doi:10.1093/gigascience/giab008
Darbellay, F., & Necsulea, A. (2020). Comparative Transcriptomics Analyses across Species, Organs, and Developmental Stages Reveal Functionally Constrained lncRNAs. *Mol Biol Evol, 37*(1), 240-259. doi:10.1093/molbev/msz212
Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., & Pritchard, J. K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics, 25*(24), 3207-3212. doi:10.1093/bioinformatics/btp579
Delprat, A., Negre, B., Puig, M., & Ruiz, A. (2009). The Transposon Galileo Generates Natural Chromosomal Inversions in Drosophila by Ectopic Recombination. *Plos One, 4*(11). doi:ARTN e7883

10.1371/journal.pone.0007883
DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., . . . Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics, 43*(5), 491-+. doi:10.1038/ng.806
Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15-21. doi:10.1093/bioinformatics/bts635
Dobzhansky, T. (1944). Chromosomal races in *Drosophila pseudoobscura* and *Drosophila persimilis*. In T. Dobzhansky & C. Epling (Eds.), *Contributions to the genetics, taxonomy, and ecology of Drosophila pseudoobscura and its relatives* (Vol. 554, pp. 47-144). Washington, DC: Carnegie Institute of Washington.
Dobzhansky, T. (1973). Is there Gene Exchange between Drosophila pseudoobsura and Drosophila persimilis in Their Natural Habitats? *The American Naturalist, 107*(954), 312-314. doi:10.1086/282833
Dobzhansky, T., & Epling, T. (1944). *Contributions to the genetics, taxonomy, and ecology of Drosophila pseudoobscura and its relatives* (Vol. 554). Washington, DC: Carnegie Institute of Washington.
Dolgin, E. S., & Charlesworth, B. (2008). The effects of recombination rate on the distribution and abundance of transposable elements. *Genetics, 178*(4), 2169-2177. doi:10.1534/genetics.107.082743
Drosophila 12 Genomes, C., Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., . . . MacCallum, I. (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature, 450*(7167), 203-218. doi:10.1038/nature06341
Du, J. K., He, X. L., Zhou, Y. M., Zhai, C. C., Yu, D. E., Zhang, S. H., . . . Wan, X. C. (2021). Gene Coexpression Network Reveals Insights into the Origin and Evolution of a Theanine-Associated Regulatory Module in Non-Camellia and Camellia Species. *Journal of Agricultural and Food Chemistry, 69*(1), 615-626. doi:10.1021/acs.jafc.0c06490
Eden, E., Lipson, D., Yogev, S., & Yakhini, Z. (2007). Discovering motifs in ranked lists of DNA sequences. *Plos Computational Biology, 3*(3), 508-522. doi:ARTN e39
10.1371/journal.pcbi.0030039
Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics, 10*. doi:Artn 48
10.1186/1471-2105-10-48
Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology, 20*(1), 238. doi:10.1186/s13059-019-1832-y
Feder, J. L., Egan, S. P., & Nosil, P. (2012). The genomics of speciation-with-gene-flow. *Trends Genet, 28*(7), 342-350. doi:10.1016/j.tig.2012.03.009
Feder, J. L., Gejji, R., Powell, T. H. Q., & Nosil, P. (2011). Adaptive Chromosomal Divergence Driven by Mixed Geographic Mode of Evolution. *Evolution, 65*(8), 2157-2170. doi:10.1111/j.1558-5646.2011.01321.x
Filteau, M., Pavey, S. A., St-Cyr, J., & Bernatchez, L. (2013). Gene coexpression networks reveal key drivers of phenotypic divergence in lake whitefish. *Mol Biol Evol, 30*(6), 1384-1396. doi:10.1093/molbev/mst053

Fortin, M. J., Dale, M. R. T., & Brimacombe, C. (2021). Network ecology in dynamic landscapes. *Proceedings of the Royal Society B-Biological Sciences, 288*(1949). doi:ARTN 20201889
10.1098/rspb.2020.1889
Fuller, Z. L., Koury, S. A., Phadnis, N., & Schaeffer, S. W. (2019). How chromosomal rearrangements shape adaptation and speciation: Case studies in Drosophila pseudoobscura and its sibling species Drosophila persimilis. *Molecular Ecology, 28*(6), 1283-1301. doi:10.1111/mec.14923
Fuller, Z. L., Leonard, C. J., Young, R. E., Schaeffer, S. W., & Phadnis, N. (2018). Ancestral polymorphisms explain the role o chromosomal inversions in speciation. *Plos Genetics, 14*(7). doi:ARTN e1007526
10.1371/journal.pgen.1007526
Galiana-Arnoux, D., Dostert, C., Schneemann, A., Hoffmann, J. A., & Imler, J. L. (2006). Essential function in vivo for Dicer-2 in host defense against RNA viruses in drosophila. *Nat Immunol, 7*(6), 590-597. doi:10.1038/ni1335
Gebert, D., Neubert, L. K., Lloyd, C., Gui, J., Lehmann, R., & Teixeira, F. K. (2021). Large Drosophila germline piRNA clusters are evolutionarily labile and dispensable for transposon regulation. *Mol Cell, 81*(19), 3965-3978 e3965. doi:10.1016/j.molcel.2021.07.011
Gerdol, M., Moreira, R., Cruz, F., Gomez-Garrido, J., Vlasova, A., Rosani, U., . . . Figueras, A. (2020). Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. *Genome Biology, 21*(1), 275. doi:10.1186/s13059-020-02180-3
Goel, M., Sun, H., Jiao, W. B., & Schneeberger, K. (2019). SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology, 20*(1), 277. doi:10.1186/s13059-019-1911-0
Guindon, S., & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology, 52*(5), 696-704. doi:10.1080/10635150390235520
Guy, L., Roat Kultima, J., & Andersson, S. G. E. (2010). genoPlotR: comparative gene and genome visualization in R. *Bioinformatics, 26*(18), 2334-2335. doi:10.1093/bioinformatics/btq413
Hastings, P. J., Lupski, J. R., Rosenberg, S. M., & Ira, G. (2009). Mechanisms of change in gene copy number. *Nature Reviews Genetics, 10*(8), 551-564. doi:10.1038/nrg2593
Hill, T., & Betancourt, A. J. (2018). Extensive exchange of transposable elements in the Drosophila pseudoobscura group. *Mobile DNA, 9*. doi:ARTN 20
10.1186/s13100-018-0123-6
Hoffmann, A. A., & Rieseberg, L. H. (2008). Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation? *Annu Rev Ecol Evol Syst, 39*, 21-42. doi:10.1146/annurev.ecolsys.39.110707.173532
Hon, T., Mars, K., Young, G., Tsai, Y. C., Karalius, J. W., Landolin, J. M., . . . Rank, D. R. (2020). Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data, 7*(1). doi:ARTN 399
10.1038/s41597-020-00743-4
Hu, G., Hovav, R., Grover, C. E., Faigenboim-Doron, A., Kadmon, N., Page, J. T., . . . Wendel, J. F. (2016). Evolutionary Conservation and Divergence of Gene Coexpression Networks

in Gossypium (Cotton) Seeds. *Genome Biol Evol, 8*(12), 3765-3783. doi:10.1093/gbe/evw280
Huang, K. C., & Rieseberg, L. H. (2020). Frequency, Origins, and Evolutionary Role of Chromosomal Inversions in Plants. *Frontiers in Plant Science, 11*. doi:ARTN 296 10.3389/fpls.2020.00296
Huang, Y., Shukla, H., & Lee, Y. C. G. (2022). Species-specific chromatin landscape determines how transposable elements shape genome evolution. *Elife, 11*. doi:10.7554/eLife.81567
Hufford, M. B., Seetharam, A. S., Woodhouse, M. R., Chougule, K. M., Ou, S., Liu, J., . . . Dawe, R. K. (2021). De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science, 373*(6555), 655-662. doi:10.1126/science.abg5289
Izuno, A., Onoda, Y., Amada, G., Kobayashi, K., Mukai, M., Isagi, Y., & Shimizu, K. K. (2022). Demography and selection analysis of the incipient adaptive radiation of a Hawaiian woody species. *Plos Genetics, 18*(1), e1009987. doi:10.1371/journal.pgen.1009987
Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., . . . Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology, 36*(4), 338-+. doi:10.1038/nbt.4060
Jalili, M., Salehzadeh-Yazdi, A., Gupta, S., Wolkenhauer, O., Yaghmaie, M., Resendis-Antonio, O., & Alimoghaddam, K. (2016). Evolution of Centrality Measurements for the Detection of Essential Proteins in Biological Networks. *Frontiers in Physiology, 7*. doi:ARTN 375 10.3339/fphys.2016.00375
Jiang, P., Wang, H., Li, W., Zang, C., Li, B., Wong, Y. J., . . . Liu, X. S. (2015). Network analysis of gene essentiality in functional genomics experiments. *Genome Biology, 16*, 239. doi:10.1186/s13059-015-0808-9
Jin, Y., Tam, O. H., Paniagua, E., & Hammell, M. (2015). TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics, 31*(22), 3593-3599. doi:10.1093/bioinformatics/btv422
Johnson, K. A., & Krishnan, A. (2022). Robust normalization and transformation techniques for constructing gene coexpression networks from RNA-seq data. *Genome Biology, 23*(1), 1. doi:10.1186/s13059-021-02568-9
Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., . . . Team, W. G. A. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature, 484*(7392), 55-61. doi:10.1038/nature10944
Kim, B. Y., Wang, J. R., Miller, D. E., Barmina, O., Delaney, E., Thompson, A., . . . Petrov, D. A. (2021). Highly contiguous assemblies of 101 drosophilid genomes. *Elife, 10*. doi:10.7554/eLife.66405
Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol, 37*(8), 907-915. doi:10.1038/s41587-019-0201-4
Kirkpatrick, M., & Barton, N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics, 173*(1), 419-434. doi:10.1534/genetics.105.047985
Kofler, R., Betancourt, A. J., & Schlotterer, C. (2012). Sequencing of Pooled DNA Samples (Pool-Seq) Uncovers Complex Dynamics of Transposable Element Insertions in Drosophila melanogaster. *Plos Genetics, 8*(1). doi:ARTN e1002487 10.1371/journal.pgen.1002487
Kofler, R., Nolte, V., & Schlotterer, C. (2015). Tempo and Mode of Transposable Element Activity in Drosophila. *Plos Genetics, 11*(7). doi:ARTN e1005406

10.1371/journal.pgen.1005406
Korunes, K. L., Machado, C. A., & Noor, M. A. F. (2021). Inversions shape the divergence of Drosophila pseudoobscura and Drosophila persimilis on multiple timescales. *Evolution*. doi:10.1111/evo.14278
Krause, S. A., Overend, G., Dow, J. A. T., & Leader, D. P. (2022). FlyAtlas 2 in 2022: enhancements to the Drosophila melanogaster expression atlas. *Nucleic Acids Res, 50*(D1), D1010-D1015. doi:10.1093/nar/gkab971
Kronenberg, Z. N., Fiddes, I. T., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, O. S., . . . Eichler, E. E. (2018). High-resolution comparative analysis of great ape genomes. *Science, 360*(6393). doi:10.1126/science.aar6343
Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., . . . Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res, 19*(9), 1639-1645. doi:10.1101/gr.092759.109
Kulathinal, R. J., Bennettt, S. M., Fitzpatrick, C. L., & Noor, M. A. F. (2008). Fine-scale mapping of recombination rate in Drosophila refines its correlation to diversity and divergence. *Proceedings of the National Academy of Sciences of the United States of America, 105*(29), 10051-10056. doi:10.1073/pnas.0801848105
Kulathinal, R. J., Stevison, L. S., & Noor, M. A. F. (2009). The Genomics of Speciation in Drosophila: Diversity, Divergence, and Introgression Estimated Using Low-Coverage Genome Sequencing. *Plos Genetics, 5*(7). doi:ARTN e1000550
10.1371/journal.pgen.1000550
Kuo, R. I., Tseng, E., Eory, L., Paton, I. R., Archibald, A. L., & Burt, D. W. (2017). Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. *BMC Genomics, 18*(1), 323. doi:10.1186/s12864-017-3691-9
Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology, 5*(2), R12. doi:10.1186/gb-2004-5-2-r12
Langfelder, P., & Horvath, S. (2007). Eigengene networks for studying the relationships between co-expression modules. *Bmc Systems Biology, 1*. doi:Artn 54
10.1186/1752-0509-1-54
Langfelder, P., Luo, R., Oldham, M. C., & Horvath, S. (2011). Is My Network Module Preserved and Reproducible? *Plos Computational Biology, 7*(1). doi:ARTN e1001057
10.1371/journal.pcbi.1001057
Larracuente, A. M., & Clark, A. G. (2014). Recent selection on the Y-to-dot translocation in Drosophila pseudoobscura. *Mol Biol Evol, 31*(4), 846-856. doi:10.1093/molbev/msu002
Lau, L. Y., Reverter, A., Hudson, N. J., Naval-Sanchez, M., Fortes, M. R. S., & Alexandre, P. A. (2020). Dynamics of Gene Co-expression Networks in Time-Series Data: A Case Study in Drosophila melanogaster Embryogenesis. *Frontiers in Genetics, 11*, 517. doi:10.3389/fgene.2020.00517
Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics, 34*(18), 3094-3100. doi:10.1093/bioinformatics/bty191
Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Proc, G. P. D. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Liao, Y., Zhang, X. W., Chakraborty, M., & Emerson, J. J. (2021). Topologically associating domains and their role in the evolution of genome structure and function in Drosophila. *Genome Research, 31*(3), 397-410. doi:10.1101/gr.266130.120

Lindsay, S. L. (1958). Food Preferences of Drosophila larvae. *Am Nat, 92*(866), 279–285.

Liu, G. W., Yong, M. Y. J., Yurieva, M., Srinivasan, K. G., Liu, J., Lim, J. S. Y., . . . Rancati, G. (2015). Gene Essentiality Is a Quantitative Property Linked to Cellular Evolvability. *Cell, 163*(6), 1388-1399. doi:10.1016/j.cell.2015.10.069

Liu, J., Yin, F., Lang, K., Jie, W., Tan, S., Duan, R., . . . Huang, W. (2022). MetazExp: a database for gene expression and alternative splicing profiles and their analyses based on 53 615 public RNA-seq samples in 72 metazoan species. *Nucleic Acids Res, 50*(D1), D1046-D1054. doi:10.1093/nar/gkab933

Liu, Z., Zhao, H., Yan, Y., Wei, M. X., Zheng, Y. C., Yue, E. K., . . . Xu, J. H. (2021). Extensively Current Activity of Transposable Elements in Natural Rice Accessions Revealed by Singleton Insertions. *Frontiers in Plant Science, 12*. doi:ARTN 745526
10.3389/fpls.2021.745526

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics, 21*(10), 597-614. doi:10.1038/s41576-020-0236-x

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*(12), 550. doi:10.1186/s13059-014-0550-8

Luisi, P., Alvarez-Ponce, D., Pybus, M., Fares, M. A., Bertranpetit, J., & Laayouni, H. (2015). Recent positive selection has acted on genes encoding proteins with more interactions within the whole human interactome. *Genome Biol Evol, 7*(4), 1141-1154. doi:10.1093/gbe/evv055

Machado, C. A., Haselkorn, T. S., & Noor, M. A. (2007). 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, 175*(3), 1289-1306. doi:10.1534/genetics.106.064758

Machado, C. A., & Hey, J. (2003). The causes of phylogenetic conflict in a classic Drosophila species group. *Proc Biol Sci, 270*(1520), 1193-1202. doi:10.1098/rspb.2003.2333

Machado, C. A., Kliman, R. M., Markert, J. A., & Hey, J. (2002). Inferring the history of speciation from multilocus DNA sequence data: the case of Drosophila pseudoobscura and close relatives. *Mol Biol Evol, 19*(4), 472-488. doi:10.1093/oxfordjournals.molbev.a004103

Mahajan, S., Wei, K. H. C., Nalley, M. J., Gibilisco, L., & Bachtrog, D. (2018). De novo assembly of a young Drosophila Y chromosome using single-molecule sequencing and chromatin conformation capture. *PLOS Biology, 16*(7), e2006348. doi:10.1371/journal.pbio.2006348

Malinsky, M., Matschiner, M., & Svardal, H. (2021). Dsuite - Fast D-statistics and related admixture evidence from VCF files. *Mol Ecol Resour, 21*(2), 584-595. doi:10.1111/1755-0998.13265

Marco, A., Konikoff, C., Karr, T. L., & Kumar, S. (2009). Relationship between gene co-expression and sharing of transcription factor binding sites in Drosophila melanogaster. *Bioinformatics, 25*(19), 2473-2477. doi:10.1093/bioinformatics/btp462

Marques, D. A., Lucek, K., Meier, J. I., Mwaiko, S., Wagner, C. E., Excoffier, L., & Seehausen, O. (2016). Genomics of Rapid Incipient Speciation in Sympatric Threespine Stickleback. *Plos Genetics, 12*(2), e1005887. doi:10.1371/journal.pgen.1005887

Martin, S. H., Davey, J. W., & Jiggins, C. D. (2015). Evaluating the Use of ABBA-BABA Statistics to Locate Introgressed Loci. *Molecular Biology and Evolution, 32*(1), 244-257. doi:10.1093/molbev/msu269

Martin, S. H., Davey, J. W., Salazar, C., & Jiggins, C. D. (2019). Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLoS Biol, 17*(2), e2006288. doi:10.1371/journal.pbio.2006288

Martin, S. H., & Van Belleghem, S. M. (2017). Exploring Evolutionary Relationships Across the Genome Using Topology Weighting. *Genetics, 206*(1), 429-438. doi:10.1534/genetics.116.194720

Matsuo, Y., Nose, A., & Kohsaka, H. (2021). Interspecies variation of larval locomotion kinematics in the genus Drosophila and its relation to habitat temperature. *BMC Biol, 19*(1), 176. doi:10.1186/s12915-021-01110-4

McGaugh, S. E., & Noor, M. A. F. (2012). Genomic impacts of chromosomal inversions in parapatric Drosophila species. *Philosophical Transactions of the Royal Society B-Biological Sciences, 367*(1587), 422-429. doi:10.1098/rstb.2011.0250

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research, 20*(9), 1297-1303. doi:10.1101/gr.107524.110

Merel, V., Boulesteix, M., Fablet, M., & Vieira, C. (2020). Transposable elements in Drosophila. *Mob DNA, 11*, 23. doi:10.1186/s13100-020-00213-z

Miller, D. E., Staber, C., Zeitlinger, J., & Hawley, R. S. (2018). Highly Contiguous Genome Assemblies of 15 Drosophila Species Generated Using Nanopore Sequencing. *G3 (Bethesda), 8*(10), 3131-3141. doi:10.1534/g3.118.200160

Mohn, N., & Spiess, E. B. (1963). Cold resistance of karyotypes in Drosophila persimilis from Timberline of California. *Evolution, 17*, 548-563.

Monaco, G., van Dam, S., Ribeiro, J. L. C. N., Larbi, A., & de Magalhaes, J. P. (2015). A comparison of human and mouse gene co-expression networks reveals conservation and divergence at the tissue, pathway and disease levels. *Bmc Evolutionary Biology, 15*. doi:ARTN 259

10.1186/s12862-015-0534-7

Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning, 52*(1-2), 91-118. doi:Doi 10.1023/A:1023949509487

Moore, B. C., & Taylor, C. E. (1986). Drosophila of Southern California: III. Gene arrangements of Drosophila persimilis. *Journal of Heredity, 77*(5), 313-323. doi:10.1093/oxfordjournals.jhered.a110248

Nachman, M. W., & Payseur, B. A. (2012). Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. *Philos Trans R Soc Lond B Biol Sci, 367*(1587), 409-421. doi:10.1098/rstb.2011.0249

Navarro, A., & Barton, N. H. (2003). Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *Evolution, 57*(3), 447-459. doi:10.1111/j.0014-3820.2003.tb01537.x

Noor, M. A., & Bennett, S. M. (2009). Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity (Edinb), 103*(6), 439-444. doi:10.1038/hdy.2009.151

Noor, M. A., Grams, K. L., Bertucci, L. A., & Reiland, J. (2001). Chromosomal inversions and the reproductive isolation of species. *Proc. Natl. Acad. Sci. USA, 98*(21), 12084-12088.

Noor, M. A. F., & Aquadro, C. F. (1998). Courtship songs of Drosophila pseudoobscura and D. persimilis: analysis of variation. *Anim Behav, 56*(1), 115-125. doi:10.1006/anbe.1998.0779

Noor, M. A. F., Garfield, D. A., Schaeffer, S. W., & Machado, C. A. (2007). Divergence between the Drosophila pseudoobscura and D-persimilis genome sequences in relation to chromosomal inversions. *Genetics, 177*(3), 1417-1428. doi:10.1534/genetics.107.070672

Noor, M. A. F., Grams, K. L., Bertucci, L. A., Almendarez, Y., Reiland, J., & Smith, K. R. (2001). The genetics of reproductive isolation and the potential for gene exchange between *Drosophila pseudoobscura* and *D. persimilis* via backcross hybrid males. *Evolution, 55*(3), 512-521.

Nosil, P., & Feder, J. L. (2012). Genomic divergence during speciation: causes and consequences. *Philos Trans R Soc Lond B Biol Sci, 367*(1587), 332-342. doi:10.1098/rstb.2011.0263

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., . . . Phillippy, A. M. (2022). The complete sequence of a human genome. *Science, 376*(6588), 44-53. doi:10.1126/science.abj6987

Nyberg, K. G., & Machado, C. A. (2016). Comparative Expression Dynamics of Intergenic Long Noncoding RNAs in the Genus Drosophila. *Genome Biol Evol, 8*(6), 1839-1858. doi:10.1093/gbe/evw116

O'Neill, K., Brocks, D., & Hammell, M. G. (2020). Mobile genomics: tools and techniques for tackling transposons. *Philos Trans R Soc Lond B Biol Sci, 375*(1795), 20190345. doi:10.1098/rstb.2019.0345

Oldham, M. C., Horvath, S., & Geschwind, D. H. (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc Natl Acad Sci U S A, 103*(47), 17973-17978. doi:10.1073/pnas.0605938103

Orr, H. A. (1987). Genetics of male and female sterility in hybrids of Drosophila pseudoobscura and D. persimilis. *Genetics, 116*(4), 555-563. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/3623079
https://www.genetics.org/content/genetics/116/4/555.full.pdf

Orr, H. A. (1995). The population genetics of speciation: the evolution of hybrid incompatibilities. *Genetics, 139*(4), 1805-1813. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/7789779
https://www.genetics.org/content/genetics/139/4/1805.full.pdf

Ortiz-Barrientos, D., Engelstadter, J., & Rieseberg, L. H. (2016). Recombination Rate Evolution and the Origin of Species. *Trends in Ecology & Evolution, 31*(3), 226-236. doi:10.1016/j.tree.2015.12.016

Ovens, K., Eames, B. F., & McQuillan, I. (2021). Comparative Analyses of Gene Co-expression Networks: Implementations and Applications in the Study of Evolution. *Frontiers in Genetics, 12*, 695399. doi:10.3389/fgene.2021.695399

Ozata, D. M., Gainetdinov, I., Zoch, A., O'Carroll, D., & Zamore, P. D. (2019). PIWI-interacting RNAs: small RNAs with big functions. *Nature Reviews Genetics, 20*(2), 89-108. doi:10.1038/s41576-018-0073-3

Paris, M., Villalta, J. E., Eisen, M. B., & Lott, S. E. (2015). Sex Bias and Maternal Contribution to Gene Expression Divergence in Drosophila Blastoderm Embryos. *Plos Genetics, 11*(10), e1005592. doi:10.1371/journal.pgen.1005592

Patlar, B., & Civetta, A. (2021). Speciation and changes in male gene expression in Drosophila. *Genome, 64*(2), 63-73. doi:10.1139/gen-2020-0025

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods, 14*(4), 417-419. doi:10.1038/nmeth.4197

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., . . . Reich, D. (2012). Ancient admixture in human history. *Genetics, 192*(3), 1065-1093. doi:10.1534/genetics.112.145037

Pavey, S. A., Collin, H., Nosil, P., & Rogers, S. M. (2010). The role of gene expression in ecological speciation. *Annals of the New York Academy of Sciences, 1206*(1), 110-129. doi:10.1111/j.1749-6632.2010.05765.x

Pembroke, W. G., Hartl, C. L., & Geschwind, D. H. (2021). Evolutionary conservation and divergence of the human brain transcriptome. *Genome Biology, 22*(1), 52. doi:10.1186/s13059-020-02257-z

Pertea, G., & Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. *F1000Res, 9*. doi:10.12688/f1000research.23297.2

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol, 33*(3), 290-295. doi:10.1038/nbt.3122

Pettersen, J. P., & Almaas, E. (2022). csdR, an R package for differential co-expression analysis. *BMC Bioinformatics, 23*(1), 79. doi:10.1186/s12859-022-04605-1

Pfeifer, B., Alachiotis, N., Pavlidis, P., & Schimek, M. G. (2020). Genome scans for selection and introgression based on k-nearest neighbour techniques. *Mol Ecol Resour, 20*(6), 1597-1609. doi:10.1111/1755-0998.13221

Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S. E., & Lercher, M. J. (2014). PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol Biol Evol, 31*(7), 1929-1936. doi:10.1093/molbev/msu136

Phadnis, N. (2011). Genetic architecture of male sterility and segregation distortion in Drosophila pseudoobscura Bogota-USA hybrids. *Genetics, 189*(3), 1001-1009. doi:10.1534/genetics.111.132324

Phadnis, N., & Orr, H. A. (2009). A Single Gene Causes Both Male Sterility and Segregation Distortion in Drosophila Hybrids. *Science, 323*(5912), 376-379. doi:10.1126/science.1163934

Pinho, C., & Hey, J. (2010). Divergence with Gene Flow: Models and Data. *Annual Review of Ecology, Evolution, and Systematics, 41*(1), 215-230. doi:10.1146/annurev-ecolsys-102209-144644

Policansky, D., & Zouros, E. (1977). Gene Differences between Sex-Ratio and Standard Gene Arrangements of X-Chromosome in Drosophila-Persimilis. *Genetics, 85*(3), 507-511. Retrieved from <Go to ISI>://WOS:A1977DE21900014
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1224583/pdf/507.pdf

Ponjavic, J., Ponting, C. P., & Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res, 17*(5), 556-565. doi:10.1101/gr.6036807

Presgraves, D. C. (2010). Speciation genetics: search for the missing snowball. *Curr Biol, 20*(24), R1073-1074. doi:10.1016/j.cub.2010.10.056

R Core Team. (2021). R: A language and environment for statistical computing: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rago, A., Werren, J. H., & Colbourne, J. K. (2020). Sex biased expression and co-expression networks in development, using the hymenopteran Nasonia vitripennis. *Plos Genetics, 16*(1). doi:ARTN e1008518

10.1371/journal.pgen.1008518

Rancati, G., Moffat, J., Typas, A., & Pavelka, N. (2018). Emerging and evolving concepts in gene essentiality. *Nature Reviews Genetics, 19*(1), 34-49. doi:10.1038/nrg.2017.74

Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., . . . Westram, A. M. (2017). Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *Journal of Evolutionary Biology, 30*(8), 1450-1477. doi:10.1111/jeb.13047

Renaut, S., Grassa, C. J., Yeaman, S., Moyers, B. T., Lai, Z., Kane, N. C., . . . Rieseberg, L. H. (2013). Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nat Commun, 4*, 1827. doi:10.1038/ncomms2833

Reyna, M. A., Haan, D., Paczkowska, M., Verbeke, L. P. C., Vazquez, M., Kahraman, A., . . . Consortium, P. (2020). Pathway and network analysis of more than 2500 whole cancer genomes. *Nature Communications, 11*(1). doi:ARTN 729

10.1038/s41467-020-14367-0

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., . . . Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature, 592*(7856), 737-+. doi:10.1038/s41586-021-03451-0

Richards, S., Liu, Y., Bettencourt, B. R., Hradecky, P., Letovsky, S., Nielsen, R., . . . Gibbs, R. A. (2005). Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. *Genome Res, 15*(1), 1-18. doi:10.1101/gr.3059305

Rieseberg, L. H. (2001). Chromosomal rearrangements and speciation. *Trends Ecol Evol, 16*(7), 351-358. doi:10.1016/s0169-5347(01)02187-5

Saetre, G. P. (2014). Genome scans and elusive candidate genes: detecting the variation that matters for speciation. *Mol Ecol, 23*(19), 4677-4678. doi:10.1111/mec.12905

Samuk, K., Manzano-Winkler, B., Ritz, K. R., & Noor, M. A. F. (2020). Natural Selection Shapes Variation in Genome-wide Recombination Rate in Drosophila pseudoobscura. *Curr Biol, 30*(8), 1517-1528 e1516. doi:10.1016/j.cub.2020.03.053

Samuk, K., & Noor, M. A. F. (2022). Gene flow biases population genetic inference of recombination rate. *G3 (Bethesda), 12*(11). doi:10.1093/g3journal/jkac236

Samuk, K., Owens, G. L., Delmore, K. E., Miller, S. E., Rennison, D. J., & Schluter, D. (2017). Gene flow and selection interact to promote adaptive divergence in regions of low recombination. *Mol Ecol, 26*(17), 4378-4390. doi:10.1111/mec.14226

Sanchez-Flores, A., Penaloza, F., Carpinteyro-Ponce, J., Nazario-Yepiz, N., Abreu-Goodger, C., Machado, C. A., & Markow, T. A. (2016). Genome Evolution in Three Species of Cactophilic Drosophila. *G3 (Bethesda), 6*(10), 3097-3105. doi:10.1534/g3.116.033779

Schaeffer, S. W., Bhutkar, A., McAllister, B. F., Matsuda, M., Matzkin, L. M., O'Grady, P. M., . . . Kaufman, T. C. (2008). Polytene chromosomal maps of 11 Drosophila species: the

order of genomic scaffolds inferred from genetic and physical maps. *Genetics, 179*(3), 1601-1655. doi:10.1534/genetics.107.086074

Schaeffer, S. W., Goetting-Minesky, M. P., Kovacevic, M., Peoples, J. R., Graybill, J. L., Miller, J. M., . . . Anderson, W. W. (2003). Evolutionary genomics of inversions in Drosophila pseudoobscura: evidence for epistasis. *Proc. Natl. Acad. Sci. USA, 100*(14), 8319-8324. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12824467

Schwarzer, J., Misof, B., & Schliewen, U. K. (2012). Speciation within genomic networks: a case study based on Steatocranus cichlids of the lower Congo rapids. *Journal of Evolutionary Biology, 25*(1), 138-148. doi:10.1111/j.1420-9101.2011.02409.x

Scully, R., Panday, A., Elango, R., & Willis, N. A. (2019). DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nature Reviews Molecular Cell Biology, 20*(11), 698-714. doi:10.1038/s41580-019-0152-0

Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., . . . Paten, B. (2020). Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature Biotechnology, 38*(9), 1044-+. doi:10.1038/s41587-020-0503-6

Shahan, R., Zawora, C., Wight, H., Sittmann, J., Wang, W., Mount, S. M., & Liu, Z. (2018). Consensus Coexpression Network Analysis Identifies Key Regulators of Flower and Fruit Development in Wild Strawberry. *Plant Physiol, 178*(1), 202-216. doi:10.1104/pp.18.00086

Sharma, S. P., Zuo, T., & Peterson, T. (2021). Transposon-induced inversions activate gene expression in the maize pericarp. *Genetics, 218*(2). doi:10.1093/genetics/iyab062

Soneson, C., Love, M. I., & Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res, 4*, 1521. doi:10.12688/f1000research.7563.2

Sridharan, V., Heimiller, J., Robida, M. D., & Singh, R. (2016). High Throughput Sequencing Identifies Misregulated Genes in the Drosophila Polypyrimidine Tract-Binding Protein (hephaestus) Mutant Defective in Spermatogenesis. *Plos One, 11*(3). doi:ARTN e0150768

10.1371/journal.pone.0150768

Stephan, W. (2010). Genetic hitchhiking versus background selection: the controversy and its implications. *Philosophical Transactions of the Royal Society B-Biological Sciences, 365*(1544), 1245-1253. doi:10.1098/rstb.2009.0278

Stevison, L. S., Hoehn, K. B., & Noor, M. A. F. (2011). Effects of Inversions on Within- and Between-Species Recombination and Divergence. *Genome Biology and Evolution, 3*, 830-841. doi:10.1093/gbe/evr081

Stevison, L. S., & Noor, M. A. (2010). Genetic and evolutionary correlates of fine-scale recombination rate variation in Drosophila persimilis. *Journal of Molecular Evolution, 71*(5-6), 332-345. doi:10.1007/s00239-010-9388-1

Tan, C. C. (1935). Salivary Gland Chromosomes in the Two Races of Drosophila Pseudoobscura. *Genetics, 20*(4), 392-402. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/17246768

https://www.genetics.org/content/genetics/20/4/392.full.pdf

Taylor, S. A., & Larson, E. L. (2019). Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nat Ecol Evol, 3*(2), 170-177. doi:10.1038/s41559-018-0777-y

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., . . . Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A, 102*(39), 13950-13955. doi:10.1073/pnas.0506758102

Thurmond, J., Goodman, J. L., Strelets, V. B., Attrill, H., Gramates, L. S., Marygold, S. J., . . . FlyBase, C. (2019). FlyBase 2.0: the next generation. *Nucleic Acids Res, 47*(D1), D759-D765. doi:10.1093/nar/gky1003

Tsaparas, P., Marino-Ramirez, L., Bodenreider, O., Koonin, E. V., & Jordan, I. K. (2006). Global similarity and local divergence in human and mouse gene co-expression networks. *Bmc Evolutionary Biology, 6*, 70. doi:10.1186/1471-2148-6-70

van Dam, S., Vosa, U., van der Graaf, A., Franke, L., & de Magalhaes, J. P. (2018). Gene co-expression analysis for functional classification and gene-disease predictions. *Briefings in Bioinformatics, 19*(4), 575-592. doi:10.1093/bib/bbw139

Van der Auwera, G. A., & O'Connor, B. D. (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra* (illustrated ed.): O'Reilly Media, Incorporated, 2020.

Via, S. (2012). Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philos Trans R Soc Lond B Biol Sci, 367*(1587), 451-460. doi:10.1098/rstb.2011.0260

Vieillard, J., Paschaki, M., Duteyrat, J. L., Augiere, C., Cortier, E., Lapart, J. A., . . . Durand, B. (2016). Transition zone assembly and its contribution to axoneme formation in Drosophila male germ cells. *Journal of Cell Biology, 214*(7), 875-889. doi:10.1083/jcb.201603086

Voigt, A., Nowick, K., & Almaas, E. (2017). A composite network of conserved and tissue specific gene interactions reveals possible genetic interactions in glioma. *PLoS Comput Biol, 13*(9), e1005739. doi:10.1371/journal.pcbi.1005739

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics, 33*(14), 2202-2204. doi:10.1093/bioinformatics/btx153

Weckselblatt, B., & Rude, M. K. (2015). Human Structural Variation: Mechanisms of Chromosome Rearrangements. *Trends in Genetics, 31*(10), 587-599. doi:10.1016/j.tig.2015.05.010

Wei, K. H. C., Lower, S. E., Caldas, I. V., Sless, T. J. S., Barbash, D. A., & Clark, A. G. (2018). Variable Rates of Simple Satellite Gains across the Drosophila Phylogeny. *Molecular Biology and Evolution, 35*(4), 925-941. doi:10.1093/molbev/msy005

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

Wei, W., Jin, Y. T., Du, M. Z., Wang, J., Rao, N., & Guo, F. B. (2016). Genomic Complexity Places Less Restrictions on the Evolution of Young Coexpression Networks than Protein-Protein Interactions. *Genome Biol Evol, 8*(8), 2624-2631. doi:10.1093/gbe/evw198

Weissensteiner, M. H., Bunikis, I., Catalan, A., Francoijs, K. J., Knief, U., Heim, W., . . . Wolf, J. B. W. (2020). Discovery and population genomics of structural variation in a songbird genus. *Nat Commun, 11*(1), 3403. doi:10.1038/s41467-020-17195-4

Wellenreuther, M., Merot, C., Berdan, E., & Bernatchez, L. (2019). Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. *Mol Ecol*, *28*(6), 1203-1209. doi:10.1111/mec.15066

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., . . . Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*(10), 1155-+. doi:10.1038/s41587-019-0217-9

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*: Springer-Verlag New York.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., . . . Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi:10.21105/joss.01686

Wolf, J. B., & Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*, *18*(2), 87-100. doi:10.1038/nrg.2016.133

Wu, C. I., & Ting, C. T. (2004). Genes and speciation. *Nature Reviews Genetics*, *5*(2), 114-122. doi:10.1038/nrg1269

Wu, L. F., Hughes, T. R., Davierwala, A. P., Robinson, M. D., Stoughton, R., & Altschuler, S. J. (2002). Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters. *Nature Genetics*, *31*(3), 255-265. doi:10.1038/ng906

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

Yang, H. W., Jaime, M., Polihronakis, M., Kanegawa, K., Markow, T., Kaneshiro, K., & Oliver, B. (2018). Re-annotation of eight Drosophila genomes. *Life Science Alliance*, *1*(6). doi:ARTN e201800156

10.26508/lsa.201800156

Yeaman, S., & Whitlock, M. C. (2011). The genetic architecture of adaptation under migration-selection balance. *Evolution*, *65*(7), 1897-1911. doi:10.1111/j.1558-5646.2011.01269.x

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

Zhang, L., Reifová, R., Halenková, Z., & Gompert, Z. (2021). How Important Are Structural Variants for Speciation? *Genes*, *12*, 1084. doi:https://doi.org/10.3390/genes12071084

Zichner, T., Garfield, D. A., Rausch, T., Stutz, A. M., Cannavo, E., Braun, M., . . . Korbel, J. O. (2013). Impact of genomic structural variation in Drosophila melanogaster based on population-scale sequencing. *Genome Research*, *23*(3), 568-579. doi:10.1101/gr.142646.112

Zimin, A. V., Puiu, D., Luo, M. C., Zhu, T., Koren, S., Marcais, G., . . . Salzberg, S. L. (2017). Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res*, *27*(5), 787-792. doi:10.1101/gr.213405.116