The Genomic Landscape of the *Drosophila nasuta* Clade

By

Dat Mai

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Integrative Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Doris Bachtrog, Chair
Professor Rasmus Nielsen
Professor Benjamin Blackman

Summer 2021

Abstract

The Genomic Landscape of the *Drosophila nasuta* Clade

by

Dat Mai

Doctor of Philosophy in Integrative Biology

University of California, Berkeley

Professor Doris Bachtrog, Chair

The *Drosophila nasuta* clade is a young, rapidly speciating clade comprising approximately a dozen species. They are widely distributed in Asia with populations also found in Hawaii, East Africa, and Oceania; and recently, *D. nasuta* has been identified as an invasive species in Brazil and is quickly spreading in South America. There are few morphological differences between species; females are indistinguishable between species while males of different species have one of three differentially patterned frons--silvery patches between their eyes--and thoracic banding patterns that are correlated to the frons pattern. There are also varying levels of reproductive isolation between species with over half of interspecific matings producing viable offspring. These characteristics make the *D. nasuta* clade a promising species group to study speciation.

The first chapter of this dissertation focuses on the development of genomic resources to study a young, non-model *Drosophila* species group. The phylogenetic relationship between species of the *Drosophila nasuta* clade have been inferred multiple times using: mitochondrial genes, X chromosome genes, courtship song, and frons patterning. However, these phylogenies have been inconsistent. We leveraged PacBio SMRT long read sequencing technology to generate a chromosome level genome assembly for *D. albomicans*, a member of the *D. nasuta* clade. Sixty eight individuals across the clade were sequenced and phylogenetic analyses on whole genome polymorphism data were used to determine the true species phylogeny. While there were multiple phylogenetic topologies across the genome—likely due to incomplete lineage sorting or widespread introgression—there were two that made up 56% of all topologies and were highly correlated with either the autosomes or X chromosome. We found that the inconsistency between the autosomal and X chromosome topology was due to introgression on the autosomes and, thus, determining that the X chromosome phylogeny reflects the true species relationship in the *D. nasuta* clade. This chapter highlights the *nasuta* clade's potential in studying the evolution of pre- and postzygotic isolation and provides a foundation and genetic resources for such endeavors.

The second chapter of this dissertation focuses on inversions between species. A hallmark process of speciation is the cessation of gene flow and accumulation of mutations between populations leading up to new species and genomic inversions are one such barrier, especially in the case of radiations like the *D. nasuta* clade. Inversions prevent recombination and allow for the accumulation of differences between inversion haplotypes. We reconstructed the ancestral karyotype using sequence homology and identified 22 inversions across the phylogeny based on the genome assemblies of *D. albomicans, D. nasuta, D. kepulauana, D. sulfurigaster albostrigata, D. sulfurigaster bilimbata, D. sulfurigaster sulfurigaster,* and *D. pallidifrons* generated in chapter 3 of this dissertation. While the overall inversion rate is consistent with previous studies in *Drosophila*, we find highly variable rates along the different branches of the phylogeny. Additionally, we find higher rates of inversions on the X chromosome relative to autosomes. Upon closer inspection of six autosomal inversions, four of them have repeat sequences associated with them. This implies the importance of ectopic recombination in generating inversions. The characterization of inversions between species in the *nasuta* clade contributes to future population genetics and functional genomics studies in the species group.

The third and final chapter of this dissertation looks at transposable elements (TEs) through a phylogenomic framework using the *D. nasuta* clade--namely how they affect gene expression and how frequently they escape TE repression and expand. We generated six high quality genome assemblies using long read technology for *D. nasuta, D. kepulauana, D. s. albostrigata, D. s. bilimbata, D. s. sulfurigaster,* and *D. pallidifrons* and improved on the *D. albomicans* assembly generated in chapter 1. Leveraging these assemblies, we generated a *de novo* repeat library for the species group and identified 147 TE families that have expanded in at least one of the species; one TE of note is a DINE element, which has shown multiple instances of expansion with thousands of copies in each genome. Additionally, we find a positive correlation between TEs that have expanded and their expression levels, which follows the expected pattern of suppression escape by the TE. Finally, we find patterns of more extreme gene expression--both elevated and downregulated--associated with TE insertions near or within genes.

Dedication

To Wilson

Acknowledgements

Graduate school has been a long road. Along the way I have had the opportunity to meet a multitude and variety of people, a good number of whom have been instrumental to helping me along my journey. Without them, I definitely wouldn’t have been able to succeed in this endeavor.

First and foremost, I thank my family and Calvin. They have been around since before grad school and have been my main support pillars. I appreciate the food-filled care packages in the mail, late night dim sum, or nice black coffee to tide me over the meh times.

I’d like to credit this journey to my undergrad professors, Dr. Wahlert and Dr. Greer. They showed me the fun in research and supported me during every step of my academic journey. May our annual meetings happen for many years to come.

I am grateful for the grad school friends I’ve made--things would have been so much more isolating and rife with bad choices without them. Extra props to Shivani Mahajan & Lauren Gibilisco; Kevin Wei & Alison Nguyen; Débora Brandt, Chenling Xu, Nina Pak, & Ana Lyons. They’re great pals.

I am glad to have the friends I’ve had since high school to keep me in touch with life beyond the lab. Shouts to Diane Lum, Gloria Lee, & Nicholas Wong. California may not be as great as where you are in NY... so I’m glad you still talk to me.

To the people who have been key to helping me with the next chapter of my life--beyond grad school--thanks a bunch. Thanks, Carina Cheng, Sumayah Rahman, Berling Chen, & Teng Lei. Without you, I would most definitely be homeless.

I would also like to thank current and past committee members for advice and feedback on my projects over the years. Thank you, Drs. Doris Bachtrog, Rasmus Nielsen, Benjamin Blackman, & Patrick O’Grady.

Finally, I thank the Bachtrog lab, both past and present. I appreciate the guidance and fun times we’ve had over the years. Thanks Shivani Mahajan, Lauren Gibilisco, Kevin Wei, Alison Nguyen, Emily Landeen, Ryan Bracewell, Kamalakar Chatla, Carolus Chan, Reema Aldaimalani, Beatriz Vicoso, Qi Zhou, Wynn Meyer, Zaak Walton, Chris Ellison, Emily Brown, Silu Wang, and, once more, Doris Bachtrog.

Chapter 1: Patterns of genomic differentiation in the *Drosophila nasuta* species complex

Dat Mai, Matthew J. Nalley, Doris Bachtrog*

Department of Integrative Biology, University of California Berkeley, Berkeley, California, United States of America.

* Corresponding Author

The Drosophila nasuta species complex contains over a dozen recently diverged species that are distributed widely across South-East Asia, and which show varying degrees of pre- and post-zygotic isolation. Here, we assemble a high-quality genome for D. albomicans using single-molecule sequencing and chromatin conformation capture, and draft genomes for 11 additional species and 67 individuals across the clade, to infer the species phylogeny and patterns of genetic diversity in this group. Our assembly recovers entire chromosomes, and we date the origin of this radiation about 2 million years ago. Despite low levels of overall differentiation, most species or subspecies show clear clustering into their designated taxonomic groups using population genetics and phylogenetic methods. Local evolutionary history is heterogeneous across the genome, and differs between the autosomes and the X chromosome for species in the sulfurigaster subgroup, likely due to autosomal introgression. Our study establishes the nasuta species complex as a promising model system to further characterize the evolution of pre- and post-zygotic isolation in this clade.

Introduction

Species radiations are responsible for most of today's biodiversity and are a prime study system to learn about the factors resulting in the origin of new species. Recent work in diverse species groups, ranging from humans, birds, fish to mosquitos, butterflies and other insects has highlighted that genealogical relationships among closely related species can be complex and can vary across the genome and among individuals (Martin et al. 2013; Brawand et al. 2014; Fontaine et al. 2015; Lamichhaney et al. 2015; Dannemann and Racimo 2018). Recently diverged species often have incomplete reproductive barriers and may hybridize. Ancestral polymorphism predating lineage splitting may also be sorted stochastically among descendant lineages (that is, incomplete lineage sorting, ILS). Phylogenetic heterogeneity can be caused both by hybridization and introgression and by incomplete lineage sorting in ancestral populations, causing some parts of the genome to have genealogies that are discordant with the species tree.

Genome-wide studies have revealed that certain genomic regions such as sex chromosomes can have distinct phylogenetic histories, possibly reflecting systematic differences in the extent of interspecific gene flow across the genome (Fontaine et al. 2015; Wong Miller et al. 2017; Fuller et al. 2018). Introgression can transfer beneficial alleles between closely related species, but interspecific gene flow can also be counteracted by natural selection at particular ‘barrier loci’

(Dannemann and Racimo 2018). Thus, the landscape of genomic divergence contains information on the evolutionary forces that contribute to the origin of new species (Martin and Jiggins 2017). Here we characterize the evolutionary history of the rapidly radiating *nasuta* subgroup of the *immigrans* species group of *Drosophila*. The *nasuta* group consists of more than a dozen closely related species or subspecies that are widely distributed across South-East Asia (Wilson et al. 1969; Kitagawa et al. 1982), and which show varying degrees of pre- and postzygotic isolation. Members from different species or subspecies often produce viable and sometimes fertile hybrids (Kitagawa et al. 1982), but show differences in their behavior and morphology (Spieth 1969; Wilson et al. 1969; Kitagawa et al. 1982).

Females of the *nasuta* species complex are indistinguishable from their external morphology. However the males can be differentiated into phenotypic groups based on markings on the frons and thorax (Wilson et al. 1969; Kitagawa et al. 1982; see **Figure 1A-F**). The first category includes species where males have a continuous silver patch on their frons and dark bands on their thorax (i.e. *D. nasuta*, *D. albomicans, D. kepulauana* and *D. kohkoa*). Species in the second category have prominent whitish orbits along the edges of their compound eyes and slightly dark thoracic bands; these include all subspecies of the *D. sulfurigaster* sp. group. *D. s. albostrigata* and *D. s. neonasuta* have broader bands than *D. s. sulfurigaster* and *D. s. bilimbata*. *D. pulaua* males have very pale white bands. The third category contains species without whitish patterns (*D. pallidifrons*, Taxon-F). The darkness of the bands on the mesopleuron on the thorax is correlated with the coloration of the frons, with flies with more bright areas on the frons showing more dark bands on the thorax. *D. niveifrons* males have an X-shaped silver patch on their forehead and no coloration on their thorax.

Species in this group also display clear differences in mating behavior (Spieth 1969), and both acoustic and visual signals appear important during courtship display (Spieth 1969). Courtship songs, caused by wing vibration of the courting males, are often species-specific and contribute to prezygotic isolation between closely related species (Gleason and Ritchie 1998). Indeed, species or species groups in the *nasuta* species clade differ in male song, both with regards to quantitative and qualitative song parameters (Shao et al. 1997; Nalley and Bachtrog unpublished). Visual stimuli have also diverged among species in this group. During courtship, males in this species group show species-specific patterns of wing displays, circling of the females, and frontal displays of the males (Spieth 1969; Kitagawa et al. 1982).

Patterns of hybrid viability and sterility are complex within the *nasuta* species group (Kitagawa et al. 1982). In general, flies with similar frons patterns often produce viable and fertile hybrids (but *D. kohkoa*, for example is clearly more reproductively isolated from other species with continuous white frons) and other crosses also sometimes produce viable offspring (in particular, *D. albomicans* females produce viable, but often sterile crosses with several species; Kitagawa et al. 1982).

Thus, levels of both pre- and postzygotic isolation differ among members of this species group, making it an ideal system to study the evolution of sexual isolation. *D. albomicans* is of special interest in this clade, because of its recently formed neo-sex chromosomes: chromosomal fusions between an autosome and both the X and Y have created a neo-sex chromosome roughly about 100,000 years ago (**Figure 1G,H**). Neo-sex chromosomes of *Drosophila* have served as a powerful tool to study the evolutionary forces driving sex chromosome differentiation (Bachtrog and Charlesworth 2002; Zhou and Bachtrog 2015; Mahajan et al. 2018).

Despite its general promise as a model system for speciation genomics, and detailed morphological, behavioral and genetic investigations, little is known about the phylogenetic relationship among members of this group, or general patterns of sequence differentiation, and the correct species branching order has remained controversial and unresolved (Yu et al. 1999; Nagaraja, Nagaraju, and Ranganath 2004). Here, we utilize whole genome sequencing to study patterns of genomic differentiation in the *nasuta* species complex. We assemble a high-quality genome of *D. albomicans* using single-molecule sequencing and chromatin conformation capture, and draft genomes for 11 additional species, and obtain genome-wide polymorphism data for a total of 67 strains of the *nasuta* group (**Table S1**, **Figure 1I**). This comprehensive data set allowed us to clarify species phylogenetic relationships, and describe overall patterns of differentiation and divergence among species in this group. As expected for such a recently diverged species group, patterns of genomic differentiation are highly heterogeneous across the genome. Detailed knowledge of background levels of genomic differentiation will provide a foundation for future studies on the genetic basis of pre- and postzygotic isolation in this clade.

Results

Assembly of D. albomicans genome and annotation

D. albomicans is a species of particular interest in this clade, due to its recently formed neo-sex chromosomes (**Figure 1H**; Zhou et al. 2012). We used a combination of single-molecule long sequencing reads, Illumina reads, and chromatin conformation capture to create a chromosome level genome assembly of *D. albomicans* (**Fig. S1**). Our final assembly is 165.8 Mb in size, with an N50 of 33.4 Mb (**Figure 2**), and with all of the major chromosomes being contained within a single scaffold (**Figure 2A**). We verified X-linked scaffolds on the basis of significant differences in read depth between males (XY) and females (XX) (**Figure 2B**). As expected, Muller element A shows half the coverage in males relative to females; Muller CD (the neo-sex chromosome), on the other hand, shows similar levels of genomic coverage in both sexes. This means that most reads from the neo-Y in males still fully map to the neo-X, indicating low levels of differentiation between the neo-sex chromosomes. Our final genome annotation contained 12,387 genes, and the repeat content is about 21%. We examined the genome for completeness using BUSCO scores (Simão et al. 2015), and found that 98% of core eukaryotic genes were present in our reference genome (**Table S2**). This assembly is a significant improvement over a previous one based on Illumina reads (**Fig. S2**; Zhou et al. 2012).

Clustering of species and population

We re-sequenced 67 individuals from 11 species across the *nasuta* species group (median sequence coverage per fly 24-fold; **Fig. S3**). Reads were aligned to the genome assembly generated for *D. albomicans*, and stringent variant calling revealed approximately 17.6 million variable sites within or between populations. We found considerable levels of genetic diversity (average pairwise diversity π) within each species, in the range 0.18% to 0.61% (Table S3), similar to that reported in other *Drosophila* populations. Genetic diversity within species does not appear to be determined by geographic range: *D. s. bilimbata* is widely scattered on many islands in the Pacific Ocean but has the lowest level of genetic diversity (0.18%), while *D. albomicans* has the highest level of diversity (0.61%) yet a more limited distribution than its close sister *D. nasuta* (Spieth 1969; Wilson et al. 1969; Kitagawa et al. 1982). Pairwise diversity between *D. nasuta* strains increases with geographic distance (isolation by distance; **Fig. S4**). We compared the proportions of shared and fixed SNPs between species in the *nasuta* group (**Figure 3A**; **Fig. S5**). Extensive sharing of genetic variation and few fixed differences among populations was evident, particularly among subspecies of the *sulfurigaster* group and *D. pulaua*, and between *D. albomicans*, *D. nasuta*, and *D. kepulauana* (**Figure 3A**), indicative of their recent divergence time. *D. niveifrons* appeared most divergent from all other species (**Figure 3A**). Principle component analysis (PCA) revealed similar patterns of clustering between species, with flies from the *sulfurigaster* group and *D. pulaua* consistently forming a cluster, and *D. albomicans*, *D. nasuta*, and *D. kepulauana* clustering (**Figure 3B**). Interestingly, one strain of *D. albomicans* (E-10815_SHL48) clusters more closely with *D. nasuta* on the autosome compared to other *D. albomicans* strains; admixture analysis reveals that this strain indeed has some *D. nasuta* ancestry (Cheng, Mailund, and Nielsen 2017; **Fig. S6**). We also find clusters of *D. pallidifrons* and *D. kohkoa* flies. *D. s. albostrigata* and *D. s. neonasuta* consistently overlap in the PCA analysis (**Figure 3B**), and also do not separate in the structure analysis (**Fig. S6**), indicating that they are genetically indistinguishable from each other. *D. s. sulfurigaster* and *D. pulaua* also fail to clearly separate in the structure plots, especially for autosomes (**Fig. S6**). *D. s. bilimbata* individuals form their own group, but *D. s. bilimbata* strain 1821.03 shows high levels of *D. s. sulfurigaster* / *D. pulaua* ancestry on the autosomes (**Fig. S6**).

Phylogenomic clustering of species

We inferred phylogenetic relationships among species using non-overlapping 500-kb genomic windows (Stamatakis 2014), and inferred consensus trees separately for the X chromosome and the autosomes (Mirarab et al. 2014; C. Zhang et al. 2018). Our species tree topology is overall consistent with groupings based on PCA, identifying the same major clades (**Figure 4**). In particular, *D. albomicans* and *D. nasuta* are sister taxa, and group with *D. kepulauana* (the *nasuta* subclade). Likewise, all the different *sulfurigaster* subspecies form a cluster together with *D. pulaua*, with *D. s. neonasuta* and *D. s. albostrigata* being intermingled in the tree (in agreement with the clustering analysis above), and with *D. s. sulfurigaster*, *D. s. bilimbata* and *D. pulaua* forming a separate group. Interestingly, however, the topology of the consensus tree for this subgroup differs for the X and the autosomes: for the autosomes, *D. s. sulfurigaster*, and *D. s. bilimbata* cluster and *D. pulaua* is the outgroup, while the X topology places *D. s. bilimbata* as the outgroup (**Figure 4**). *D. pallidifrons* is most closely related to Taxon F and *D. kohkoa*, and they form the outgroup to the *sulfurigaster* clade, and *D. niveifrons* is most distantly related to all other flies investigated (**Figure 4**).

Using molecular clock estimates, we dated several nodes that define major groups and distinct species. Our inferred date for the basal node suggests that this species group started to diverge about 2 MY ago. Assuming a neutral mutation rate of 3.46 x 10^{-9} per year (Keightley et al. 2009a), we estimate that *D. nasuta* and *D. albomicans* diverged roughly 0.6 MY ago (Ks=0.030), and split from *D. kepulauana* about 0.7 MY ago (Ks=0.034). The *nasuta* clade diverged from the *sulfurigaster* clades roughly 1 MY ago (Ks=0.047) and about 1.8 MY ago from *D. niveifrons* (Ks=0.089). Thus, sequence divergence confirms that species within the *nasuta* group split only very recently, consistent with patterns of incomplete pre- and postzygotic isolation in this clade.

Heterogeneity in patterns of ancestry across the genome

While we generally find strong support for the inferred species tree, it conceals rampant phylogenetic complexity that is evident when examining the evolutionary history of more defined genomic regions. In particular, we analyzed the distribution of ancestry across the genome for the species (using a randomly selected individual from each group) by constructing trees in 500-kb (or 50-kb) sliding windows (**Figure 5**, **Fig. S7**). Consistent with the inferred consensus trees, we find that the most prevalent topologies differ between the X and the autosomes. The most common topology is found in 35% of the windows (19% of the X windows, and 42% of the autosomal windows), and the second most common topology (21% of windows) dominates on the X chromosomes (51% of windows on X, vs. 12% of windows on autosomes). The third and fourth most common topologies are found in only 4% and 3% of windows, mostly on the autosomes. Conflicting signals in the distribution of ancestry across the genome may reflect incomplete lineage sorting and/or gene flow.

Increased introgression on the X chromosome between *D. s. sulfurigaster* and *D. pulaua* or autosomal introgression between *D. s. sulfurigaster* and *D. s. bilimbata* could account for the observed discrepancy of X-linked and autosomal topologies (Fontaine et al. 2015; **Figure 6**). The X chromosome often has a disproportionately large effects on hybrid sterility (the large X-effect; Masly and Presgraves 2007; Presgraves 2018), and autosomes may thus introgress more easily across species boundaries. Introgression will reduce sequence divergence between the species exchanging genes (E. Y. Durand et al. 2011; Fontaine et al. 2015). Thus, gene trees constructed from non-introgressed sequences should show deeper divergences than those constructed from introgressed sequences. To identify the correct species branching order, we inferred the length of autosomal topologies that support each of the possible groupings between *D. s. sulfurigaster*, *D. s. bilimbata* and *D. pulaua*. If autosomal introgression resulted in conflicting phylogenetic signals, we would expect that topologies supporting the majority X chromosome grouping (that is, (*D. s. bilimbata*, (*D. s. sulfurigaster*, *D. pulaua*))) to show higher divergence times than those supporting the majority autosomal topology (Fontaine et al. 2015). To estimate divergence times, we chose a random *D. s. sulfurigaster*, *D. s. bilimbata* and *D. pulaua* strain and followed the procedure outlined in (Fontaine et al. 2015). In particular, there are three possible topologies and two divergence times for each tree (T_1 and T_2) for this trio (**Figure 6A**). We compared mean values of T_1 and T_2 between the three possible topologies, only focusing on trees derived from the autosomes (since many confounding factors differ between the X chromosome and autosomes (Fontaine et al. 2015). Indeed, the set of (autosomal) trees supporting the majority X chromosome topology (*D. s. bilimbata*, (*D. s. sulfurigaster*, *D. pulaua*)) had longer branches, as measured by both T_1 and T_2 ($p<0.05$ and $p<10^{-4}$), than those supporting the majority autosomal tree (**Figure 6B**). This indicates that the species branching order inferred from the X chromosome is likely the correct topology, and that extensive autosomal introgression has resulted in a different majority phylogeny for the autosomes.

To identify genomic regions that have introgressed between species in the recent past, we used the G_{min} statistics (Geneva et al. 2015). G_{min} measures the ratio of the minimum pairwise sequence distance between species to the average pairwise distance between species, and is sensitive to genealogical configurations resulting from recent gene flow where the minimum pairwise divergence (and thus G_{min}) is small relative to the mean pairwise distance (**Fig. S8**). A total of 11.9% of autosomal 50-kb windows (294 out of 2464 windows) support significant introgression based on G_{min} between *D. s. sulfurigaster* and *D. s. bilimbata* but only 0.1% of windows on the X (1 out of 669 windows). In contrast, we find a similar fraction of introgressed windows on the X and autosomes for the *D. s. sulfurigaster* and *D. pulaua* comparison: 7.7% of significant windows on autosomes and 5.4% on the X. Thus, patterns of introgression, as inferred by the G_{min} statistic, indicate pervasive introgression at autosomes between *D. s. sulfurigaster* and *D. s. bilimbata*. Note, however, that most of the small autosomal G_{min} values are caused by *D. s. bilimbata* strain 1821.03 (**Fig. S8**), which also show signatures of mixed ancestry in the structure analysis (**Fig. S6**). We also used the genealogy-based (ABBA-BABA) test, summarized by the *D* and f_D statistic (E. Y. Durand et al. 2011; Martin, Davey, and Jiggins 2015), to evaluate the distribution of shared derived variants between *D. s. sulfurigaster* and *D. s. bilimbata* on the X versus autosomes. Assuming a (((*D. s. sulfurigaster*, *D. pulaua*), *D. s. bilimbata*), *D. pallidifrons*) tree topology, we found significantly elevated values for both statistics on autosomes relative to the X chromosome (**Figure 6C**). This is indicative of a significant excess of shared derived sites between *D. s. sulfurigaster* and *D. s. bilimbata* on autosomes relative to the X, and provides complementary support for a history of increased levels of introgression on autosomes, potentially explaining the topological differences between the autosomal and X chromosome phylogeny. Indeed, we find that regions of the genome that support the alternative topology (*D. pulaua*, (*D. s. sulfurigaster*, *D. s. bilimbata*)) show elevated levels of introgression, as estimated by f_D (**Figure 6D**, $p<10^{-4}$).

Discussion

Drosophila has long served as a prominent model in speciation research, from describing macro-evolutionary patterns of diversification to identifying the molecular players involved in species incompatibilities (Dobzhansky 1937; Muller 1942; Orr 1993; Castillo and Barbash 2017). A large body of work to understand the genetic basis of reproductive isolation has focused on *D. melanogaster* and its sibling species (Presgraves 2003; Brideau et al. 2006; Bayes and Malik 2009; Ferree and Barbash 2009; Phadnis and Orr 2009). These studies benefit from the amazing repertoire of genetic tools available in this model organism, and have allowed the dissection of hybrid incompatibilities at the molecular and cellular level. However, *D. melanogaster* and its siblings have split >5 MY ago (Tamura, Subramanian, and Kumar 2004), and have accumulated a large number of hybrid incompatibilities since their reproductive isolation (Presgraves 2003;

Masly and Presgraves 2007). To identify the evolutionary forces and molecular pathways involved in the initial processes of species formation, it is necessary to investigate systems at the earliest stages of divergence (Phadnis and Orr 2009; Wong Miller et al. 2017). The *nasuta* radiation is therefore a group system to address questions on the genomics of speciation and adaptive radiations.

The *nasuta* species complex shows dramatic differences in patterns of pre- and postzygotic isolation, including divergence in courtship song and mating behavior, and male coloration (Spieth 1969; Wilson et al. 1969; Kitagawa et al. 1982). Yet many species pairs in this clade can form viable and often fertile hybrids, making it an ideal system to study the genetic basis of reproductive isolation. Our analyses establish phylogenetic relationships in this clade, and describe its evolutionary history, thereby providing a foundation for further detailed investigations of pre- and postzygotic barriers to gene flow. In addition, *D. albomicans* contains a recently formed sex chromosome, and genome-wide investigation of its young neo-X and neo-Y can yield important information about the initiation of sex chromosome divergence (Zhou and Bachtrog 2012), and its contribution to the formation of species boundaries (Kitano et al. 2009; Bracewell et al. 2017)(Kitano et al. 2009; Bracewell et al. 2017).

We generated a chromosome-level high-quality genome assembly for *D. albomicans* and reference-based "pseudogenomes" for the other species in the *nasuta* species group, to resolve phylogenetic relationships in this clade, and describe global patterns of differentiation and gene flow. In addition to having all euchromatic chromosome arms contained within a single scaffold, our assembly also recovers large parts of repeat-rich regions. In particular, we assembled 4 Mb of the repeat-rich dot chromosomes, about 1.25 Mb of the pericentromeric region on Muller B, and roughly 10 Mb of repeat-rich unmapped scaffolds (UH1-5, see **Figure 2B**) that presumably correspond to pericentromeric, heterochromatic regions. In total, our assembly contains about 18 Mb of sequence that is composed mainly of repetitive DNA (defined as 50% or more bp repeat-masked in 10-kb windows). Many genome assemblies, and in particular those using short-read sequencing data, are highly fragmented, and repeat-rich regions are typically missing (Simpson and Pop 2015). Yet, several recent studies have suggested that repetitive DNA, or genes interacting with repeats and heterochromatin, play an important role in the evolution of species boundaries. For example, several of the known "speciation genes" in *Drosophila* associate with satellite DNA and repeats. Hmr HMR and Lhr LHR interact with heterochromatin at centromeres and telomeres, and are needed for transposable element repression (Brideau et al. 2006); Zhr ZHR is a protein that localizes to a chromosome-specific satellite (Ferree and Barbash 2009) and *OdsH* is encodes a heterochromatin-associated protein that binds to the repeat-rich Y chromosome (Bayes and Malik 2009). Additionally, transposable elements have been found to be mis-expressed in hybrids between closely related species, including *Drosophila* (Lopez-Maestre et al. 2017), fish (Dion-Côté et al. 2014), mammals (O'Neill, O'Neill, and Graves 1998), or plants (Wu et al. 2015). Finally, the rapid evolution of centromeric satellite DNA and the centromere-specific histone protein CENP-A has led to the proposal that these two components evolve under genetic conflict, and may result in hybrid incompatibilities (Henikoff, Ahmad, and Malik 2001; Brown and O'Neill 2010). Homologous chromosomes may compete for inclusion in the oocyte, and centromere DNA may act as a selfish element and exploit asymmetric female meiosis to promote transmission to the egg. Co-evolution of CENP-A may restore meiotic parity, but could result in segregation problems in hybrids (Henikoff, Ahmad, and Malik 2001; Brown and O'Neill 2010; Rosin and Mellone 2017). Work in monkeyflowers provides empirical support for the centromere drive hypothesis (Fishman and Saunders 2008). Interspecific monkeyflower hybrids exhibit strong transmission advantage of one parental allele via female meiosis, and divergence of centromere-associated repeats is thought to be responsible for this drive (Fishman and Saunders 2008). Centromere drive has also been detected in mice. Here, selfish centromeres exploit asymmetry of the meiotic spindle and preferentially orient towards the egg pole, thereby achieving preferential transmission into the next generation (Akera et al. 2017; Iwata-Otsubo et al. 2017). A candidate meiotic driver in a centromere-linked region that shows a moderate increase in transmission frequency has also been found in *Drosophila* using a quantitative sequencing approach (Wei et al. 2017). Together, these studies provide empirical support that repetitive DNA can play an important role in the evolution of reproductive isolation. High quality genomes will be necessary to study the impact of heterochromatin and repetitive DNA on the evolution of species boundaries.

Previous studies have obtained conflicting results on the phylogenetic relationships among members of the *nasuta* species group (summarized in Yu et al. 1999). These phylogenies were based on both phenotypic data, such as hybrid sterility (Kitagawa et al. 1982), courtship song (Shao et al. 1997), male frons coloration (Yu et al. 1999), or genetic markers, such as isozymes (Kitagawa et al. 1982), mitochondrial loci (Yu et al. 1999), or a handful of nuclear genes (Bachtrog 2006). Our phylogenomic approach reveals that while phylogenetic relationships vary dramatically across the genome, we find overall strong support for the inferred species trees. Our analysis, using both population genetic and phylogenetic inferences, reveals consistent species groupings. *D. albomicans, D. nasuta,* and *D. kepulauana* form one cluster. These species all show similar male frons coloration (**Figure 4**), and produce viable (though partially sterile) offspring. Another cluster consists of *D. pulaua, D. s. sulfurigaster, D. s. bilimbata, D. s. albostrigata* and *D. s. neonasuta*, and most crosses between these species result in viable hybrids (Kitagawa et al. 1982). *D. s. albostrigata* and *D. s. neonasuta* have been described as different subspecies (Yu et al. 1999 but are genetically indistinguishable in our analysis. Previous studies have typically placed *D. pulaua* as the sister group to the *D. sulfurigaster* semi-species, but our genomic analysis clearly places *D. s albostrigata* and *D. s. neonasuta* as the sister species to *D. s. sulfurigaster*, *D. s. bilimbata,* and *D. pulaua*. These taxa also show differences in their frons colorations: *D. s albostrigata* and *D. s. neonasuta* have thicker frons markings than *D. s. bilimbata* and *D. s. sulfurigaster*, and *D. pulaua* males have very faintly marked frons (**Figure 4**). *D. pallidifrons,* Taxon F and *D. kohkoa* form a distinct cluster*,* and are the sister to the *sulfurigaster* species group, and *D. niveifrons* forms the outgroup to this radiation.

Interestingly, however, signals involved in pre-zygotic isolation (that is, courtship song, mating behavior and male frons coloration) do not always follow the species phylogeny. For example, frons marking on male forehead seems to have evolved convergently in different groups (see **Figures 1 and 4**). The silvery markings on the frons were either present in an ancestor of the *nasuta* species complex, and modified or lost in some species, or gained independently in different clades. *D. pallidifrons*, which is most closely related to *D. kohkoa*, completely lacks silvery markings on its forehead, while *D. kohkoa* males have a continuous silver patch on their frons similar to *D. albomicans* / *D. nasuta*. Interestingly, *D. pallidifrons* is also the only species in this group in which the male never faces the female in his courtship (Spieth 1969), which may suggest that the frons marking and courtship display co-evolved. *D. pulaua*, on the other hand, is very closely related to *D. s. bilimbata* and *D. s. sulfurigaster*, yet its frons are extremely faintly marked, and male courtship song is also drastically different in this species relative to all the *D. sulfurigaster* flies (Nalley and Bachtrog, unpublished). Introgression between lineages, or independent sorting of ancestral variation may be responsible for convergent evolution of signals involved in pre-zygotic isolation.

Intriguingly, we observed a large amount of phylogenetic discordance between trees generated from the autosomes and X chromosome for *D. s. sulfurigaster*, *D. s. bilimbata,* and *D. pulaua*. The autosomes, which make up the majority of the genome, largely supported the grouping of *D. s. bilimbata* and *D. s. sulfurigaster* being sister species, while on the X chromosome, *D. pulaua* and *D. s. sulfurigaster* are more often placed as sister species. Our analysis suggests that the most common topology on the X reflects the true species branching order, and introgression on the autosomes has contributed to the incongruent topologies between the X chromosome and autosomes in this species clade. Lower rates of introgression on the X are expected, since X chromosomes from different species generally have disproportionately large effects on hybrid sterility (the large X-effect; Masly and Presgraves 2007; Presgraves 2018). The large X-effect results from the hemizygous expression of recessive X-linked hybrid sterility factors in XY hybrids and the higher density of hybrid sterility factors on the X relative to the autosomes. Thus, strong selection against hybrid sterility factors would disproportionately eliminate incompatible X-linked variation in species hybrids. Indeed, reduced introgression on the X chromosomes has been reported in multiple systems. For example, hybridizing subspecies of rabbits show elevated levels of differentiation on the X compared to autosomes (Carneiro et al. 2014). Likewise, the X chromosomes of house mouse subspecies is more highly differentiated than the autosomes (Phifer-Rixey, Bomhoff, and Nachman 2014). Interspecific gene flow has also been found to be lower on X chromosomes in various *Drosophila* clades (Phifer-Rixey, Bomhoff, and Nachman 2014; Turissini and Matute 2017; Meiklejohn et al. 2018). Thus, our data support the notion that X chromosomes are less permeable to cross species boundaries. Extensive autosomal introgression between *D. s. bilimbata* and *D. s. sulfurigaster* paradoxically has the effect that most of the trees derived from autosomes do not recover the correct species branching order. This resembles patterns of genomic differentiation between mosquito species (Fontaine et al. 2015). Mosquito species also show discordant X-linked and autosomal phylogenies, with the X chromosome reflecting the species branching order while pervasive autosomal introgression groups non-sister species together (Fontaine et al. 2015).

Materials and Methods

Fly strains

We investigated a total of 67 *nasuta* group fly strains, and one *D. immigrans* strain as an outgroup. **Table S1** gives an overview of the species and strains used, and their geographic location. We chose the inbred *D. albomicans* 15112-1751.03 strain to generate a high-quality genome assembly using PacBio sequencing and Hi-C scaffolding.

PacBio DNA extraction and genome sequencing.

We used a mix of 15112-1751.03 females and extracted high molecular weight DNA using a QIAGEN Gentra Puregene Tissue Kit (Cat #158667). DNA was sequenced on the PacBio RS II platform. In total, this produced 11.6-Gb spanning 531,638 filtered subreads with a mean read length of 12,992-bp.

Chromatin-conformation capture

Hi-C libraries were created from sexed female third instar larvae of *D. albomicans*, adapted from (Stadler, Haines, and Eisen 2017). Single larvae were first homogenized, washed, and fixed with final concentration of 1% formaldehyde for 30 min. Fixed chromatin was then digested overnight with HpyCH4IV at 37°C. The resulting sticky ends were then filled in and marked with biotin-14-dCTP, and dilute blunt end ligation was performed for 4 hours at room temperature. Cross-links were then reversed by incubation at 65°C with Proteinase K. DNA was purified through phenol/chloroform extraction and sheared using a Covaris instrument S220. Biotinylated fragments were enriched using streptavidin beads and subsequent washes. Library preparation (end repair, A-tailing, adapter ligation, library amplification) was performed off the DNA on the streptavidin beads. The final amplified library was purified using Ampure XP beads.

Whole-Genome Re-sequencing of nasuta group flies

We extracted DNA from all flies from Table S1 using either Illumina TruSeq or Nextera libraries. Illumina TruSeq Nano libraries were prepared from 100 ng genomic DNA according to Illumina's protocol for 350-bp inserts. Libraries were pooled and sequenced on a HiSeq 4000 with 100-bp paired-end reads. Nextera libraries were prepared from genomic DNA, following Illumina's protocol with the following modification: reaction volumes were scaled to 10 ng input DNA. Two-sided Ampure XP size selections removed fragments <200-bp and minimized fragments >800-bp. Libraries were pooled and sequenced on a HiSeq 4000 with 100-bp paired-end reads or 150-bp single-end reads.

Genome Assembly and Annotation

The genome assembly was generated as described in (Michael et al. 2018). Briefly, long reads were assembled into contigs using Minimap and Miniasm (Li 2016). This draft assembly was polished three times with RACON (Vaser et al. 2017) and once with Pilon (Walker et al. 2014). Juicer (N. C. Durand et al. 2016) and 3D-DNA (Dudchenko et al. 2017) were used to process Hi-C reads and reorder contigs from the draft assembly based on levels of short range interactions.

Blocks of ordered contigs which showed short-range interactions were stitched together into chromosome level scaffolds. Juicer's bash script was modified to run on our cluster and job scheduling system. 3D-DNA was used with the following options: "-m haploid -t 10000 -s 0 -c 3." We looked at synteny between our scaffolded assembly and a previously published *D. albomicans* genome assembly (Zhou et al. 2012) using MUMmer3 (Kurtz et al. 2004). Scaffolds from our assembly were assigned to Muller elements based on synteny. To confirm that the sex chromosome, Muller A, was correctly assembled, we mapped 20x male and female *D. albomicans* reads with BWA (Li and Durbin 2009) using default options and obtained coverage data for 10-kb windows using bedtools genomecov (Quinlan and Hall 2010) and an in-house Python script. Female coverage was also compared to male/female coverage to identify un-collapsed heterozygosities in our assembly (that is, regions where both haplotypes were assembled independently). Un-collapsed haplotypes can be identified based on reduced genomic coverage (by half; Mahajan et al. 2018), and were removed from our assembly, and resulting gaps in our scaffolds were stitched over. The final genome assembly was annotated using Maker (Campbell et al. 2014). RNA-seq data from adult tissues (male and female head, 3rd instar larvae, carcass; and ovary, spermatheca, accessory glands, and testis) was mapped to the *D. albomicans* genome assembly with HiSat2 version 2.1.0 (Kim, Langmead, and Salzberg 2015) using default parameters and the -dta option. A transcriptome assembly was then generated with the alignments using StringTie version 1.3.3b (Pertea et al. 2015) with default parameters. Finally, fasta sequences of the transcripts were extracted and used as the input for Maker.

SNP Calling and Filtering

Repeat libraries for *D. albomicans* 15112-1751.03 were generated using RepeatModeler version 1.0.5 (Smith and Hubley 2008) and REPdenovo (Chu, Nielsen, and Wu 2016) using default parameters. RepeatModeler was run with default parameters. REPdenovo was run with the following parameters: "MIN_REPEAT_FREQ 3, RANGE_ASM_FREQ_DEC 2, RANGE_ASM_FREQ_GAP 0.8, K_MIN 30, K_MAX 50, K_INC 10, K_DFT 30, READ_LENGTH 100, READ_DEPTH 185.099490, THREADS 20, GENOME_LENGTH 172728670, ASM_NODE_LENGTH_OFFSET -1, MIN_CONTIG_LENGTH 100, IS_DUPLICATE_REPEATS 0.85, COV_DIFF_CUTOFF 0.5, MIN_SUPPORT_PAIRS 20, MIN_FULLY_MAP_RATIO 0.2, TR_SIMILARITY 0.85, and RM_CTN_CUTOFF 0.9". The *D. albomicans* genome was then repeat masked with RepeatMasker version 3.3.0 (Smith, Hubley, and Green 2013) using default parameters. Reads from each fly strain were mapped separately to the *D. albomicans* genome. Read alignment files of strains from the same species were combined. We then call SNPs and indels for each strain using GATK's haplotype caller (DePristo et al. 2011). SNPs were filtered out with the following cutoffs (Gilks et al. 2016): "QD < 2.0", "MQ < 58.0", "FS > 60.0", "SOR > 3.0", "MQRankSum < -7.0", and "ReadPosRankSum < -5.0"—SNPs that fail to meet these thresholds are subsequently masked. These SNPs were used to perform phylogenetic analyses. However, they were pruned using PLINK1.9 (Chang et al. 2015) to minimize the effects of LD in our clustering analyses and demographic inference using the following option: "--indep-pairwise 5kb 50 0.1".

Phylogenetic reconstruction and analysis

To create a phylogeny, we generated pseudo-genomes for each strain by replacing sites on the *D. albomicans* genome assembly with their called SNPs. Sites that are heterozygous and where there is less than 20x coverage were masked, and the reference *D. albomicans* genome was excluded from this analysis, due to reference genome biases. The pseudo-genomes were split into 50-kb bins, and a maximum likelihood (ML) phylogeny was created for each bin using RAxML 8.2.11 (Stamatakis 2014), and a consensus tree was created with ASTRAL-III (C. Zhang et al. 2018). We used FigTree (https://github.com/rambaut/figtree/) to visualize the phylogeny. To test for heterogeneity in evolutionary history across the genome, we randomly selected one representative strain for each species, and calculated topologies in 50-kb or 500-kb windows, as described above. To calculate tree heights in the *sulfurigaster* subgroup, we followed an approach outlined in (Fontaine et al. 2015). We randomly selected (four times) one representative strain for *D. s. sulfurigaster, D. pulaua, D. s. bilimbata,* and *D. pallidifrons* and generated phylogenies using non-overlapping 50-kb windows along the autosomes with RAxML using the same parameters as mentioned above. With the topology, ((a,b),c), we calculated the more shallow divergence time (T_2) using the equation, $\frac{d_{ab}}{2}$, and the more deep divergence time (T_1) using the equation, $\frac{d_{ac}+d_{ab}}{4}$, where d_{ab} is the distance between strains a and b in branch lengths. We used the phytools R package (Revell 2012) to infer the topologies and obtain terminal branch length for each phylogeny.

Divergence time estimates

We used the set of coding sequences (CDS) from the genome annotation to derive Ks (the number of synonymous substitutions per site) values between species. To obtain the coding sequences from non-*D. albomicans* species, we used the corresponding sites from the pseudogenome used to create a phylogeny. Ks values were calculated using *KaKs_Calculator* (Z. Zhang et al. 2006). We used a neutral mutation rate estimate of 3.46×10^{-9} per base per generation, which was experimentally determined from *D. melanogaster* (Keightley et al. 2009a). The species studied here have a generation time that is slightly longer than *D. melanogaster* and we therefore used an intermediate estimate of the number of generations per year for Drosophilids (7 generations; Cutter 2008) to convert the mutation rate to time-based units (2.42×10^{-8} mutations per base per year).

Population genetic analysis

We used Ohana (J. Y. Cheng, Mailund, and Nielsen 2017) with default options to quantify population structure, and calculate admixture proportions between species in the two major clades found in the phylogenetic analysis: the *albomicans* subclade consisting of *D. albomicans, D. nasuta,* and *D. kepulauana* as well as the *sulfurigaster* subclade consisting of *D. s. albostrigata, D. s. neonasuta, D. s. bilimbata, D. s. sulfurigaster,* and *D. pulaua*. FlashPCA (Abraham and Inouye 2014) was used to perform PCA with all strains. To test for introgression in the *sulfurigaster* subgroup, we calculated the G_{min} and ABBA-BABA statistics (E. Y. Durand et al. 2011; Geneva et al. 2015; Martin, Davey, and Jiggins 2015). Aligned reads from *D. s. bilimbata, D. s. sulfurigaster* and *D. pulaua* were processed in 50-kb windows with the POPBAM package (Garrigan 2013), and

G_{min} was calculated using POPBAMTools (https://github.com/geneva/POPBAMTools). We also calculated the D and f_D statistic (Green et al. 2010; Martin, Davey, and Jiggins 2015), to test for introgression between (((*D. s. bilimbata*, *D. s. sulfurigaster*), *D. pulaua*), *D. pallidifrons*). The genome was split into 50-kb windows and a Wilcoxon test was used to determine if the median values are statistically different between X-linked and autosomal windows. We calculated values of average pairwise diversity π along the genome using non-overlapping 50-kb windows. Mean and median values of the entire genome for species with more than one sequenced individual are reported. Software to calculate both the D and f_D statistic as well as π was obtained from (https://github.com/simonhmartin/genomics_general).

Acknowledgments

We thank Masayoshi Watada for providing *nasuta* flies and Carolus Chan for help generating data. This research was supported by NIH grants (R01GM076007, R01GM101255 and R01GM093182) to D. Bachtrog. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant.

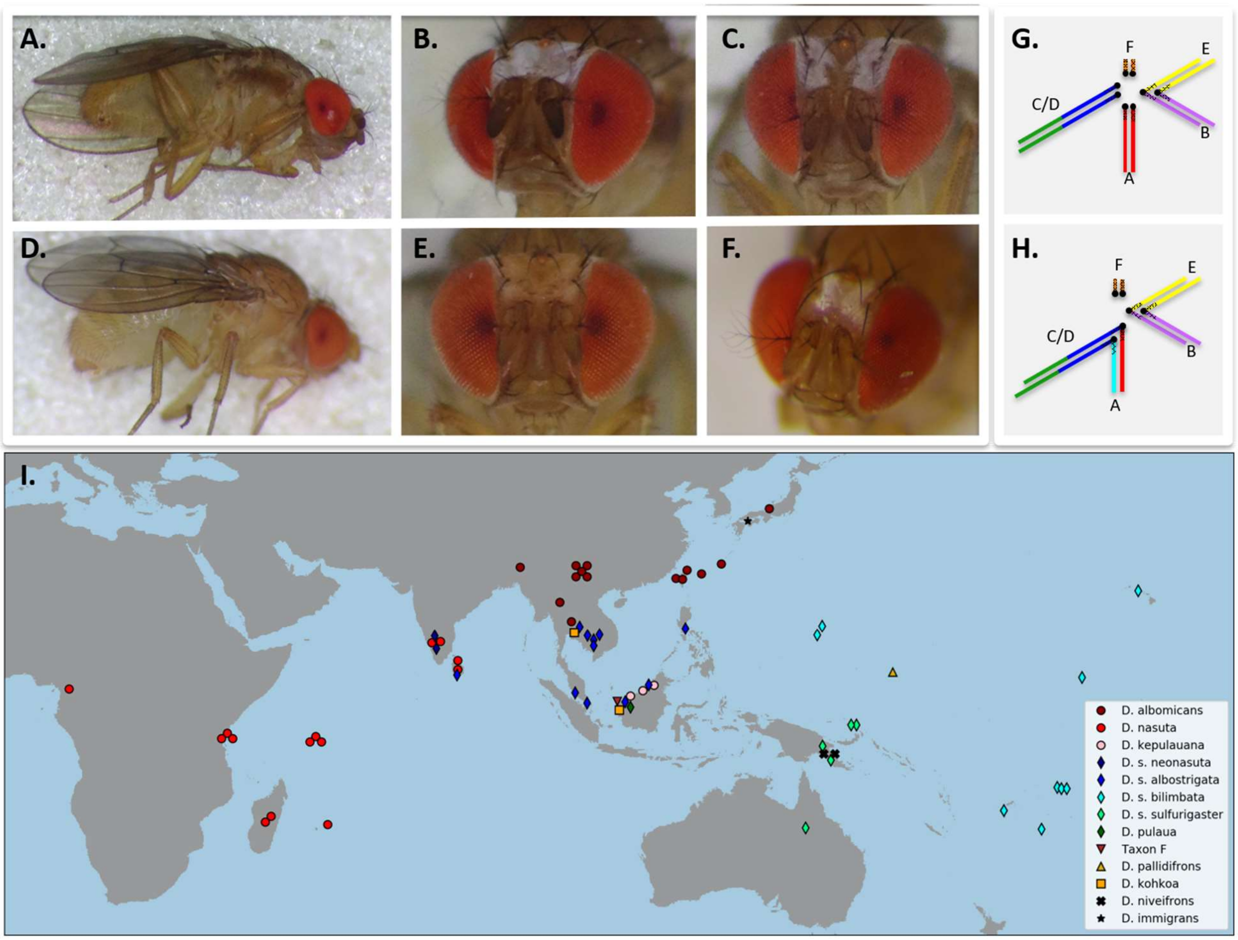
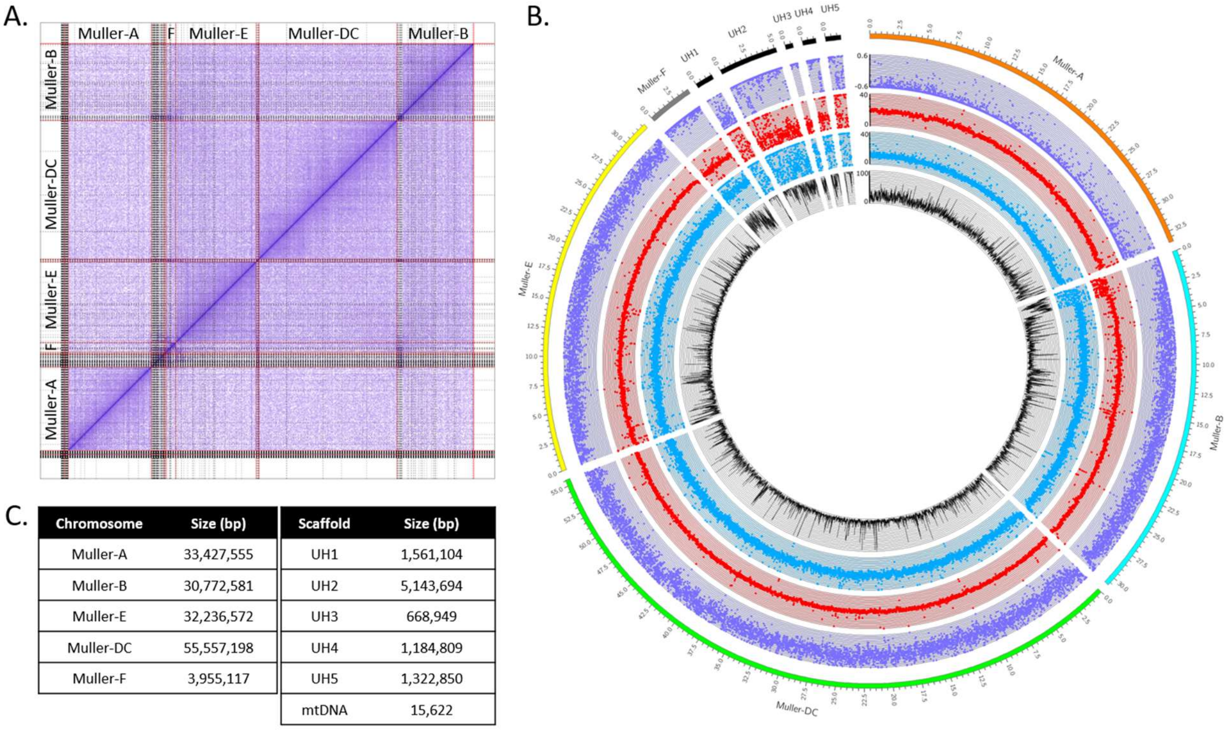


Figure 1. Morphology, karyotype and distribution of species in the *nasuta* subgroup. (A-F.) Male flies of the *nasuta* subgroup differ with regards to their morphology. (A, B.) *D. albomicans*; (C.) *D. s. albostrigata*; (D, E.) *D. pulaua*; (F.) *D. niveifrons*. (G, H.) Karyotypes of members of the *nasuta* group. Muller elements A-F are color-coded. (G.) All species (apart from *D. albomicans*) have a acrocentric X chromosome (Muller A), a metacentric autosome (Muller B/E fusion), and a large acrocentric autosome (Muller C/D fusion), and the small dot chromosome (Muller F) (H.) In *D. albomicans*, a neo-sex chromosome formed by the fusion of Muller C/D to both the X and Y chromosome. (I.) Sampling locations of species and strains investigated. Note that for flies with overlapping sampling locations, the markers where slightly shifted on the map for visualization.



Chromosome	Size (bp)
Muller-A	33,427,555
Muller-B	30,772,581
Muller-E	32,236,572
Muller-DC	55,557,198
Muller-F	3,955,117

Scaffold	Size (bp)
UH1	1,561,104
UH2	5,143,694
UH3	668,949
UH4	1,184,809
UH5	1,322,850
mtDNA	15,622

Figure 2. Assembly of *D. albomicans* genome. A. Hi-C scaffolding of contigs. Gray lines denote PacBio contigs, and red lines indicate different chromosomes. B. Coverage analysis of chromosomes (Muller elements). Genomic reads from *D. albomicans* 15112-1751.03 males and females were mapped to the genome (20x coverage each); each point represents the mean coverage in non-overlapping 10-kb windows (blue: male coverage, red: female coverage, purple: log2(male/female) coverage). The black line shows the mean repeat content (% repeat-masked bp along 10-kb windows). Unmapped scaffolds are highly repeat-rich and presumably correspond to pericentromeric regions. C. Assembled size of the different chromosomes, and of various unmapped scaffolds (UH1-5), and the mitochondrial DNA.

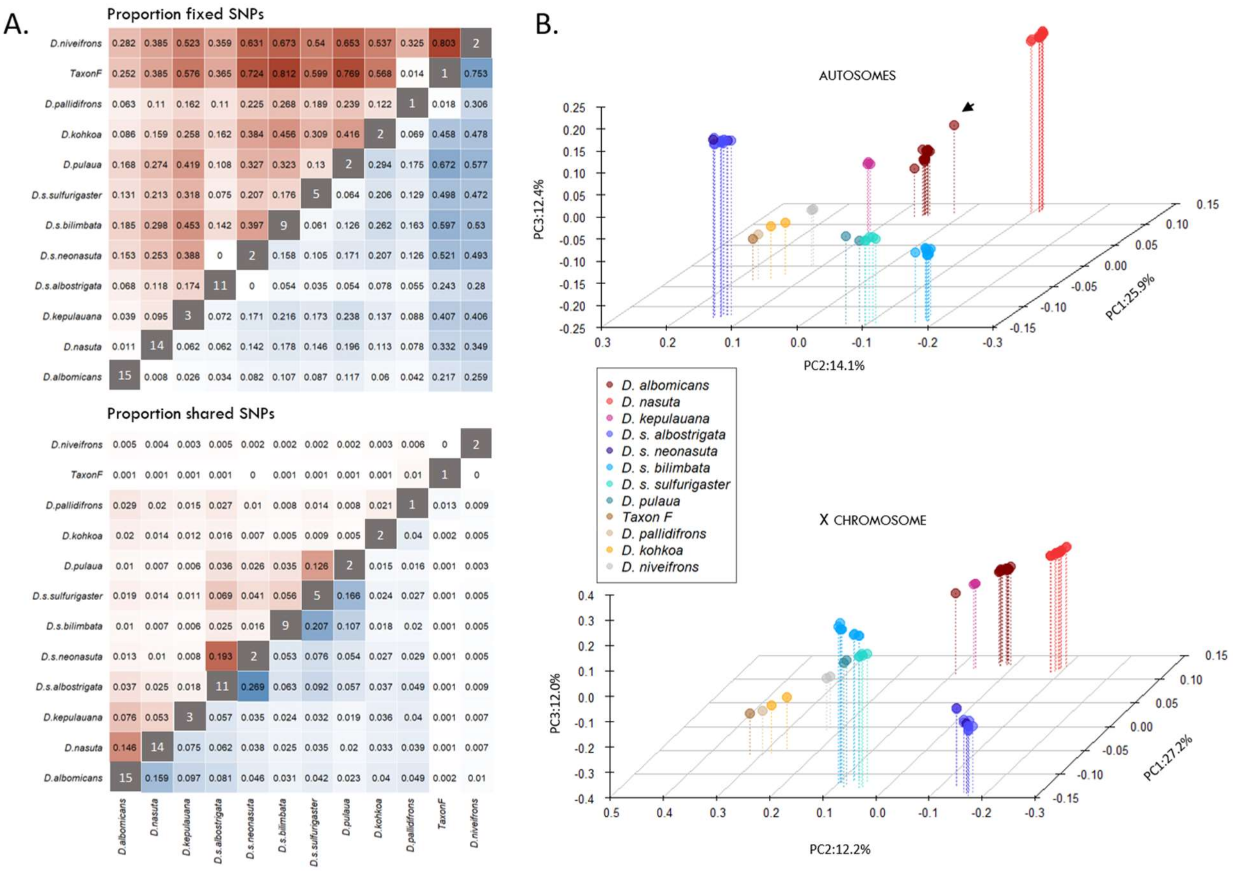


Figure 3. Patterns of genome-wide differentiation in the *nasuta* group. A. Proportion of fixed (top) and shared (bottom) SNPs in the *nasuta* group between the X chromosomes (red) and autosomes (blue). Darker shading indicates larger values. The values in the diagonal indicate the sample size. B. Principle component analysis of autosomal (top) and X-linked (bottom) SNPs in the *nasuta* species group. The black arrow indicates *D. albomicans* E-10815_SHL48.

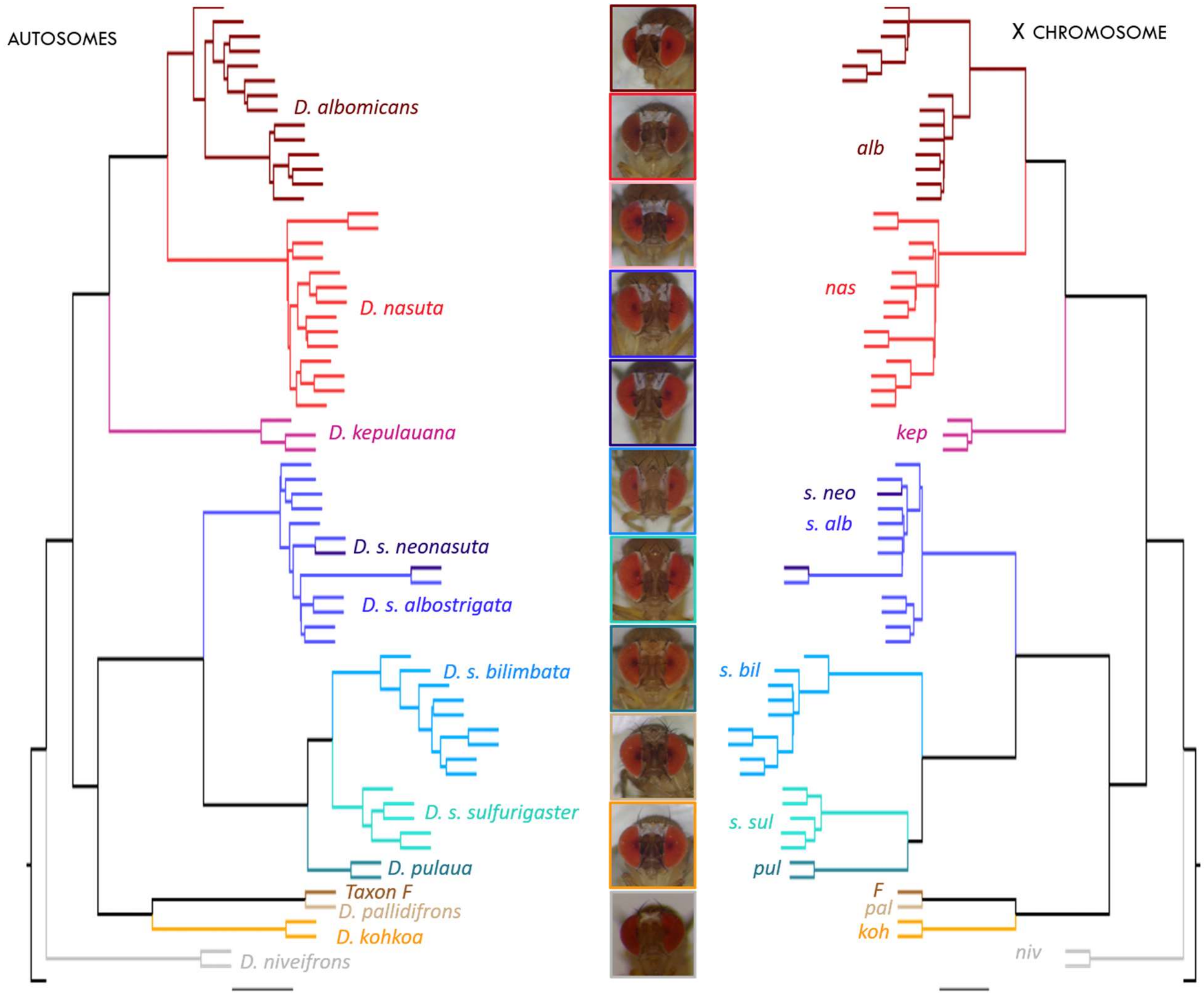


Figure 4. Phylogenetic relationships among species of the *nasuta* group. The autosome phylogeny (left) has the same species level topology as the X chromosome phylogeny (right) with the exception of *D. s. bilimbata, D. s. sulfurigaster,* and *D. pulaua*. The colored lines correspond to all the strains that belong to the same species group.

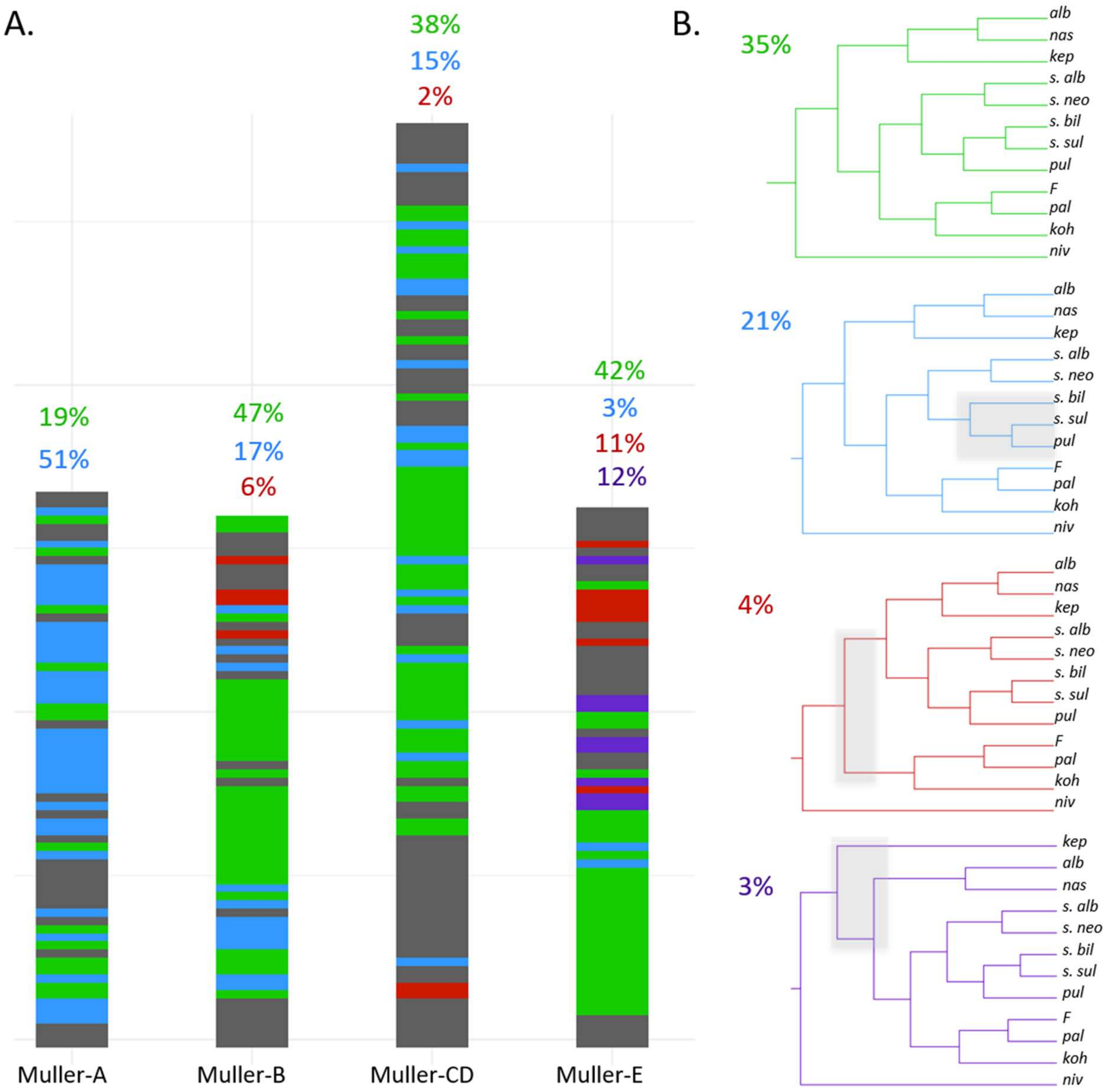


Figure 5. Local evolutionary history in the *nasuta* group varies across the genome. A. Tree topology across the genome. For each 500-kb window, we color-code the topology recovered from that region (colors correspond to topologies in B). Note that while tree 1 (green) dominates on the autosomes, tree 2 (blue) dominates on the X. Coordinates are in terms of *D. albomicans* genome. Grey regions show alternative topologies. B. Common topologies. The four most common trees are shown. The value in the top left corner is the percentage of all 500-kb windows that recovers that topology.

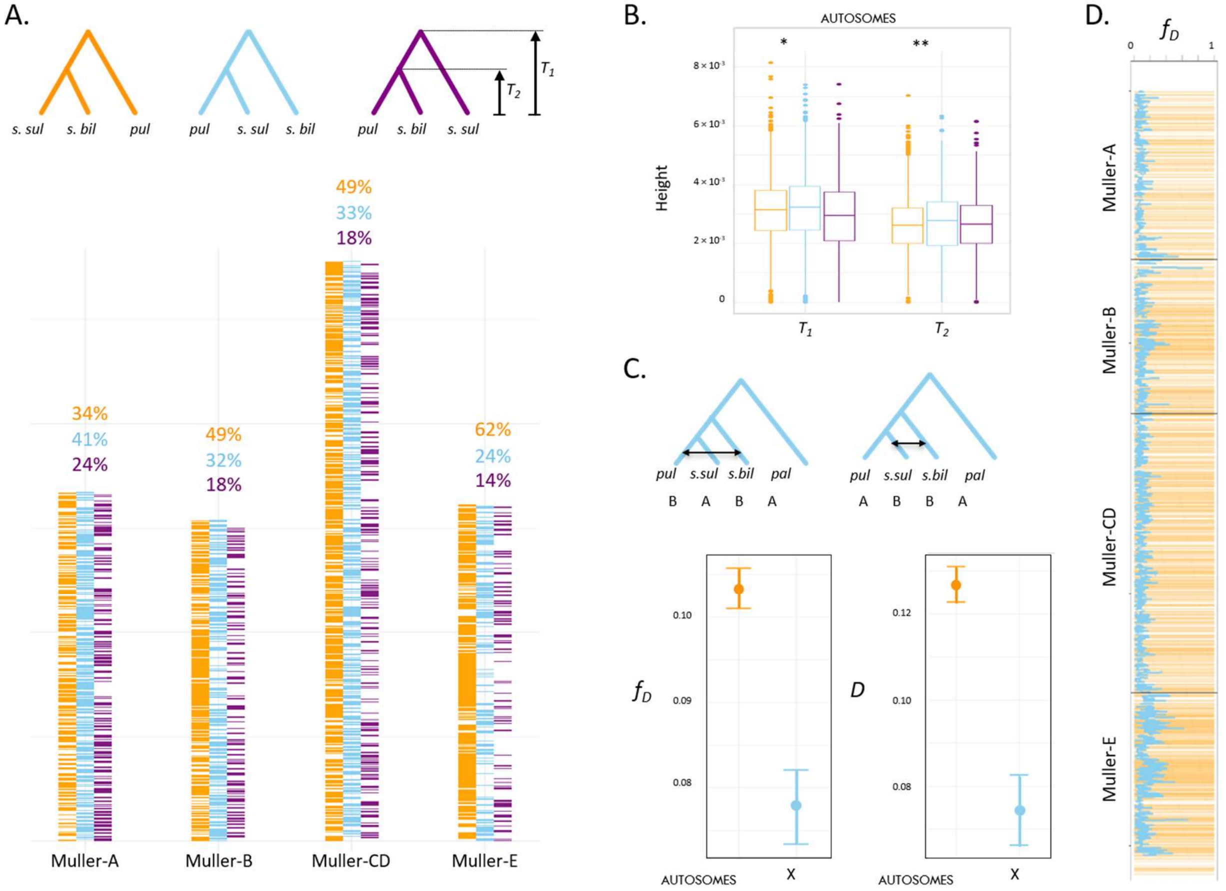


Figure 6. Autosomal introgression in the *sulfurigaster* clade. A. Tree topology across the genome for *D. s. bilimbata, D. s. sulfurigaster*, and *D. pulaua* in 50-kb windows. Topologies are color-coded, and tree heights T_1 and T_2 are indicated. B. Tree height suggests that the majority X chromosome topology is the true phylogeny. T_1 and T_2 are shown for autosomal trees inferred from 50-kb windows. Trees with X majority relationship (*s.bil*, (*s.sul*, *pul*)) have significantly higher T_1 and T_2 than (*s.bil*, *s.sul*), *pul*) trees (p=0.0037 and p=2.55 x 10^{-11}, Wilcoxon test), which is consistent with widespread introgression on the autosomes. C. ABBA-BABA statistics (D and f_D) to test for introgression between *D. s. bilimbata*, *D. s. sulfurigaster* and *D. pulaua* on the autosomes and the X chromosome (vertical bar shows the SE). Both test statistics are higher on the autosome compared to the X (D: p= 1.47x 10^{-9}; f_D: p= 8.61 x 10^{-11}; Wilcoxon test). D. Genomic regions that show the autosome majority topology ((*s.bil*, *s.sul*), *pul*) show higher levels of introgression (as measured by f_D; p< 2.2x 10^{-16}; Wilcoxon test). Shown is the autosomal tree topology across the genome (in yellow, as in panel A) across the genome and f_D (blue line) in 50-kb windows.

Table S1. *Drosophila* species and strains utilized.

Species	**Strain**	**Obtained From**	**Collection Site**
D. albomicans	15112-1751.00	UCSD Stock Center	Okinawa, Japan
D. albomicans	15112-1751.01	UCSD Stock Center	Alisha, Taiwan
D. albomicans	15112-1751.02	UCSD Stock Center	Penghu Islands
D. albomicans	15112-1751.03	UCSD Stock Center	Nankang, Taiwan
D. albomicans	15112-1751.05	UCSD Stock Center	Ishigaki Island, Japan
D. albomicans	15112-1751.07	UCSD Stock Center	Chiang Dao, Thailand
D. albomicans	15112-1751.08	UCSD Stock Center	Chakkarat, Thailand
D. albomicans	FKC20	EHIME Stock Center	Fukui, Japan
D. albomicans	KM070	Qi Zhou	Kunming, China
D. albomicans	KM126	Qi Zhou	Kunming, China
D. albomicans	KM134	Qi Zhou	Kunming, China
D. albomicans	KM165	Qi Zhou	Kunming, China
D. albomicans	KM55	Qi Zhou	Kunming, China
D. albomicans	SHL	EHIME Stock Center	Shillong, India
D. albomicans	NOU98	Masayoshi Watada	Noumea, New Caledonia
D. immigrans	15111-1731.13	UCSD Stock Center	Ehime, Japan
D. kepulauana	15112-1761.01	UCSD Stock Center	Sarawak, Malaysia
D. kepulauana	15112-1761.02	UCSD Stock Center	Brunei, Borneo
D. kepulauana	15112-1761.03	UCSD Stock Center	Ulu Temburong, Borneo
D. kohkoa	15112-1771.00	UCSD Stock Center	Chakkarat, Thailand
D. kohkoa	15112-1771.01	UCSD Stock Center	Sarawak, Malaysia
D. nasuta	15112-1781.00	UCSD Stock Center	Mysore, India
D. nasuta	15112-1781.01	UCSD Stock Center	Seychelle Isles, France
D. nasuta	15112-1781.02	UCSD Stock Center	Seychelle Isles, France
D. nasuta	15112-1781.06	UCSD Stock Center	Mombasa Kenya
D. nasuta	15112-1781.07	UCSD Stock Center	Mombasa Kenya
D. nasuta	15112-1781.08	UCSD Stock Center	Antanarivo, Madagascar
D. nasuta	15112-1781.09	UCSD Stock Center	Antanarivo, Madagascar

D. nasuta	15112-1781.11	UCSD Stock Center	Sri Lanka
D. nasuta	15112-1781.12	UCSD Stock Center	Kandy, Sri Lanka
D. nasuta	15112-1781.13	UCSD Stock Center	Yaounde, Cameroon
D. nasuta	E-19502_MBA31	EHIME Stock Center	Mombasa Kenya
D. nasuta	E-19503_G86	EHIME Stock Center	Mauritius
D. nasuta	E-19504_NHO4	EHIME Stock Center	Nagarahole, India
D. nasuta	E-19505_SEZ11	EHIME Stock Center	Seychelle Isles, France
D. niveifrons	LAE276	Masayoshi Watada	Lae, Papua New Guinea
D. niveifrons	LAE221	Masayoshi Watada	Lae, Papua New Guinea
Taxon F	B208	Masayoshi Watada	Kunching,Sarawak,Malaysia
D. pallidifrons	PN175_E-19901	Masayoshi Watada	Ponape Micronesia
D. pulaua	O-30	Masayoshi Watada	Lae, Papua New Guinea
D. pulaua	15112.1801.00	UCSD Stock Center	Sarawak, Malaysia
D. s. albostrigata	15112-1771.04	UCSD Stock Center	Rizal, Phillipines
D. s. albostrigata	cambodia_1	Doris Bachtrog	Cambodia
D. s. albostrigata	cambodia_3	Doris Bachtrog	Cambodia
D. s. albostrigata	15112-1811.00	UCSD Stock Center	Kuala Lumpur, Malaysia
D. s. albostrigata	15112-1811.01	UCSD Stock Center	Akreiy Ksatr Commune
D. s. albostrigata	15112-1811.02	UCSD Stock Center	Chakkarat, Thailand
D. s. albostrigata	15112-1811.03	UCSD Stock Center	Sarawak, Malaysia
D. s. albostrigata	15112-1811.04	UCSD Stock Center	Siem Reap, Cambodia
D. s. albostrigata	15112-1811.05	UCSD Stock Center	Brunei, Borneo
D. s. albostrigata	15112-1811.07	UCSD Stock Center	Singapore
D. s. albostrigata	15112-1811.08	UCSD Stock Center	Kandy, Sri Lanka
D. s. bilimbata	15112-1821.00	UCSD Stock Center	Oahu, Hawai'i
D. s. bilimbata	15112-1821.02	UCSD Stock Center	Palmyra Islands
D. s. bilimbata	15112-1821.03	UCSD Stock Center	Savai'i, Samoa
D. s. bilimbata	15112-1821.04	UCSD Stock Center	Upolu, Samoa
D. s. bilimbata	15112-1821.05	UCSD Stock Center	Tongatapu, Tonga Islands
D. s. bilimbata	15112-1821.06	UCSD Stock Center	Viti Levu, Fiji

D. s. bilimbata	15112-1821.08	UCSD Stock Center	Guam, Mariana Islands
D. s. bilimbata	15112-1821.09	UCSD Stock Center	Upolu, Samoa
D. s. bilimbata	15112-1821.10	UCSD Stock Center	Guam
D. s. neonasuta	15114-1861.00	UCSD Stock Center	Mysore, India
D. s. neonasuta	E-20702_CJB53	EHIME Stock Center	Coinbatore, India
D. s. sulfurigaster	WAU-18	Masayoshi Watada	WAU, Papua New Guinea
D. s. sulfurigaster	15112-1831.00	UCSD Stock Center	Queensland, Australia
D. s. sulfurigaster	15112-1831.01	UCSD Stock Center	Kavieng, New Ireland
D. s. sulfurigaster	15112-1831.02	UCSD Stock Center	Wau, Papua New Guinea
D. s. sulfurigaster	15112-1831.04	UCSD Stock Center	Kavieng, New Ireland

Table S2. The number of BUSCOs found in the *D. albomicans* genome assembly. Nearly all complete, single copy genes were found.

Busco Stats		
Complete	1046	98.1 %
Complete and Single Copy	1037	97.3 %
Complete and Duplicated	9	0.8 %
Fragmented	2	0.2 %
Missing	18	1.7 %
Total Searched	1066	100 %

Table S3. Mean and median levels of average pairwise diversity (π) in the different species.

Species	median π (%)	mean π (%)
D. albomicans	0.61	0.60
D. nasuta	0.49	0.47
D. kepulauana	0.58	0.57
D. s. albostrigata	0.52	0.51
D. s. neonasuta	0.49	0.48
D. s. bilimbata	0.18	0.18
D. s. sulfurigaster	0.40	0.39
D. pulaua	0.31	0.33
D. kohkoa	0.44	0.44
D. niveifrons	0.29	0.30

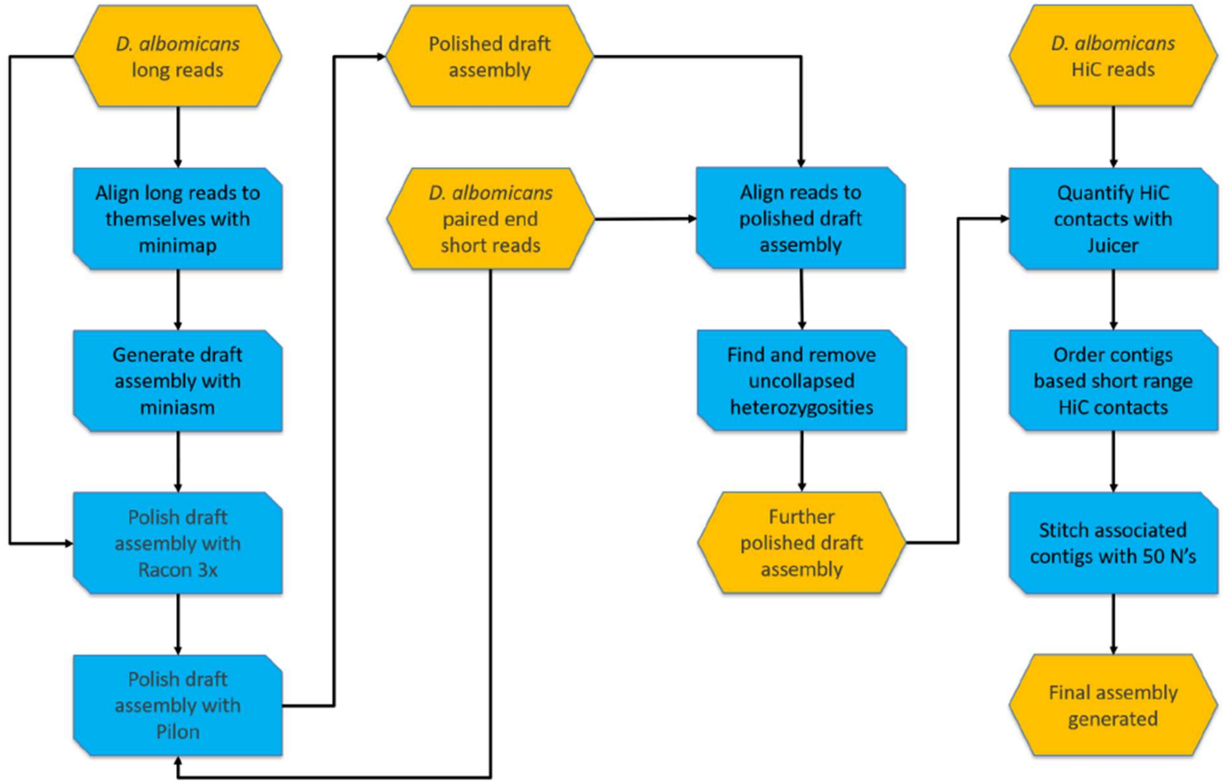
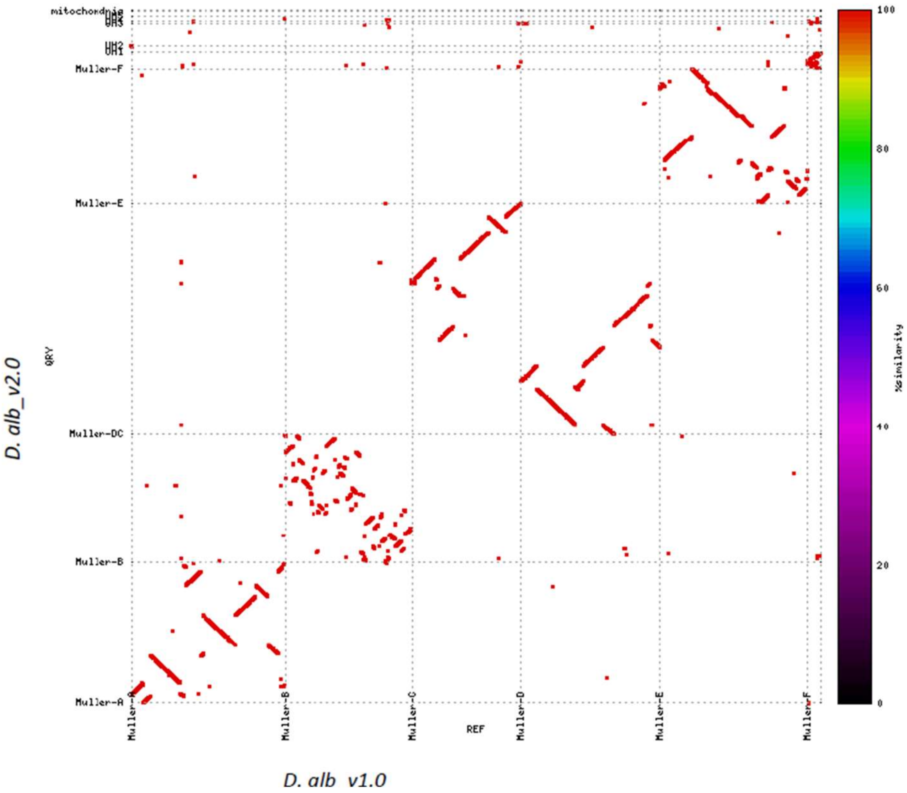

D. albomicans long reads
Align long reads to themselves with minimap
Generate draft assembly with miniasm
Polish draft assembly with Racon 3x
Polish draft assembly with Pilon
Polished draft assembly
D. albomicans paired end short reads
Align reads to polished draft assembly
Find and remove uncollapsed heterozygosities
Further polished draft assembly
D. albomicans HiC reads
Quantify HiC contacts with Juicer
Order contigs based short range HiC contacts
Stitch associated contigs with 50 N's
Final assembly generated

Figure S1. Assembly pipeline. The yellow hexagons represent data and the blue rectangles aree the computational steps performed to obtain a high-quality genome assembly of *D. albomicans.*



D. alb_v1.0

N50= 34,018
MAX SCAFFOLD SIZE= 268,929
GENOME SIZE= 149,497,550
NUMBER OF SCAFFOLDS= 10,263
GC CONTENT= 38.06%

D. alb_v2.0

N50= 33,427,555
MAX SCAFFOLD SIZE= 55,557,198
GENOME SIZE= 165,846,051
NUMBER OF SCAFFOLDS= 11
GC CONTENT= 38.15%

Figure S2. Comparison to previous assembly of *D. albomicans* (D.alb_v1.0) to current assembly (D.alb_v2.0). D.alb_v1.0 was based on an Illumina assembly, and scaffolding of contigs using the *D. virilis* genome as a reference.

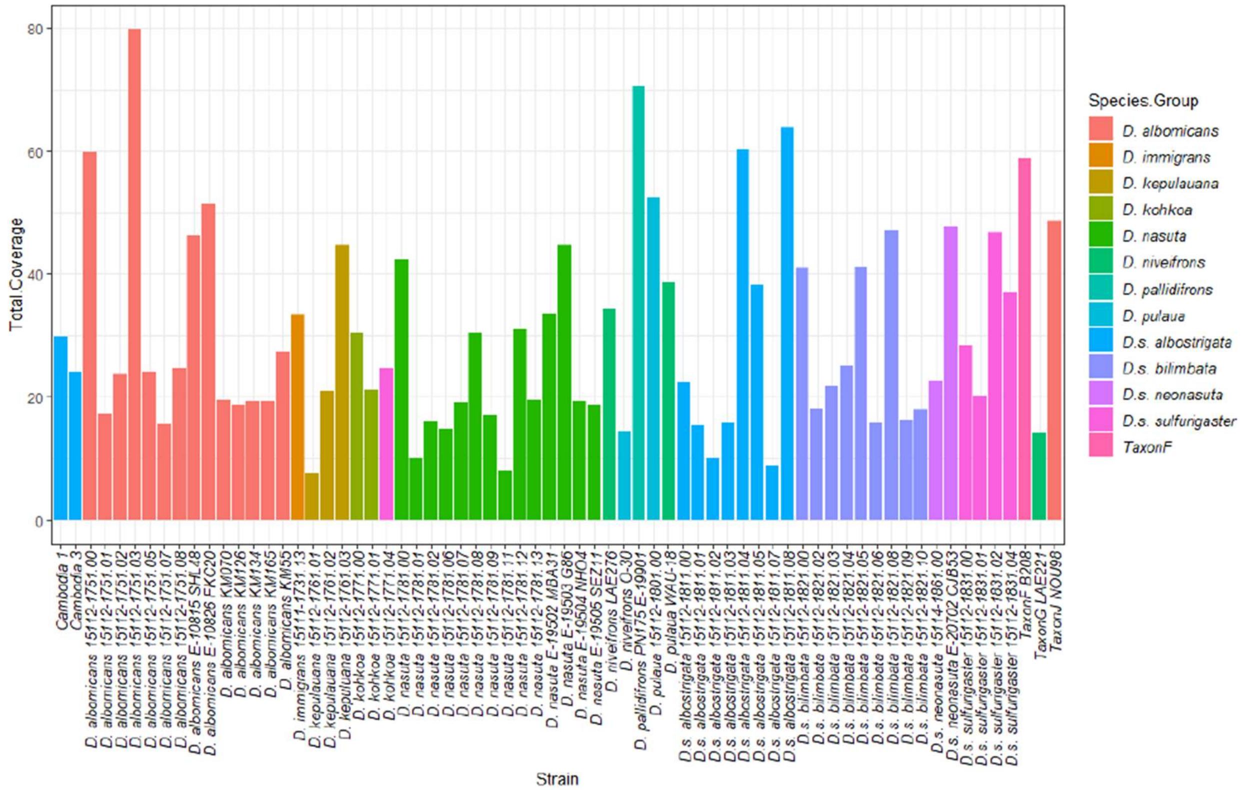


Figure S3. Sequencing read depth for each line investigated. Coverage ranges from 7.5x to 79.7x among strains.

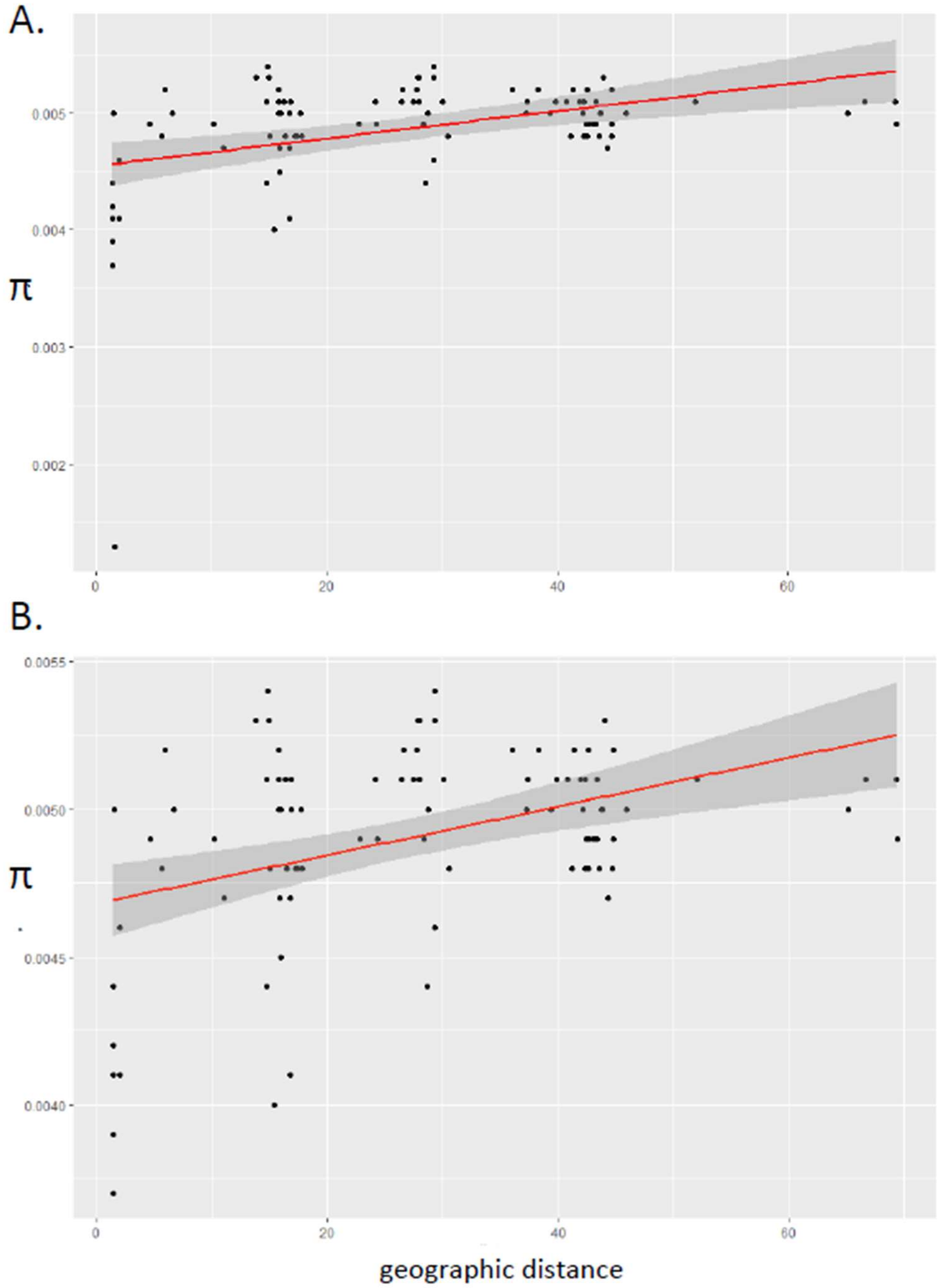
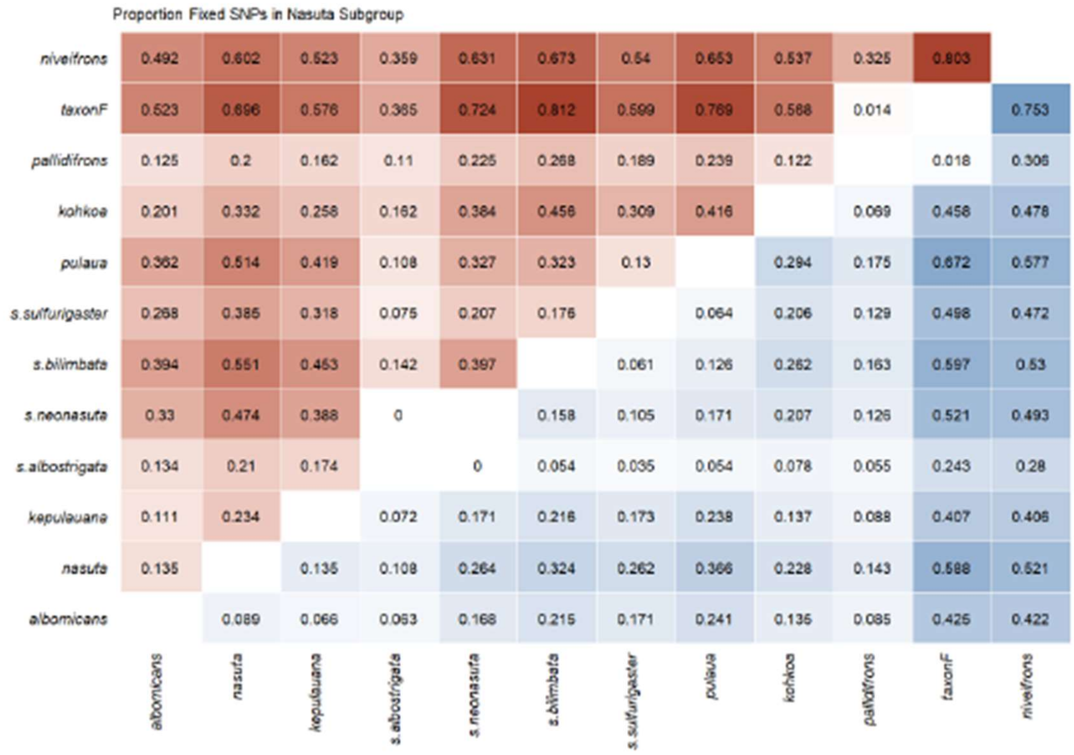
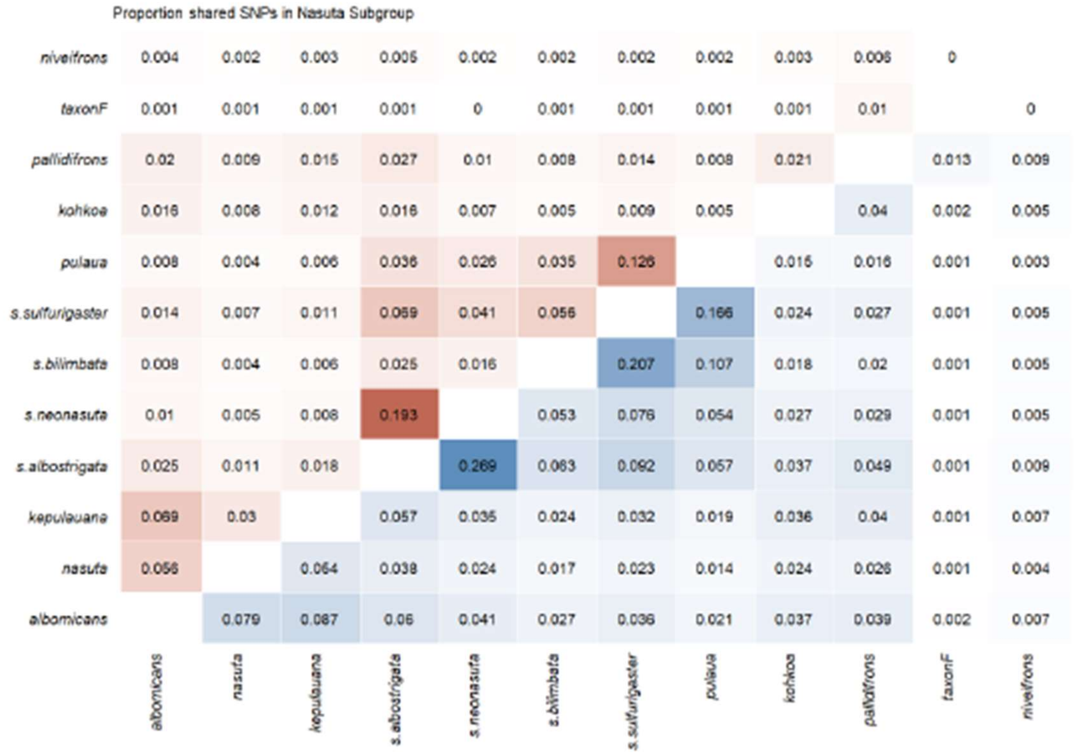


Figure S4. Mean pairwise diversity versus geographic distance for *D. nasuta.* A. Shown is mean π in 50-kb sliding windows versus geographic distance for all pairwise *D. nasuta*strain comparisons. Geographic distance is the Euclidean distance between GPS coordinates. Genetic distance correlates with geographic distance (adjusted R-squared = 0.1393; p-value < 0.0001587). **B.** Same as A, but outliers removed (adjusted R-squared = 0.1524; p-value < 8.374 x 10^{-5}).



Proportion Fixed SNPs in Nasuta Subgroup

	albomicans	nasuta	kepulauana	s.albostrigata	s.neonasuta	s.bilimbata	s.sulfurigaster	pulaua	kohkoa	pallidifrons	taxonF	niveifrons
niveifrons	0.492	0.602	0.523	0.359	0.631	0.673	0.54	0.653	0.537	0.325	0.803	
taxonF	0.523	0.696	0.576	0.365	0.724	0.812	0.599	0.769	0.568	0.014		0.753
pallidifrons	0.125	0.2	0.162	0.11	0.225	0.268	0.189	0.239	0.122		0.018	0.306
kohkoa	0.201	0.332	0.258	0.162	0.384	0.456	0.309	0.416		0.069	0.458	0.478
pulaua	0.362	0.514	0.419	0.108	0.327	0.323	0.13		0.294	0.175	0.672	0.577
s.sulfurigaster	0.268	0.385	0.318	0.075	0.207	0.176		0.064	0.206	0.129	0.498	0.472
s.bilimbata	0.394	0.551	0.453	0.142	0.397		0.061	0.126	0.262	0.163	0.597	0.53
s.neonasuta	0.33	0.474	0.388	0		0.158	0.105	0.171	0.207	0.126	0.521	0.493
s.albostrigata	0.134	0.21	0.174		0	0.054	0.035	0.054	0.078	0.055	0.243	0.28
kepulauana	0.111	0.234		0.072	0.171	0.216	0.173	0.238	0.137	0.088	0.407	0.406
nasuta	0.135		0.135	0.108	0.264	0.324	0.262	0.366	0.228	0.143	0.588	0.521
albomicans		0.089	0.066	0.063	0.168	0.215	0.171	0.241	0.136	0.085	0.425	0.422



Proportion shared SNPs in Nasuta Subgroup

	albomicans	nasuta	kepulauana	s.albostrigata	s.neonasuta	s.bilimbata	s.sulfurigaster	pulaua	kohkoa	pallidifrons	taxonF	niveifrons
niveifrons	0.004	0.002	0.003	0.005	0.002	0.002	0.002	0.002	0.003	0.006	0	
taxonF	0.001	0.001	0.001	0.001	0	0.001	0.001	0.001	0.001	0.01		0
pallidifrons	0.02	0.009	0.015	0.027	0.01	0.008	0.014	0.008	0.021		0.013	0.009
kohkoa	0.016	0.008	0.012	0.016	0.007	0.005	0.009	0.005		0.04	0.002	0.005
pulaua	0.008	0.004	0.006	0.036	0.026	0.035	0.126		0.016	0.016	0.001	0.003
s.sulfurigaster	0.014	0.007	0.011	0.069	0.041	0.056		0.166	0.024	0.027	0.001	0.005
s.bilimbata	0.008	0.004	0.006	0.025	0.016		0.207	0.107	0.018	0.02	0.001	0.005
s.neonasuta	0.01	0.005	0.008	0.193		0.053	0.076	0.054	0.027	0.029	0.001	0.005
s.albostrigata	0.025	0.011	0.018		0.269	0.063	0.092	0.057	0.037	0.049	0.001	0.009
kepulauana	0.069	0.03		0.057	0.035	0.024	0.032	0.019	0.036	0.04	0.001	0.007
nasuta	0.056		0.064	0.038	0.024	0.017	0.023	0.014	0.024	0.026	0.001	0.004
albomicans		0.079	0.087	0.06	0.041	0.027	0.036	0.021	0.037	0.039	0.002	0.007

Figure S5. Patterns of genome-wide differentiation in the *nasuta* group. Proportion of fixed (top) and shared (bottom) SNPs in the *nasuta* group between the X chromosomes (red) and autosomes (blue). Darker shading indicates larger values. Flies are down-sampled to two individuals in *D. nasuta* and *D. albomicans.*

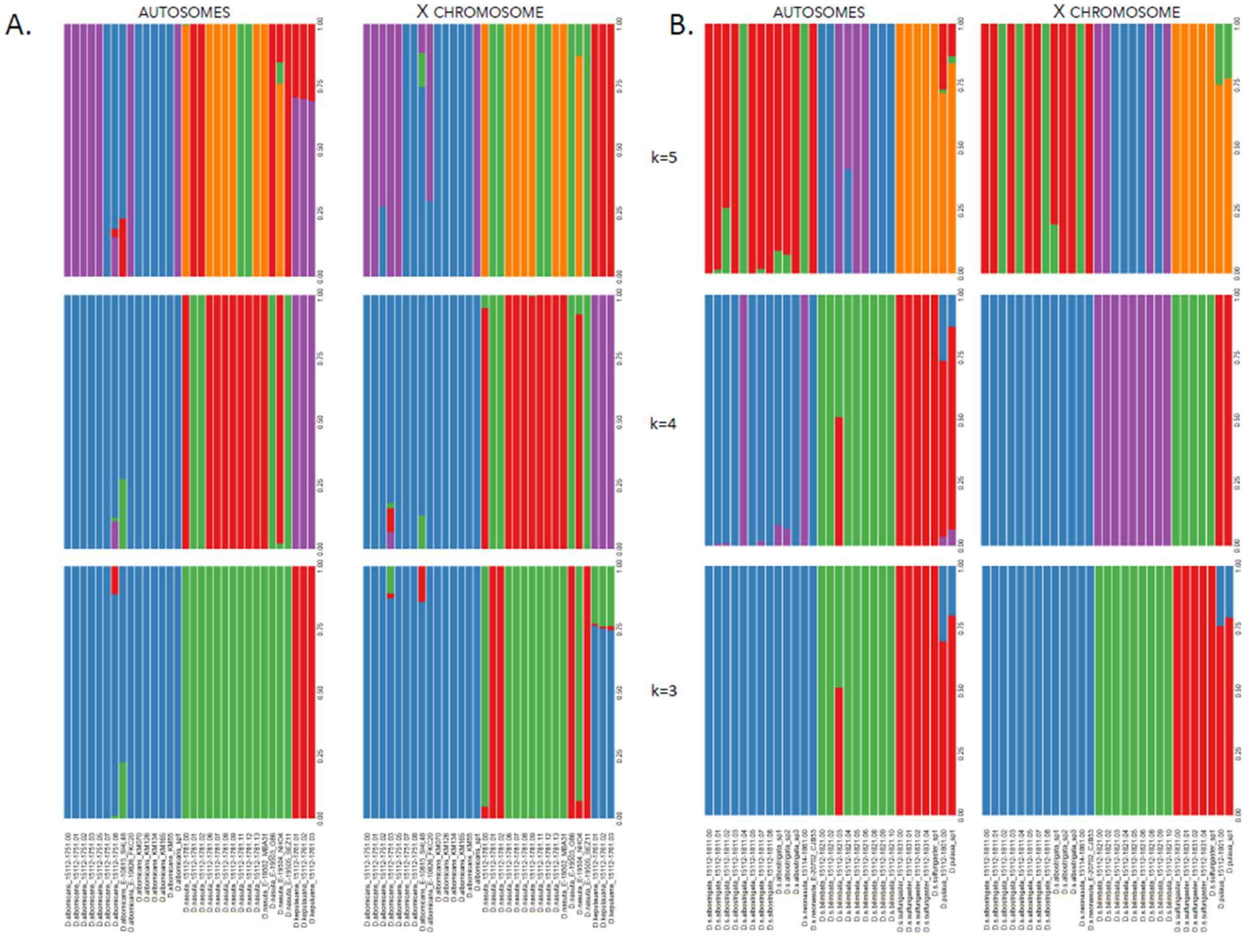


Figure S6. Structure analysis. Structure bar plots representing k=3, k=4, and k=5 populations for **A.** *D. albomicans, D. nasuta, and D. kepulauana* and **B.** for *D. s. albostrigata, D. s. neonasuta, D. s. sulfurigaster,* and *D. pulaua.*

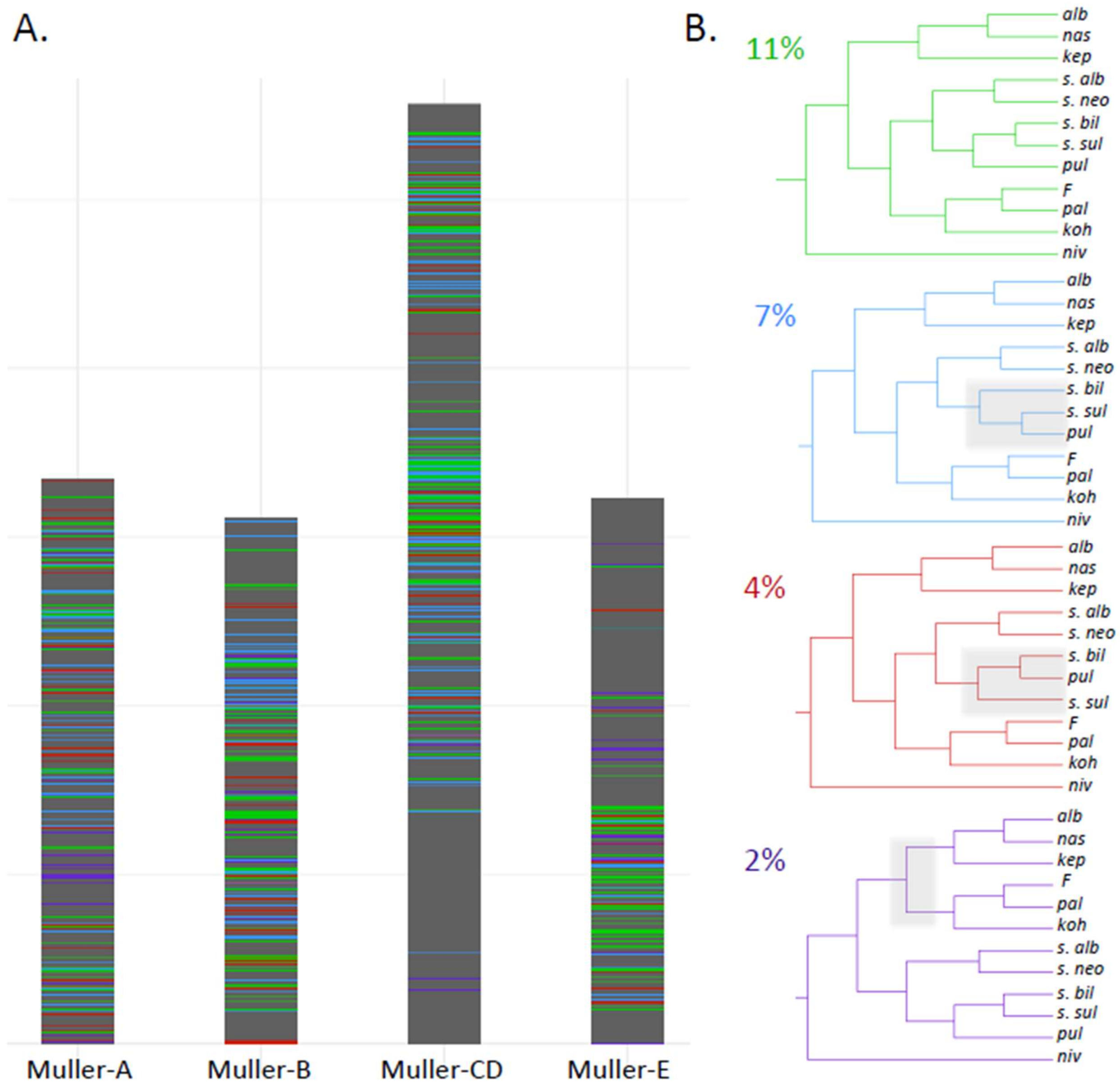


Figure S7. Local evolutionary history in the *D. nasuta* group varies across the genome. A. Tree topology across the genome. For each 50-kb window, we color-code the topology recovered from that region (colors correspond to topologies in B). Note that while tree 1 (green) dominates on the autosomes, tree 2 (blue) dominates on the X. Coordinates are in terms of *D. albomicans*genome. Greyy regions show alternative topologies. **B.** Common topologies. The four most common trees are shown. The value in the top left corner is the percentage of all 50-kb windows that recovers that topology.

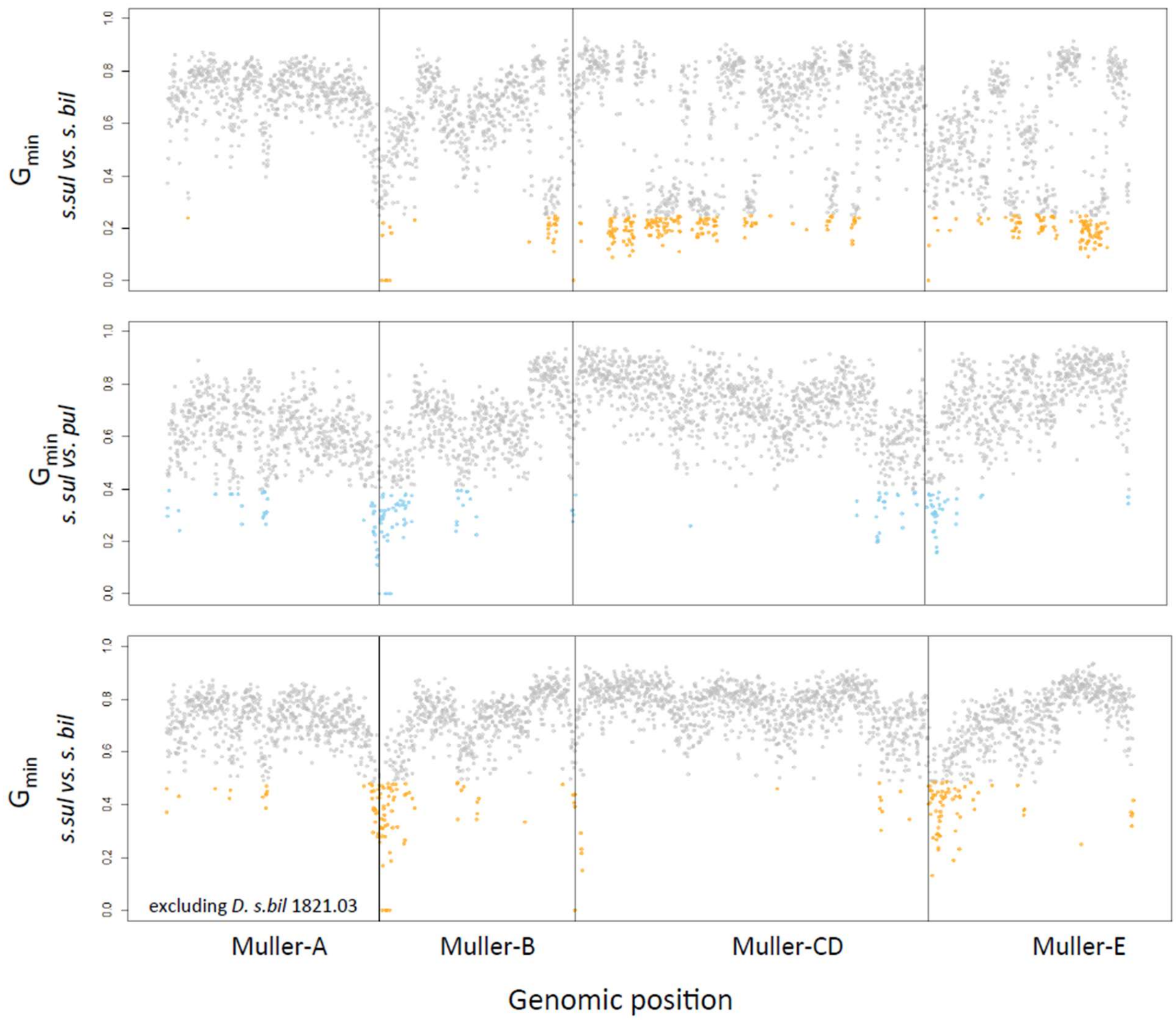


Figure S8. G_{min} across the genome. Shown is the G_{min} statistic across the genome (in 50-kb windows) between *D. s. bilimbata* and *D. s. sulfurgiaster* (top); *D. s. sulfurigaster*and *D. pulaua* (middle) and between *D. s. bilimbata* (but excluding strain D. s. bil 1821.03) and *D. s. sulfurigaster* (bottom). Significant windows are color-coded.

Chapter 2: Molecular characterization of inversion breakpoints in the *Drosophila nasuta* species group

Dat Mai, Doris Bachtrog

Department of Integrative Biology, University of California Berkeley, Berkeley, California, United States of America.

Chromosomal inversions are fundamental drivers of genome evolution. In the *Drosophila* genus, inversions have been widely characterized cytologically, and may play an important role in local adaptation. Here, we characterize chromosomal inversions in the *Drosophila nasuta* species group using chromosome-level, reference-quality assemblies of seven species and subspecies in this clade. Reconstruction of ancestral karyotypes allowed us to infer the order in which the 22 identified inversions occurred along the phylogeny. We found a higher rate of inversions on the X chromosome, and heterogeneity in the rate of accumulation across the phylogeny. We molecularly characterize the breakpoints of six autosomal inversions, and found that repeated sequences are associated with inversion breakpoints in four of these inversions, suggesting that ectopic recombination is an important mechanism in generating inversion. Characterization of inversions in this species group provides a foundation for future population genetic and functional studies in this recently diverged species group.

Introduction

Inversion polymorphisms have been studied extensively in *Drosophila* genetics since their first discovery over a century ago (Sturtevant 1917). Chromosomal inversions were first identified as a suppressors of recombination in *Drosophila melanogaster* (Sturtevant 1917), and characterized subsequently in detail as structural alterations in polytene chromosomes across the *Drosophila* genus (Sperlich and Pfreim 1986; Krimbas and Powell 1992).

Over the past century, chromosomal inversions have been recognized as a ubiquitous evolutionary phenomenon. Inversions are present in virtually all species and can have wide-ranging evolutionary effects. Inversions can help maintain coadapted gene complexes, reduce gene flow in hybrid zones, or restrict recombination between diverging sex chromosomes (Hoffmann and Rieseberg 2008). In addition to modifying the recombination landscape along a chromosome, inversions can also directly alter the structure or expression of genes found near inversion breakpoints (Calvete et al. 2012; Guillén and Ruiz 2012).

Despite being ubiquitous in nature and their putative widespread consequences, the evolutionary forces maintaining inversions are typically poorly understood. Several lines of evidence suggest that many inversions found in Drosophila and other species are adaptive. In particular, inversions often show seasonal, altitudinal and/or latitudinal clines, and polymorphic inversions are often associated with fitness-related traits (Theodosius Dobzhansky 1944; T. Dobzhansky 1948; Hoffmann and Rieseberg 2008; Hoffmann, Sgrò, and Weeks 2004).

Genome-wide alignments between species allow us not only to detect the presence of chromosomal inversions but also to identify and characterize inversion breakpoint regions (Feuk et al. 2005; Richards et al. 2005; Ranz et al. 2007; Corbett-Detig, Cardeno, and Langley 2012; Fuller et al. 2017). Breakpoint sequences may shed light on the causes generating the inversion as well as on the functional consequences that the inversion might have had.

Here we identify chromosomal inversions and characterize their breakpoints in the *D. nasuta* subgroup. This species group contains about a dozen species that are distributed across South-East Asia. The karyotype of *D. nasuta* species consists of the X (Muller A), a large metacentric autosome (chromosome 2; Muller B, E) and a large acrocentric autosome (chromosome 3; Muller C, D), and the small dot chromosome. In *D. albomicans*, chromosome 3 fused to both the X and the Y chromosome, forming a neo-sex chromosome. Inversion polymorphism of the nasuta species group has been studied using cytogenetic techniques (Lambert 1982; Pope 1987; Casu 1990), and species in this group were found to be highly polymorphic for chromosomal inversions. However, no systematic characterization of inversions at the molecular level exists. Here we take advantage of high-quality chromosome-level genome assemblies for molecular characterization of inversions in the *D. nasuta* subgroup.

Results

Karyotype evolution in the D. nasuta subgroup

Ancestral linkage groups are conserved across the Drosophila genus and termed Muller elements (Muller 1940). Most flies in the *D. nasuta* subgroup have a conserved karyotype with an acrocentric X (Muller A), a large metacentric autosome (Muller B and Muller E) and an acrocentric autosome (Muller C and Muller D), and the small dot chromosome (Muller F). In *D. albomicans*, the acrocentric autosome fused to both the ancestral X and Y, forming a neo-X and neo-Y chromosome. Our high-quality assemblies recovered each chromosome arm as a single contig (Mai, Nalley, and Bachtrog 2020). **Figure 1** gives an overview of global syntenic relationships across the species investigated, based on the location of protein-coding genes; genes are assigned to Muller elements (and color-coded accordingly). Consistent with previous studies within the *Drosophila* genus, syntenic comparisons on all Muller elements reveal a rich history of intrachromosomal reshuffling of genes (Bhutkar et al. 2008; **Figure 1**). Interestingly, while Muller C and Muller D genes are mixed up along the telocentric chromosome 3, no shuffling of Muller B and Muller E genes occurred on the metacentric chromosome 2. This suggests that paracentric inversions are more frequent in this group than pericentric inversions, consistent with observations in other *Drosophila* groups (Sperlich and Pfreim 1986; Krimbas and Powell 1992).

Identification of inversions using whole-genome alignments

We used whole-chromosome alignments to identify inversions on each chromosome arm for species of the *D. nasuta* subgroup. We used MUMmer to compare the chromosomes of each species and used breaks in synteny to map inversion breakpoints (Kurtz et al. 2004; **Figure 2**). For each chromosome arm, we identified syntenic segments and we used GRIMM to find the minimum number of rearrangements required to account for the order and orientation of syntenic segments along the phylogeny (Tesler 2002). In total, we identify 22 large chromosomal inversions along the major chromosomes (**Figure 3**). We identify eight inversions on the ancestral X chromosome (Muller A), six inversions on the metacentric chromosome 2 (three on Muller B and three on Muller E), and eight inversion on the telocentric chromosome 3 (Muller C and Muller D). Thus, while encompassing only a single Muller element and thus being substantially smaller than other chromosomes, the X has a similar number of inversions. Higher rates of X-linked inversions have also been found in primates (Porubsky et al. 2020). The chromosomal inversions identified vary dramatically in size, ranging from 3.9-18.0 Mb, and contain hundreds or thousands of genes (**Table 1**).

Phylogenetic reconstruction of inversion

We reconstructed the evolution of inversions in the nasuta clade along the phylogeny using parsimony. **Figure 3** shows the inferred occurrence of inversions along different branches. Our sequenced strains of *D. albomicans* and *D. nasuta* differ by two overlapping inversions on Muller C/D (which forms the neo-sex chromosome in D. albomicans), but are otherwise co-linear. Their sister species *D. kepulauana* harbors two additional inversions on Muller C/D, and one on Muller B and Muller E, and this entire clade shares three inversions on Muller A.

The sister species *D. s. sulfurigaster* and *D. s. bilimbata* and entirely collinear, and a single shared inversion on Muller B distinguishes them from their sister clade *D. s. albostrigata*. The *sulfurigaster* clade has four inversions on Muller A in common, and on each on Muller B, E and C/D. Their sister species *D. pallidifrons* has one inversion on Muller A, CD and E, and two inversions on Muller C/D occurred in the common ancestor of the *sulfurigaster* flies and *D. pallidifrons*.

Overall, we find the average inversion rate to be 5.2 inversions per million years, consistent with previously found inversion rates in *Drosophila* (Lemeunier and Ashburner 1984; Powell 1997; Vieira et al. 1997; Bartolomé and Charlesworth 2006; Papaceit, Aguadé, and Segarra 2006; González, Casals, and Ruiz 2007; Ranz et al. 2007; Bhutkar et al. 2008) . However, there is high variation in the inversion rate per branch on the phylogeny (**Figure 3**). In particular, almost 1/3 of all inversions were identified on the short branch leading to species of the *sulfurigaster* species group.

In addition, inversions appear to be more common on the X chromosome compared to autosomes. While encompassing only about 1/5 of the total genome size, the X chromosome harbors more than 1/3 of all the inversions detected (**Figure 3**). Again, higher rates of inversions on the X chromosome are consistent with previous observations in *Drosophila* (Cheng and Kirkpatrick 2019.

Molecular characterization of breakpoints

Localizing the precise inversion breakpoints can be informative for several reasons. Inversions may directly impact gene structure or gene expression, and the identification of inversion breakpoints might provide insights into the molecular mechanisms by which inversions arise. We therefore carefully characterized all the inversion breakpoints on Muller B and Muller E.

We identify three inversions on Muller B (**Figure 4**). Inversion B1 (which occurred along the *D. kepulauana* branch) and B3 (occurring along the *D. s. bilimbata/D. s. sulfurigaster* branch) are about 18 Mb in size. B1 and B3 occurred at homologous positions in the genome, and both of their breakpoints are located within the histone gene cluster. Nonallelic homologous recombination could promote recurrent generation of inversions at the histone cluster, but it is also possible that this inversion was inherited from a common ancestor. Inversion B2 is about 11Mb long and shared by all *sulfurigaster* flies. One breakpoint of this inversion is located next to HP1 (Su(var)205), an important structural component of heterochromatin, but no repeated sequences are found at the breakpoints of the inverted chromosome (**Figure 4**).

Muller E harbors three inversions (**Figure 5**). Inversion E1 is about 10 Mb in size and occurred along the lineage leading to *D. kepulauana*. One of the breakpoints occurred at an approximately 1.6 kb repeat-masked region with no known homology aside from a 64 bp stretch that is homologous to R1-3_DF—a non-LTR retrotransposon. The other breakpoint is in a unique region (**Figure 5; Figure S1**).

Inversion E2 is about 10 Mb in size and occurred in the sulfurigaster lineage shared by *D. s. albostrigata, D. s. bilibmata*, and *D. s. sulfurigaster*. Both inversion breakpoints lie inside repetitive regions that are 1.2 kb and 13 kb in size (**Figure 5**; **Figure S2**). The breakpoint has a duplication of the Clbn gene along the *sulfurigaster* lineage while all other species in the *nasuta* clade have only a single copy of Clbn, suggesting that the inversion created a duplicate copy of this gene. Duplications of non-repetitive DNA at inversion breakpoints can be caused by staggered single-strand DNA breaks and repair by non-homologous end-joining (Ranz et al. 2007; Guillén and Ruiz 2012).

Inversion E3 occurred on the lineage leading to *D. pallidifrons* and is about 5 Mb in size. One inversion breakpoint is found inside a large (over 1 Mb long) repeat island, which in the *D. pallidifrons* genome is comprised of a number of transposons, over half of which are hAT elements. The other breakpoint is located within the tandemly duplicated multicopy gene CG31436. Thus, repeated sequences are found recurrently at inversion breakpoints in the *D. nasuta* species group.

Discussion

Inversion polymorphism has been studied for over a century in *Drosophila*. Inversions can have profound biological influences (see introduction), but the evolutionary processes maintaining inversions are typically poorly understood. Individuals heterozygous for inversions may suffer reduced fertility by producing nonfunctional gametes during meiosis. These fertility effects are expected to be less pronounced in *Drosophila*, since males generally lack recombination, and aberrant recombinant products contribute preferentially to the polar body nurse cells in females (Sturtevant and Beadle 1936; Reis et al. 2018).

Large differences in rearrangement rates have been reported between species and between chromosomes in *Drosophila*. We find dramatic variation in the rate of chromosomal inversions among lineages. Seven out of the 22 inversions identified map to the short branch that leads to flies of the *sulfurigaster* species complex, but only a single inversion on the branch leading to *D. albomicans* or *D. nasuta* (Figure 3). This is in agreement with previous observations in *Drosophila*, which found that rates of chromosomal inversions can differ by over an order of magnitude even among closely related species and between Muller's elements (Lemeunier and Ashburner 1984; Powell 1997; Vieira et al. 1997; Bartolomé and Charlesworth 2006; Papaceit, Aguadé, and Segarra 2006; González, Casals, and Ruiz 2007; Ranz et al. 2007; Bhutkar et al. 2008). This asymmetry in rates of inversions could result from differences in fitness effects or the efficacy of selection to establish new inversions, or from differences in mutation rates among lineages.

The molecular mechanisms of how inversions are generated are incompletely understood, and may differ among species or chromosomes. Inversions can be generated by nonallelic homologous recombination between repeated sequences, or by chromosome breakage and erroneous repair of the break by nonhomologous end-joining (Sonoda et al. 2006). Most inversion breakpoints in the *melanogaster* subgroup are associated with inverted duplication of genes or other non-repetitive sequences (Ranz et al. 2007). The presence of inverted duplications associated with inversion breakpoint regions was suggested to result from staggered breaks, followed by non-homologous end-joining. On the other hand, several studies in the *Drosophila* subgroup have suggested that repetitive elements are associated with the formation of inversion, suggesting an important role of ectopic exchange (Cáceres et al. 1999; Richards et al. 2005; Fonseca et al. 2012). In the *D. nasuta* subgroup, we find evidence for both processes.

Reuse of inversion breakpoints in *Drosophila* has been reported at both the cytological and molecular level (Theodosius Dobzhansky and Socolov 1939; Krivshenko 1963; Coluzzi et al. 1979; Pevzner and Tesler 2003; Zhao et al. 2004; Murphy et al. 2005; Richards et al. 2005; Goidts et al. 2005; Bhutkar et al. 2008; Fuller et al. 2018). We find that the histone gene cluster, which is located on two separate regions on Muller B in flies of the *nasuta* subgroup was involved in the generation of inversions in two separate lineages (though we cannot rule out that this inversion was segregating in a common ancestor of this species group). This resembles findings in great apes, where a high rate of homoplasy of inversions was observed (Porubsky et al. 2020). Reuse of inversion breakpoints might be due to mutational bias if these regions are particularly prone to breakage, or driven by selection if a specific breakpoint position affects the intrinsic fitness of a new arrangement (McBroome, Liang, and Corbett-Detig 2020). Mutations caused by inversion breakpoints may have diverse consequences, from gene disruptions to generation of new gene duplicates or transfer of regulatory sequences from one gene to another. We identify one instance of a gene duplication generated by an inversion on Muller E in the *sulfurigaster* clade.

Chromosomal inversions can maintain linkage among alleles that are favored by natural selection and inversions that are associated with complex polygenic phenotypes are known from a variety of taxa (Hoffmann and Rieseberg 2008). Species from the *nasuta* clade have recently diverged, but differ in various morphological and behavioral phenotypes (Kitagawa et al. 1982; Spieth

1969). It will be of great interest to address the role of chromosomal inversions in contributing to phenotypic differences and local adaptation.

Methods

Genome Assemblies & Annotations

We used chromosome-level assemblies for seven species of the *D. nasuta* species group, which are described elsewhere (Wei, Mai et al., in preparation). Table S1 lists the strains that were investigated. All assemblies are highly contiguous (N50s ranging from 34 Mb to 38 Mb) and very complete (with BUSCO scores ranging from 98.5% to 99.7%), and total assembly sizes ranging from 161 Mb to 163 Mb. For each species, the euchromatic portion of all Muller elements are assembled as a single contig, and only highly repetitive pericentromeric fragments could not be placed on the assembly (the assemblies comprise between 77 to 282 scaffolds with a mean of 157 scaffolds).

Gene annotations for each species (from Wei, Mai et al., in preparation) were clustered with *D. virilis* gene annotations using OrthoDB (Kriventseva et al. 2019). We then assign a name to each gene based on clustering with *D. virilis* annotations and their homology to *D. melanogaster* genes. An average of 10,534 std dev = 40) genes were assigned to a D. melanogaster gene; 10,336 of these are single copy genes (std dev = 54) and 198 are duplicated genes (std dev = 41).

Inversion Along Phylogeny

MUMmer was used to determine the inversion status between all genome assemblies using the *D. albomicans* assembly as the reference (Kurtz et al. 2004). Sequences between each inversion breakpoint are assigned a numeric identifier and an optional negative sign to denote an inverted status relative to *D. albomicans* for each genome, which are then represented by an ordered sequence of these identifiers. The numeric sequence for each genome is then used as input for GRIMM to determine the optimal rearrangement scenario along a phylogeny, and identify the most likely ancestral genome structure (Tesler 2002).

We generated an "ancestral genome" by using the *D. albomicans* genome assembly, and 'un-inverting' all the inversions occurring along the branches leading to *D. albomicans* (that is, inversion A1, A2, A3, CD1; see **Figure 3**). To call inversion breakpoints, we took the mean between the end of the alignment on one side of the inversion and the start of the alignment on the other side of the alignment from MUMmer coordinates (**Table S2**).

All genome assemblies were then aligned to the ancestral genome using MUMmer. For each chromosome, all different inversion breakpoints are used to demarcate regions along the chromosome and inversions along the genome were then estimated using GRIMM based on the order and orientation of these regions relative to the ancestral genome.

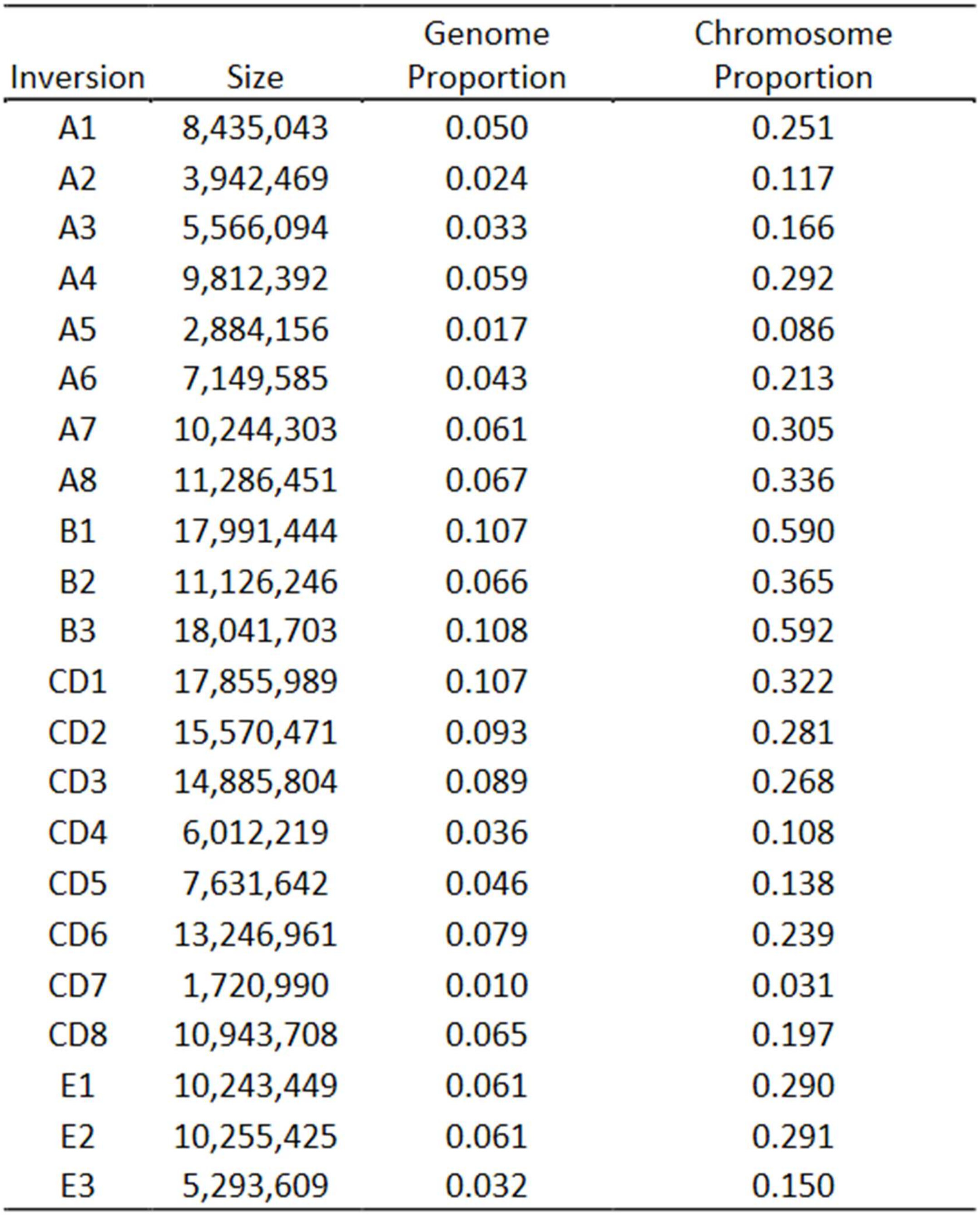
Inversion Rates

To test the rates of inversions along the phylogeny, we also calculate the branch lengths along the phylogeny. They are determined using the same method as Mai et al. 2019, using mean Ks values between species, a neutral mutation rate estimate of 3.46×10−9 per base per generation, and a 7 generation per year estimate for *Drosophila* (Z. Zhang et al. 2006; Cutter 2008; Keightley et al. 2009b). Internal branch lengths are calculated by subtracting the divergence time between sister species from the divergence time between the mean of the sister species and an outgroup. For example, let d_{AB} be the divergence time between species A and B. Given the phylogeny ((A, B), C), the length of the branch from the root node to the shared node between A and B is calculated by $\frac{d_{AC} + d_{BC}}{2} - d_{AB}$.

The overall inversion rate is calculated using the total number of inversions on the phylogeny divided by the total length, in millions of years, of the phylogeny (in other words, the sum of all branches along the phylogeny). The inversion rate per branch is calculated by dividing the number of inversions occurring on the branch by the branch length.

Acknowledgements

We acknowledge Ryan Bracewell for discussion.



Inversion	Size	Genome Proportion	Chromosome Proportion
A1	8,435,043	0.050	0.251
A2	3,942,469	0.024	0.117
A3	5,566,094	0.033	0.166
A4	9,812,392	0.059	0.292
A5	2,884,156	0.017	0.086
A6	7,149,585	0.043	0.213
A7	10,244,303	0.061	0.305
A8	11,286,451	0.067	0.336
B1	17,991,444	0.107	0.590
B2	11,126,246	0.066	0.365
B3	18,041,703	0.108	0.592
CD1	17,855,989	0.107	0.322
CD2	15,570,471	0.093	0.281
CD3	14,885,804	0.089	0.268
CD4	6,012,219	0.036	0.108
CD5	7,631,642	0.046	0.138
CD6	13,246,961	0.079	0.239
CD7	1,720,990	0.010	0.031
CD8	10,943,708	0.065	0.197
E1	10,243,449	0.061	0.290
E2	10,255,425	0.061	0.291
E3	5,293,609	0.032	0.150

Table 1. Inversion sizes along the *nasuta* phylogeny

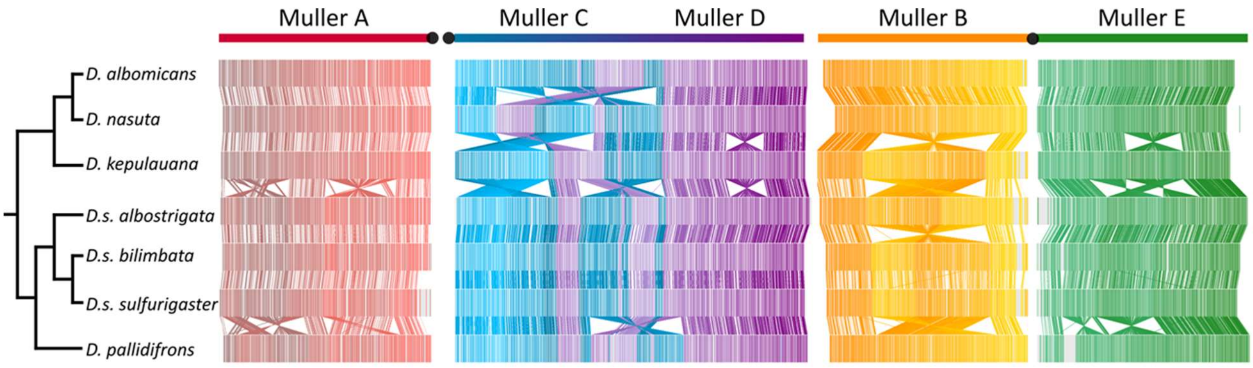


Figure 1. – Phylogenetic relationship between species investigated and chromosomal synteny based on alignments of orthologous single-copy genes. Genes are color-coded according to their assignments to Muller elements. Note that Muller A and C/D are fused in *D. albomicans*.

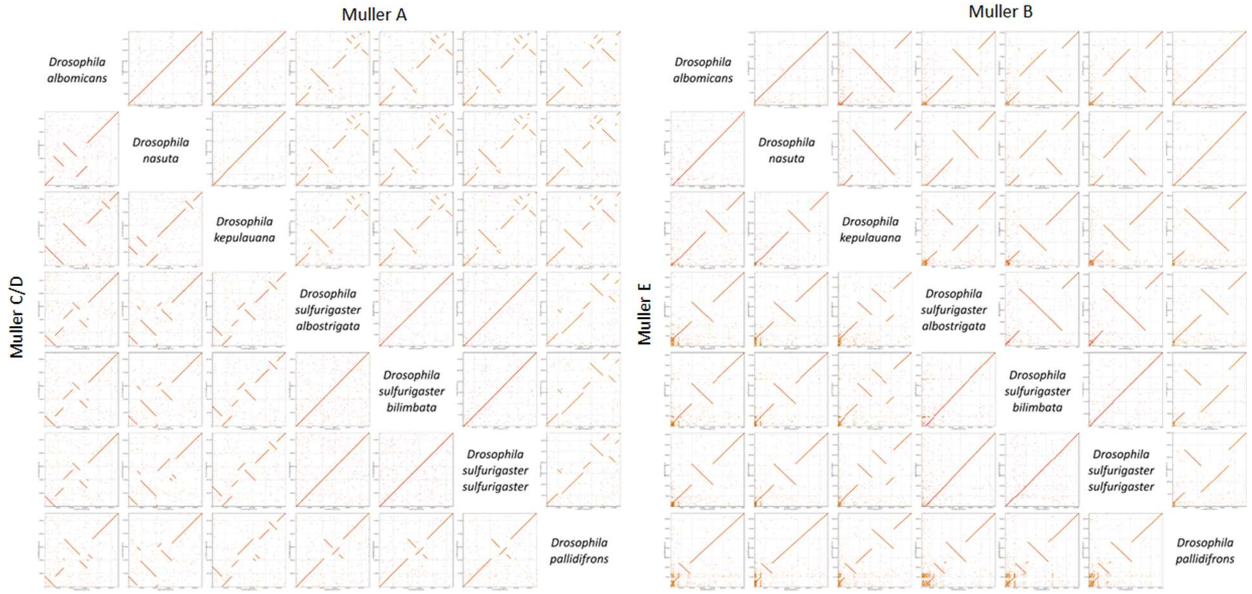


Figure 2. – Dotplot between species for all major chromosome arms (Muller elements A through E). The pericentromere of each chromosome arm is placed at the bottom left corner of each subplot. Note that chromosomes are not drawn to scale (i.e. Muller C/D is approximately twice as large as all other chromosome arms).

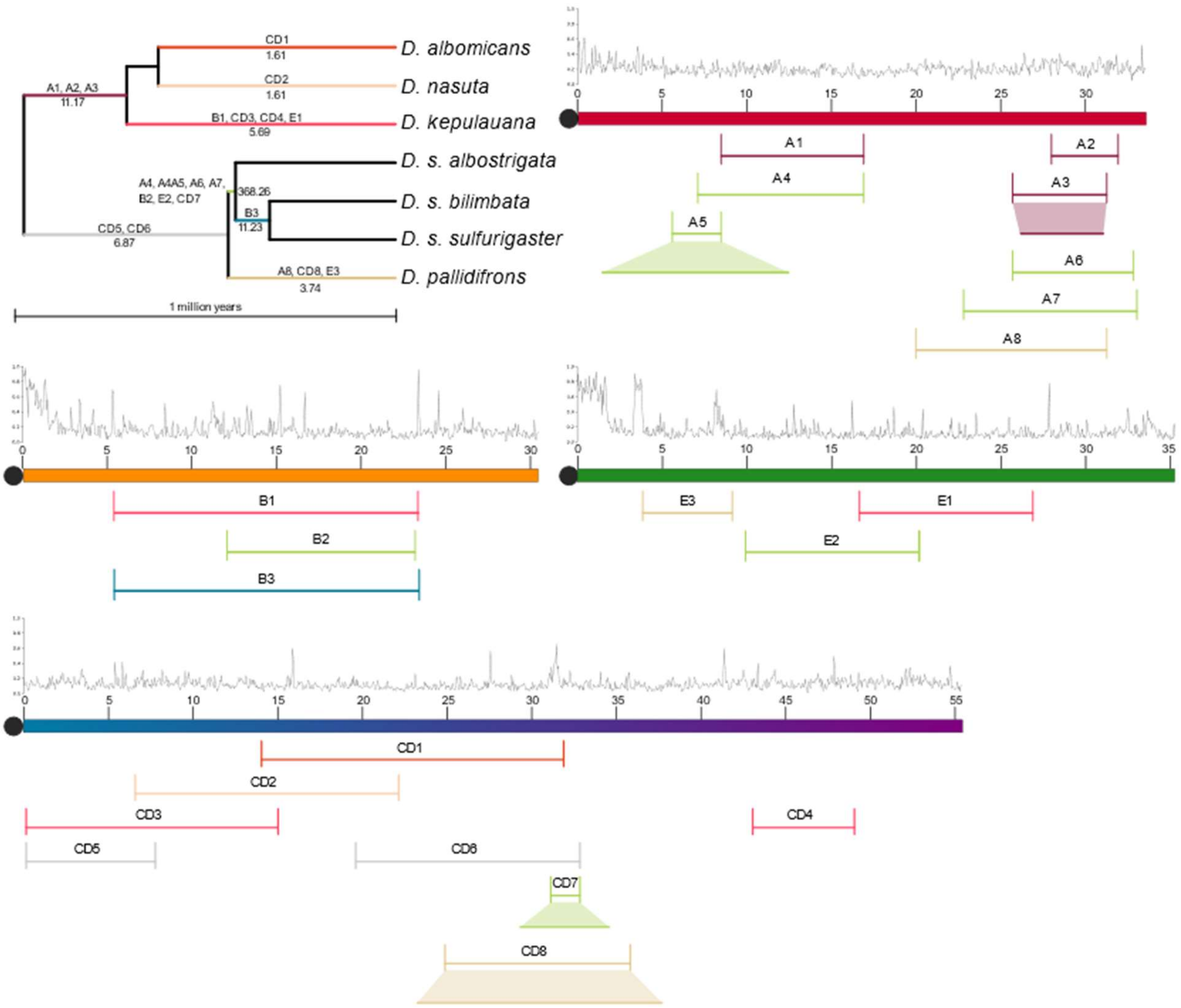


Figure 3. – Major inversions along the phylogeny based on parsimony reconstruction. The approximate location of inversions along the ancestral chromosomes is indicated, and their size (the size of nested inversions is indicated under the shaded region). Inversions are color coded by the branch on the phylogeny where they occurred. Dots indicate location of the centromere and ticks mark every 5 Mb. The repeat content (fraction repeat masked in 50-kb windows) is shown above each chromosome. The number of inversions per million years is shown along each branch (bottom) of the phylogeny.

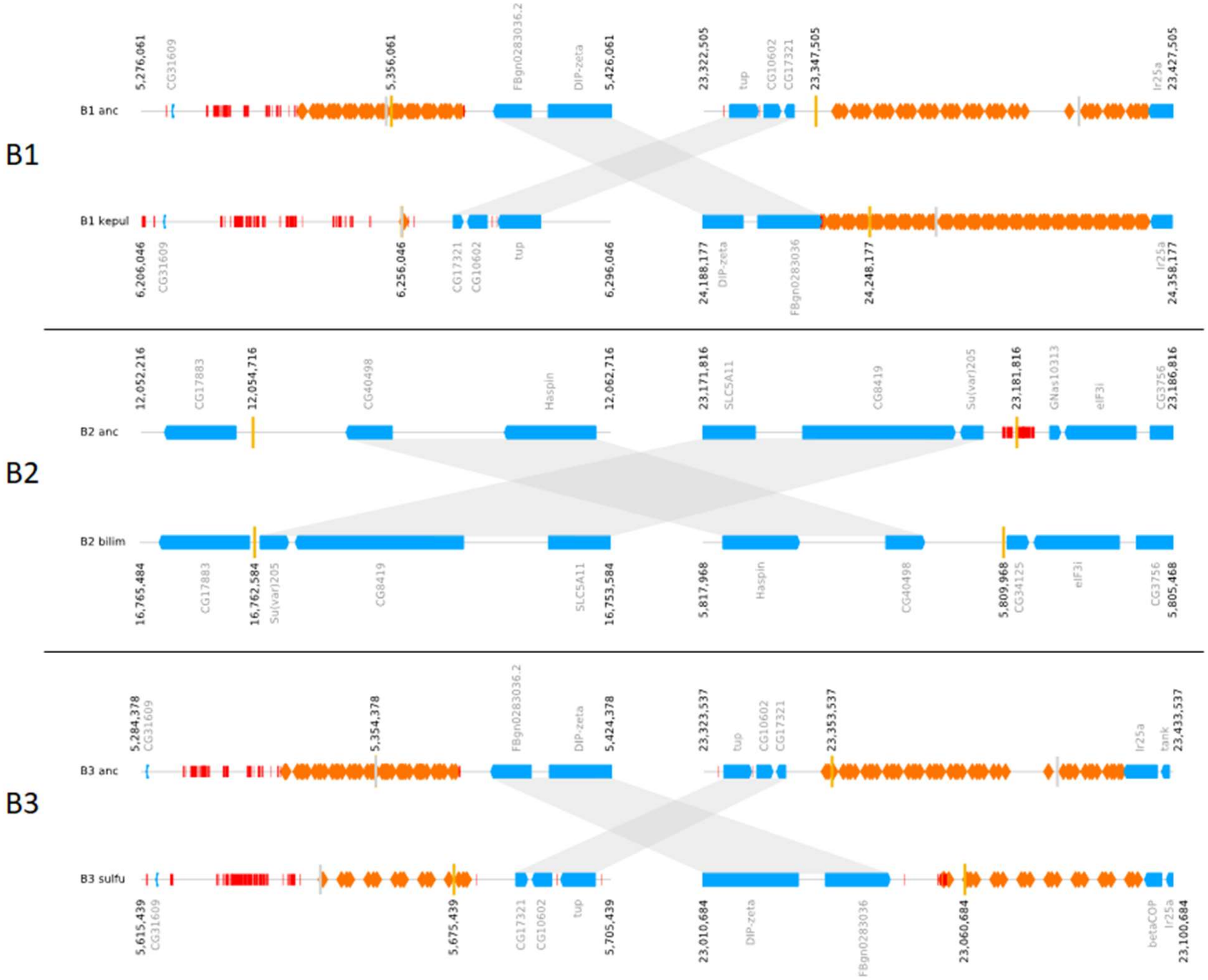


Figure 4. – Inversion breakpoints on Muller B. Blue boxes indicate protein-coding genes, orange boxes indicate histone genes, and red boxes indicate repeats. Approximate breakpoint coordinates are given (yellow line), and homologous regions inside the inversion breakpoints are shown by grey shading.

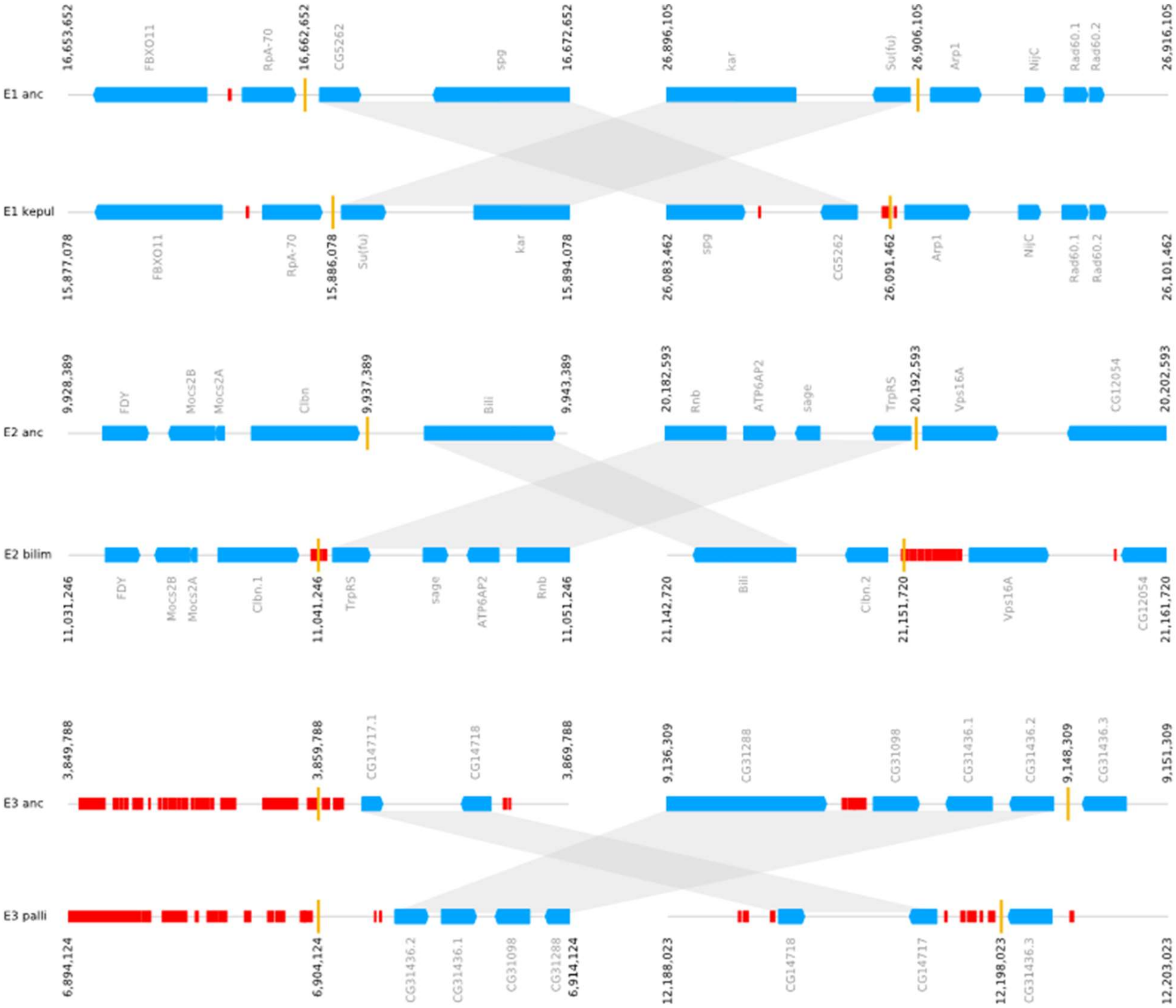
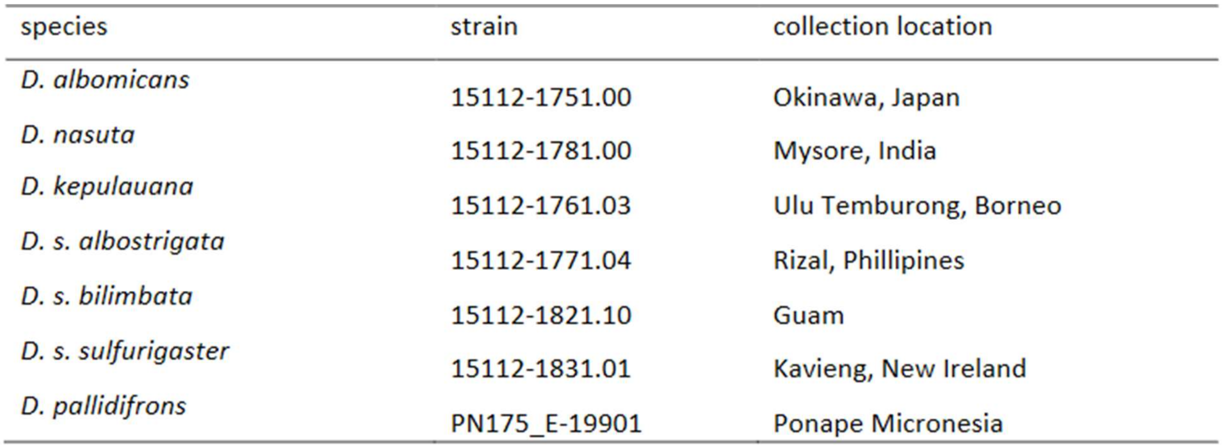
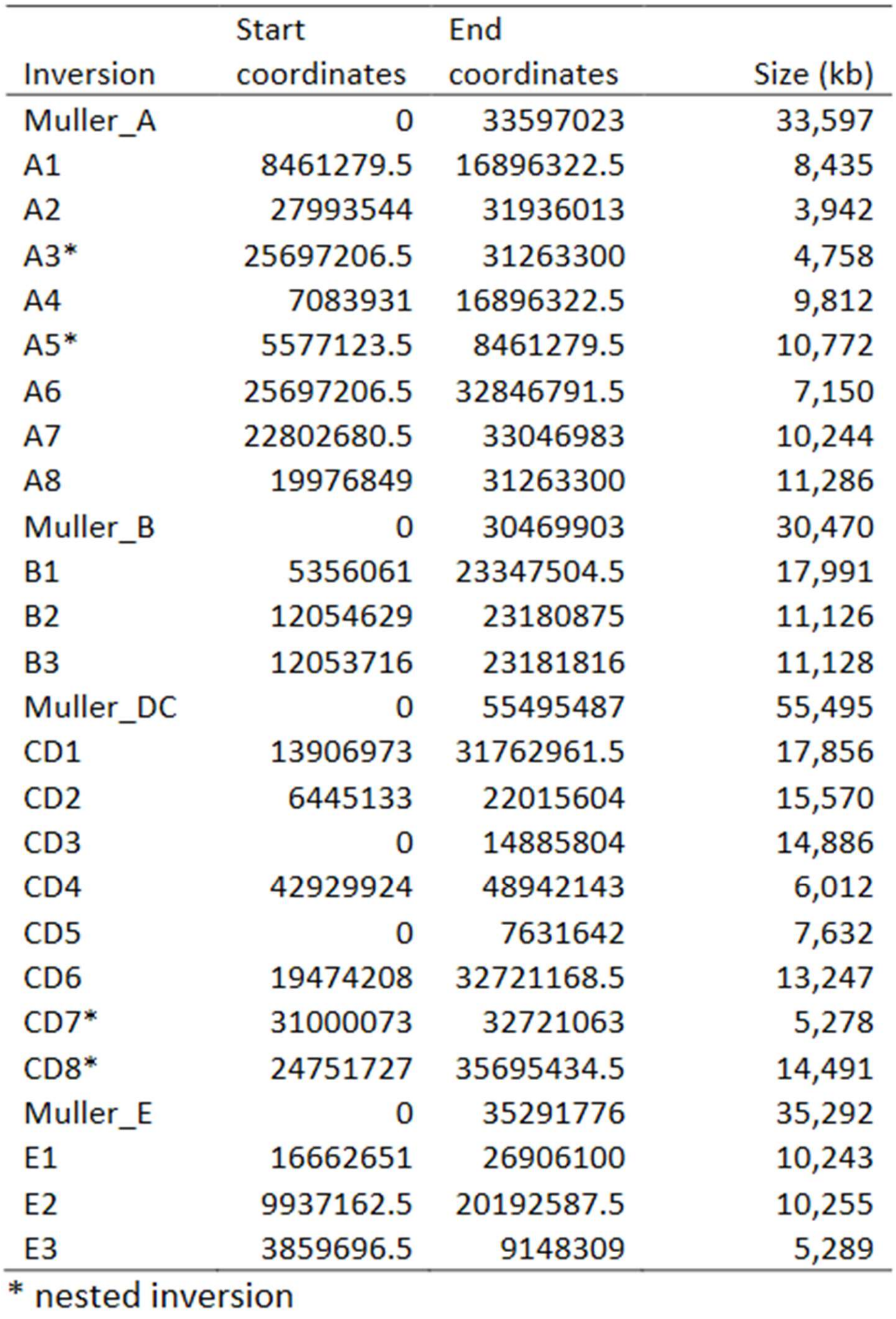


Figure 5. – Inversion breakpoints on Muller E. For legend, see **Fig. 4**



species	strain	collection location
D. albomicans	15112-1751.00	Okinawa, Japan
D. nasuta	15112-1781.00	Mysore, India
D. kepulauana	15112-1761.03	Ulu Temburong, Borneo
D. s. albostrigata	15112-1771.04	Rizal, Phillipines
D. s. bilimbata	15112-1821.10	Guam
D. s. sulfurigaster	15112-1831.01	Kavieng, New Ireland
D. pallidifrons	PN175_E-19901	Ponape Micronesia

Table S1. Strains investigated



Inversion	Start coordinates	End coordinates	Size (kb)
Muller_A	0	33597023	33,597
A1	8461279.5	16896322.5	8,435
A2	27993544	31936013	3,942
A3*	25697206.5	31263300	4,758
A4	7083931	16896322.5	9,812
A5*	5577123.5	8461279.5	10,772
A6	25697206.5	32846791.5	7,150
A7	22802680.5	33046983	10,244
A8	19976849	31263300	11,286
Muller_B	0	30469903	30,470
B1	5356061	23347504.5	17,991
B2	12054629	23180875	11,126
B3	12053716	23181816	11,128
Muller_DC	0	55495487	55,495
CD1	13906973	31762961.5	17,856
CD2	6445133	22015604	15,570
CD3	0	14885804	14,886
CD4	42929924	48942143	6,012
CD5	0	7631642	7,632
CD6	19474208	32721168.5	13,247
CD7*	31000073	32721063	5,278
CD8*	24751727	35695434.5	14,491
Muller_E	0	35291776	35,292
E1	16662651	26906100	10,243
E2	9937162.5	20192587.5	10,255
E3	3859696.5	9148309	5,289

* nested inversion

Table S2. Inferred approximate inversion breakpoints

Chapter 3: Dynamics and impacts of transposable element proliferation during the Drosophila nasuta species group radiation

Authors: Kevin H.-C. Wei, Dat Mai, Kamalakar Chatla, Doris Bachtrog

Department of Integrative Biology, University of California Berkeley, Berkeley, California, United States of America.

Transposable element (TE) mobilization is a constant threat to genome integrity. Eukaryotic organisms have evolved robust defensive mechanisms to suppress their activity, yet TEs can escape suppression and proliferate, creating strong selective pressure for host defense to adapt. This genomic conflict fuels a never-ending arms race that drives the rapid evolution of TEs and recurrent positive selection of genes involved in host defense; the latter has been shown to contribute to postzygotic hybrid incompatibility. However, how TE proliferation impacts genome and regulatory divergence remains poorly understood. Here, we report the highly complete and contiguous (N50=33.8Mb - 38.0Mb) genome assemblies of seven closely-related *Drosophila* species that belong to the *nasuta* species group - a poorly studied group of flies that radiated in the last 2 million years. We constructed a high quality de novo TE library and gathered germline RNA-seq data, which allowed us to comprehensively annotate and compare insertion patterns between the species, and infer the evolutionary forces controlling their spread. We find a strong negative association between TE insertion frequency and expression of genes nearby; this likely reflects survivor-bias from reduced fitness impact of TE inserting near lowly expressed, non-essential genes, with limited TE-induced epigenetic silencing. Phylogenetic analyses of insertions of 147 TE families reveal that 53% of them show recent amplification in at least one species. The most highly amplified TE is a non-autonomous DNA element DINE which has gone through multiple bouts of expansions with thousands of full length copies littered throughout each genome. Across all TEs, we find that TEs expansions are significantly associated with high expression in the expanded species consistent with suppression escape. Altogether, our results shed light on the heterogenous and context-dependent nature in which TEs affect gene regulation and the dynamics of rampant TE proliferation amidst a recently radiated species group.

Introduction

Eukaryotic genomes are littered with transposable elements (TEs). TEs are selfish genetic elements that self-replicate via copy and paste or cut and paste mechanisms. Despite their abundance and ubiquity in genomes (Kidwell 2002), they can be highly deleterious especially when active. When they transpose, TEs can create double strand breaks and disrupt reading frames when inserted into genes (Hedges and Deininger 2007). Even when transpositionally inactive, they can induce non-allelic exchange due to sequence homology which can create devastating genome rearrangements (Athma and Peterson 1991; Kidwell and Holyoake 2001; Xiao, Li, and Peterson 2000; Zhang et al. 2011).

To combat their deleterious activity, eukaryotic genomes have evolved intricate defense pathways to inactivate TEs both transcriptionally and post-transcriptionally (for review see Ozata et al. 2019). Post-transcriptional silencing generally involves small RNA-targeted degradation of TE transcripts (for reviews see Czech et al. 2018; Ozata et al. 2019; Wang and Lin 2021). Transcriptional inactivation is achieved through compaction of the chromatin environment into a dense and inaccessible state, known as heterochromatin (for reviews see (Richards and Elgin 2002; Elgin and Reuter 2013). This involves di- and tri-methylation to the histone H3 tail at the 9th lysine (H3K9me2/3), which in turn recruits neighboring histones to be methylated allowing heterochromatin to spread across broad domains (Nakayama et al. 2001; Lachner et al. 2001; Bannister et al. 2001; Hall et al. 2002). Interestingly, this spreading mechanism can also have the unintended effect of silencing genes nearby TE insertions (Choi and Lee 2020). Therefore, in addition to disrupting coding sequences, TE insertions can further impair gene function by disrupting gene expression (Hollister and Gaut 2009; Lee 2015; Lee and Karpen 2017).

However, even with strong repressive mechanisms, defense against TEs appears to be an uphill battle. TEs are among the most rapidly changing components of eukaryotic genomes. TE content can differ drastically even between closely related species and has been shown to be a key contributor to genome size disparities. In *Drosophila*, the P-element, a DNA transposon originating from *D. willistoni*, invaded both *D. melanogaster* (Anxolabéhère, Kidwell, and Periquet 1988; Daniels et al. 1990) and subsequently *D. simulans* (Kofler et al. 2015). Both of these cross-species invasions occurred rapidly within the last century and resulted in world-wide sweeps of the P-element in wild populations. Mobilization events are accompanied by reduction in host fertility and viability (Kidwell, Kidwell, and Sved 1977; Kidwell and Novy 1979; Schaefer, Kidwell, and Fausto-Sterling 1979), which in turn creates strong selective pressure for the host to evolve an updated repressive mechanism (Simkin et al. 2013; Kelleher and Barbash 2013). Such dynamics create an evolutionary arms-race between host suppression mechanisms and TE suppression escape, and is thought to underlie the recurrent adaptive evolution of many proteins involved in the TE silencing pathways (Parhad and Theurkauf 2019; Luo et al. 2020). Rapid evolution of TEs and the repressive pathways have even been implicated in establishing postzygotic reproductive isolation between closely related *Drosophila* species (Kliman et al. 2000; Garrigan et al. 2012; Brand et al. 2013).

Beyond their deleterious potential, TEs can also be sources of novelty in the genome (Kidwell and Lisch 1997). TEs, or parts of their sequences, have been co-opted for gene regulatory functions such as promoters and enhancers (Jacques, Jeyakani, and Bourque 2013; Merenciano et al. 2016; Sundaram and Wysocka 2020). Their recurrent transpositions across nascent sex chromosomes also mediated the evolution of dosage compensation chromosome-wide (Ellison and Bachtrog 2013; Zhou et al. 2013). Insertions of TEs to the proximity of genes have also been shown to create functional chimeric retrogenes (Buzdin 2004; Xing et al. 2006). In mammals, KRAB-zinc finger transcription factors have repeatedly co-opted the transposase protein encoded by DNA transposons, allowing for the diversification of their binding targets (Cosby et al. 2021). Lastly, in flies, domesticated retrotransposons insert at chromosome ends for telomere extension thus alleviating the need for telomerase to solve the end-replication problem (Traverse and Pardue 1988; Biessmann et al. 1990; Levis et al. 1993). Therefore, TEs do not just force the host defense to adapt in order to suppress their activity, but they can also be beneficial drivers of genome evolution (Kidwell and Lisch 1997; Casacuberta and González 2013).

While TEs can have multi-faceted influences on the genome and its evolution, the dynamics of TE amplification and suppression escape remain poorly understood, especially outside of select model species. This is in part due to the inherent challenge associated with studying highly repetitive sequences, an issue that became particularly problematic during the boom of short-read sequencing technologies in the last two decades. Most TE-derived short reads (typically less than 150bps) cannot be uniquely assigned to a region of the genome, which causes errors in mapping and breakages in genome assemblies (Bourque et al. 2018; O'Neill, Brocks, and Hammell 2020). Numerous approaches have been devised that take advantage of different features of short-read sequencing platforms (e.g. paired sequencing) to call insertions (Linheiro and Bergman 2012; Cridland et al. 2013; Rahman et al. 2015; McGurk and Barbash 2018; Wei, Gibilisco, and Bachtrog 2020), but such methods are nevertheless limited by short read lengths, often producing inconsistent results (Vendrell-Mir et al. 2019). With the advent of long-read (5kb+) sequencing technologies from Oxford Nanopore and PacBio, many of these issues can finally be circumvented (Hotaling et al. 2021). The use of such technologies have already led to drastic improvements of genome assemblies across highly repetitive genomes in, for example, flies (Mahajan et al. 2018; Bracewell et al. 2019; Chakraborty et al. 2021), mosquitoes (Matthews et al. 2018), mammals (Bickhart et al. 2017), and humans (Nurk et al. 2021)..

Highly contiguous genomes with well-represented repeat content permit comprehensive analyses of TE insertions across the genome. Multiple such high quality genomes further enable analyses of the dynamics of TE proliferation through a comparative and phylogenomics framework. Therefore, to illuminate how TEs proliferate and potentially drive genome evolution and speciation, we used long-read technologies to generate high quality genome assemblies of seven closely related *Drosophila* species (**Figure 1A,B**) that belong to the *nasuta* group. These species group radiated in the last two million years (Kitagawa et al. 1982; Bachtrog 2006; Ranjini and Ramachandra 2013; Mai, Nalley, and Bachtrog 2020) and is widely distributed across Asia, with some populations found in eastern Africa, Oceania, and Hawaii (Wilson et al. 1969; Mai, Nalley, and Bachtrog 2020), and *D. nasuta* has recently been identified as an invasive species in Brazil that is spreading quickly in South America (Vilela and Goñib 2015). While most of the species are geographically isolated, they have varying levels of reproductive isolation (Kitagawa et al. 1982); over half of interspecific crosses produce viable offspring. With these high quality genomes, we sought to systematically understand how TE insertions around genes affect gene expression, and how frequently TEs escape repression and expand. To answer these questions, we generated a library of a common set of high quality TE consensus sequences from de novo TE calls across the genome assemblies. With this library, we identified species-specific TE insertions and found that TEs frequently expand, likely due to suppression escape, with >50% of TEs showing evidence of lineage-specific expansion in at least one species. Species-specific TEs are disproportionately found near lowly expressing genes and rarely have impact on gene expression. Lastly, we show that silencing of expanding TEs can lead to silencing of neighboring genes.

Results

High quality genome assemblies across seven species

Genome assemblies for females of seven species in the *nasuta* clade—*D. albomicans*, *D. nasuta*, *D. kepulauana*, *D. sulfurigaster albostrigata*, *D. sulfurigaster bilimbata*, *D. sulfurigaster sulfurigaster*, and *D. pallidifrons*—were generated using Nanopore and Hi-C reads (**Table 1; Figs. S1-S7**). The methodology for preparing reads was adopted from Bracewell et al. and applied across all species: error-correct Nanopore reads with canu, generate contig assembly with wtdbg2 and flye, polish assembly with racon and pilon, remove contigs that belong to other organisms with BLAST, and stitch contig assemblies using Hi-C reads as input for Juicer and 3d-dna (Altschul et al. 1990; Walker et al. 2014; Durand et al. 2016; Dudchenko et al. 2017; Koren et al. 2017; Vaser et al. 2017; Bracewell et al. 2019; Kolmogorov et al. 2019; Ruan and Li 2020). However, there is no universal pipeline to generate ideal assemblies; the assembly pipeline for different flies underwent various adjustments for optimal results (see Methods). Overall, we generated consistent assemblies for each species using an average of 30.3x long read coverage (std dev = 10; **Table S1**), resulting in a mean N50 of 35.9 Mb (std dev = 1.4 Mb; **Table 1**), assembly size of 166.6 Mb (std dev = 2.59 Mb; **Table 1**), and BUSCO score of 99.3% (std dev = 0.4%; **Table S2**).

We leveraged the chromosome level genome assemblies alongside RNA-seq data from *D. albomicans* and *D. nasuta* (Zhou and Bachtrog 2012) to annotate genes across all species (**Table 1**). An average of 12,513 genes were annotated per species (std dev = 128.48), which is lower than the number of genes annotated in other *Drosophila* species (Drosophila 12 Genomes

Consortium et al. 2007). In order to analyze homologous genes, we clustered genes between species with OrthoDB and found 9,413 genes shared across all species (Kriventseva et al. 2019).

Generating a curated de novo TE library

For each genome, we used RepeatModeler2 to generate a de novo TE library, which we then used to annotate the genome (Flynn et al. 2020). This resulted in between 18.8%-23.3% of the genomes being masked (**Table S3**). Further, high repeat content near chromosome ends show that these near-chromosome length scaffolds include some heterochromatin and pericentromeric regions. Expectedly, gene density and repeat density are negatively correlated (**Figure 1C**).

One major challenge with de novo TE identification using standard computational methods is that the resulting TE libraries are littered with redundant and fragmented entries. Furthermore, we find that secondary structures such as nested insertions or fragment duplications (**Figure S8**) are frequently identified as unique TE entries in the libraries. To improve the de novo TE library and to generate a common set of TE consensus sequences across all the *nasuta* subgroup, we devised a pipeline that utilizes multiple steps and metrics (**Figure 2A**). After an initial de novo TE library call with RepeatModeler2 for each of the genomes, we demarcated the euchromatin/pericentromere boundaries (**Figure 1C**). Reasoning that recently active TEs are more likely to be intact and surrounded by unique sequences in the euchromatin, we then ran RepeatModeler2 for a second time on only the euchromatic portions of the genome assemblies. The resulting TE libraries were then merged across all the species generating a library of 1818 entries.

We then used CD-HIT2 to group the entries into clusters based on sequence similarity (Fu et al. 2012). By default, CD-HIT2 outputs the longest sequence in each cluster. While this means full length entries will be favored over fragmented entries (when both exist in the library), entries with nested structures or chimeric TEs will be selected in favor of full-length but shorter elements. Therefore, in addition to sequence length, we evaluate each TE in each cluster based on two additional metrics to preferentially select representative and full length TE consensus sequences. While increasing entry lengths, chimeric TEs are unlikely to be frequently found in the genome; we therefore blasted the TE entries to the genome and tallied the number of times hits cover 80% of the length of the entries. In addition, we blasted the TEs to themselves to determine internal redundancy; entries with internal duplications or nested insertions will have a high self-blast score. We then selected the representative sequence as the longest sequence with high numbers of near full length blast hits and low self-blast score. We then repeated this step one more time to further remove redundancies in the library. After these two rounds of clustering with CD-HIT2, the TE library size was reduced to 351 consensus sequences. Afterwards, we merged TE sequences that make up a larger, full length element through patterns of co-occurrences in the genome. This resulted in a substantially reduced library with 318 entries.

The TEs generated from RepeatModeler2 are, by default, assigned to a TE category. To validate these assignments, we used ClassifyTE to reannotate the TE library (Panta et al. 2021). There is a 50% concordance between the annotations from RepeatModeler2 and ClassifyTE. Entries that were different between the two annotations were assigned the default category from RepeatModeler2. Gypsy elements make up the majority of the TE library, consisting of 82 entries (25.8%) followed by unknown families (57 entries, 17.9%; **Figure 2B**). All other TE families make up less than 5% of the TE library. The pattern of high number of Gypsy families is similar to those in other *Drosophila* species (Mérel et al. 2020).

TE Insertion patterns across the genome

Using the refined *nasuta* group-specific TE library, we annotated TE insertions in each genome assembly using RepeatMasker (see Materials and methods). We classified full length insertions as annotations that cover at least 80% of the entry in the library; insertions covering less than 80% and are over 200bp are classified as truncated insertions. In addition, we merged annotations that are contiguous or overlapping, which can be due to nested insertions or remaining redundancies in the repeat library. The number of full length TEs range from 3489 to 4544 (**Figure 2C**) and the majority (73.6% on average) fall within euchromatic regions of the genomes, (**Figure S9C**), similar to previous reports (Biémont and Vieira 2005; Drosophila 12 Genomes Consortium et al. 2007). Truncated insertions are nearly 2x as numerous (ranging between 7164- 8273; **Figure 2C**). As expected given their mosaic nature, the merged annotations have the largest fractions fall within the heterochromatic regions (**Figure S9**). With the exception of four TE families, all are found in low to intermediate copy numbers with fewer than 100 copies in any given genome, consistent with previous findings (**Figure 2D**). Interestingly, one TE stands out as having thousands of copies across all the genomes (**Figure 2D**, arrowhead, see section below).

To evaluate if and how TEs impact gene function, we looked at TE insertion patterns with respect to neighboring genes (**Figure 2E**, **Figure S9A**). On average, TE insertions are 18.5 kb away from the nearest genes; 41.5-46.9% of insertions are within 5kb of genes (**Figure 2E**). Of the 12,362-12,718 genes annotated, 2,887 to 3,343 have insertions within or nearby (<1kb 5' or 3'). Of those, 48.9% to 50.1% of genes have insertions within introns, which would not affect the reading frame (**Figure 2G**).

TE insertions are associated with low expression of nearby genes

To systematically examine the impact, if any, of TE insertions on gene expression, we generated ovarian and testes mRNA-seq for five of the seven species investigated (excluding *D. s. bilimbata* and *D. s. albostrigata)*. Genes with TE insertions nearby or within are over-represented for lowly expressed genes, in both testis and ovaries (p < 2.2e-16 Wilcoxon rank sum test; **Figure 2G**; **Figure S10**). Genes with insertions less than 2kb upstream have the lowest expression in both the testes and ovaries (**Figure 2G**). Genes with TEs inserted further away (2-5kb) also have significantly lower expression, though to a lesser extent (**Figure 2G, Figure S10**). Moreover, we find that gene expression is inversely correlated with the number of TE insertions (**Figure 2H**). This negative relationship holds for insertions found within, upstream, and downstream of genes. Interestingly, ovarian expression appears to be more negatively associated with TE insertions, with no expression in ovaries of nearly half of the genes with TEs inserted nearby (**Figure 2G, H**).

Due to the spreading of heterochromatin, TE insertions can induce epigenetic silencing at neighboring genes (Choi and Lee 2020). Therefore, prima facie, these results are consistent with the epigenetic silencing of genes due to neighboring TE insertion. To further test this, we reasoned that if TE insertions are inducing downregulation of surrounding genes, orthologous genes without insertions should be more highly expressed. To test this possibility, we compared the expression of orthologs when insertions are found in one species but not the other. Curiously, we do not find that expression between orthologs changes significantly depending on the presence of insertions nearby or within (**Figure 2F**, **Figure S11**). This suggests that insertions within/nearby genes are not systematically downregulating expression. Instead, TEs appear to preferentially insert and/or accumulate around lowly expressed genes.

Survivor bias likely drives anti-correlation between TE insertion and gene expression

To elucidate the source of the negative association between gene expression and TE insertions nearby, we looked at all TE insertions found around/within the 9413 genes with orthologs across all species. To ensure that only unique insertions are counted and ancestral insertions are counted only once, we removed insertions belonging to the same TE family that are within 100bp relative to the neighboring genes. Further, we removed pericentric genes from these analyses to avoid their high local TE counts driving correlations. For these gene orthologs, we indeed find a significant negative correlation between TE insertion counts and averaged gene expression in both testes and ovaries (**Figure 3A, B**; p < 2.2e-16). Similar correlations are also found when looking at the proportion of bases covered by TEs around and within genes (**Figure S12**). Curiously, the negative correlation of ovarian expression is significantly stronger than that of testes expression (**Figure 3A, B**; p < 1e-8, Pearson and Filon's z).

We then looked at the extent of correlation between species-specific TE insertion counts to gene expression across species. If TEs insert independently at different genes and are down-regulating nearby genes in one species, we expect no cross-species correlations. Instead, significantly negative correlations are observed between all pairwise comparisons (**Figure 3D**), although the within-species correlations are significantly more negative than between-species correlations (**Figure 3D**, outlined boxes). Further, insertion-induced epigenetic down-regulation to neighboring genes is expected to increase expression divergence between species, since genes with insertions are expected to be more lowly expressed than their orthologs without insertions. We do not find any significant correlation between insertion counts around genes in one species and their expression fold-differences when compared to orthologs without insertions (**Figure 3E**). However, when comparing the distribution of TEs between species, we find that the number of TE insertions at/near genes are correlated between many of the species (**Figure 3F**). Especially between more closely related species pairs, the correlation of insertions are highly significant, suggesting that TEs have a tendency to independently insert and/or accumulate near the same genes in different genomes. Thus, between-species correlations in TE counts vs. gene expression (**Figure 3D**), and low interspecific expression divergence (**Figure 3E**) may in part be explained by the same genes being targeted by TEs in different species. Biased insertion counts near lowly expressed genes could be due to insertion bias or survival bias. The former can result from TEs preferentially targeting specific genomic features to insert such as promoters and accessible chromatin; the latter is likely the result of low fitness consequences due to insertions near lowly expressed genes.

TE insertions associated with extreme expression changes in a small number of genes

TE insertions do not appear to have pervasive silencing effects on neighboring genes (**Figures 2G, 3B-C, 4D**). However, there are known cases where individual TE insertions modulate gene regulation of nearby genes. To identify such cases, we compared the expression of each gene in each species to the average expression across all species (**Figure 3G**, **Figure S13**). For the vast majority of genes with/nearby insertions, their expression does not deviate from the cross-species average. However, interestingly, we notice multiple cases where insertions are associated with substantially lowered gene expression. Examining the small fraction of genes with expression less than half of the cross-species average, we find that there are between 55-167 genes in each species showing low expression and nearby/intronic insertions (**Table S4**). Consistent with TE-induced epigenetic silencing, these genes with reduced expression are significantly overrepresented for genes with TE insertions in almost every species, and in both ovaries and testes (**Figure 3H**).

To determine whether TE insertions are inducing epigenetic silencing of nearby genes in some of these cases, we selected on one of the more significantly downregulated genes, *CG12768*, which has an insertion in the first intron (**Figure 3I**) and shows the lowest expression in *D. albomican* testes (**Figure 3G**, inset). Accompanying its low expression in *D. albomicans*, we find elevated enrichment of H3K9me3 at the intronic insertion as well as across the gene body, exons and 5' region (see below for ChIP analysis). Notably, this insertion did not appear to completely silence the gene, as abundant RNA-seq reads still map to the second exon, albeit substantially lower than other species (**Figure S14**).

Interestingly, TE insertions are not just associated with highly downregulated genes: we find that highly upregulated genes in a species (>2-fold higher than species mean) can also be significantly over-represented by genes with insertions. While not significant in all species, up-regulated genes have proportionally more TE insertions in all comparisons (**Figure 3G**). For example, the gene *Gyc88E* in *D. albomicans* has an intronic insertion in the first exon and is the highest expressed orthologs in the testes (2.16-fold higher than the next highest; **Figure S15**). Therefore, TE insertions appear to be associated with increased expression divergence through both down- and up-regulation of nearby genes.

H3K9me3 spreading around TE insertions near genes

To evaluate the extent to which epigenetic silencing of TEs can lead to reduction in expression of neighboring genes, we analyzed available ChIP-seq data for the repressive heterochromatic histone modification H3K9me3 in *D. albomicans* male 3rd instar larvae (Wei and Bachtrog 2019). We examined the extent of H3K9me3 spreading from TE insertions with different distances to the closest gene; to avoid TEs inside the pericentromeric or telomeric heterochromatin, we analyzed only those >5Mb from the chromosome ends. Insertions over 5kb from genes show the highest H3K9me3 enrichment in neighboring regions (**Figure 3A**, top). TEs that are closer to genes (within 5kb of genes), on the other hand, show lower levels of heterochromatin spreading. Less heterochromatin spreading from TE insertions nearby genes is consistent with opposing effects of heterochromatin formation and gene expression; transcriptionally active chromatin near genes may impede the spreading of silencing heterochromatin. Looking more closely, we find that high H3K9me3 enrichment is observed in the immediate vicinity up and downstream of the insertions and quickly drops off within 100bp (**Figure 4A**, bottom). Interestingly, this rapid decline from highly elevated H3K9me3 enrichment is observed regardless of insertion distance. Therefore, despite a narrower spreading range of TEs close to genes, the silencing effect in the immediate vicinity is similar to those far from genes, and may explain the paucity of insertions within 100bp of genes (**Figure 2E**) and exons (**Figure S9A**).

To address whether heterochromatin spreading from TEs reduces expression of nearby genes, we evaluated the extent of H3K9me3 enrichment surrounding TE insertions that are nearby genes with different expression levels in testes. Insertions were partitioned by their proximity to genes with low (< 8TPM) and high expression (>8 TPM). Insertions around low TPM genes show a higher H3K9me3 enrichment and spreading than those around high TPM genes (**Figure 4B**, top). While these differences are consistent with epigenetic silencing of genes induced by neighboring TEs, they could also reflect high transcriptional activity opposing heterochromatin spreading from nearby TEs. Given the lack of systematic downregulation between genes with insertions and their orthologs (**Figure 2F**), yet overrepresentation of TE insertions in genes that are downregulated (**Figure 3H**), our data suggest that both forces are at play.

Recurrent and rampant amplifications of DINEs

The most abundant TE, accounting for 2.1-3.8 Mb across all the species, is a 770 bp repeat which shows homology to the Drosophila INterspersed Element (DINE) - a non-autonomous DNA transposon that is highly species-specific (Locke et al. 1999). DINE's are widespread in the *Drosophila* genus, with hundreds to thousands of copies identified across a wide range of *Drosophila* species (Yang and Barbash 2008). They appear particularly abundant in the *nasuta* species complex, with 1501-3202 full length and 4863-6793 truncated DINE insertions identified across species.

Phylogenetic analysis of individual TE insertions can reveal about their evolutionary history, including the timing of when a particular TE likely was transcriptionally active. To study the explosion of DINE elements in the *nasuta* species group, we determined their phylogenetic relationship, using near-full length copies with the addition of insertions found in the *D. immigrans* genome as the outgroup (**Figure 5A**). We find a complex phylogenetic tree where the majority of DINEs do not show species-specific clustering. Instead, insertions from different species in the *nasuta* subgroup are highly intermingled, indicating that the bulk of DINE amplification predated the radiation of this species complex (**Figure 5A**). Most of the elements are likely currently inactive given the lack of species-specific clusters and long terminal branches (**Figure S16**).

While most DINEs in the *nasuta* subgroup likely originated from old expansion events, we nevertheless identified multiple instances of species-specific clustering. First, we find that the *D. immigrans* DINEs form a monophyletic clade with short branch lengths, suggesting a relatively recent, *immigrans*-specific expansion of this element. Second, we identified multiple clusters of *D. pallidifrons* insertions throughout the tree, including one large branch containing 142 out of 400 (subsampled) DINE insertions. *D. pallidifrons* DINEs within this branch contain several distinct clusters with short branch lengths, suggesting that multiple copies of DINE are currently (or have been recently) amplifying in the genome (**Figure 5B**). Expansions of DINE in *D. pallidifrons* and *D. immigrans* are consistent with a small number of elements (if not a single copy) escaping silencing, which subsequently generated a large number of insertions. Interestingly, multiple smaller clusters of *D. pallidifrons* DINE expansion (**Figure 5,** green arrows) are also found in distant branches across the phylogeny, suggesting that other DINE lineages may have reactivated (see discussion). Lastly, though less obvious, smaller scale copy number increases of DINE can also be observed in other species, such as the large numbers of *D. albomicans*, *D. nasuta* and *D. kepulauana* DINEs within the *D. pallidifrons* cluster that suggest both species-specific insertion events as well as older insertions events in their common ancestor. Similarly, small scale expansion events are also observed for the *sulfurigaster* species complex.

To better understand the sequence changes that may have precipitated the expansions, we first generated consensus sequences for DINEs in *D. immigrans*, across the *nasuta* subgroup, and in specifically the *D. pallidifrons* cluster (**Figure 5B**) from the *D. pallidifrons* genome. We then compared them to the previously reported consensus sequences from other *Drosophila* species (**Figure 5C**). While DINEs are between 300-400bps in the other species, they double to 695 and 726 bp in *D. immigrans* and the *nasuta* group, respectively. However, they still contain many of the main features such as the presence of sub-terminal inverted repeats, microsatellite regions consisting of variable lengths of simple repeats and 3' stem loop. Conservation can be found across the core sequence near the 5'. Nearly all the sequence length increase can be found in the middle disordered region where alignment is poor even between *D. virilis* and *D. melanogaster*. We note that there are several indels and SNPs that differentiate between consensus from the *nasuta* group consensus and the *pallidifrons* cluster. However, many of these mutations are found in DINEs that are outside of the expanded clusters.

Frequent expansion likely due to suppression escape

Given the pattern of proliferation of the DINEs, we were curious as to the frequency in which TEs can escape suppression and expand. We therefore generate phylogenetic trees of 147 TEs where we can find more than 20 copies across all seven species; expansions were identified as branches showing significant lineage and/or species-specific clustering (**Figure 6A-D**). We find that 78 TEs show significant species-specific clustering in at least one species, suggesting TE proliferation occurs frequently in different species (**Figure 6A**). In most cases, individual TE expansions do not reach beyond 50 copies. Expansion occurs across all types of elements although in different ways (**Figure 6A**). For example, for a variant of the Gypsy LTR retrotransposon, expansions are observed in four species as well as prior to the *sulfurigaster* semi-species split (**Figure 6B**). In contrast, for Merlin, a DNA transposon, expansions are observed in *D. pallidifrons* and *D. nasuta* and prior to the *D. albomicans/D. nasuta /D. kepulauana* species split (**Figure 6C**). Lastly, a rolling circle element expanded in *D. pallidifrons* and two of the *sulfurigaster* species (**Figure 6D**). Strikingly, there are 47 expanded TE families in *D. pallidifrons* which accounts for its higher repeat content compared to the other species (**Figure 6C-E**) and may suggest increased tolerance to TE load and/or reduced genomic defense.

To determine whether these expansions resulted from escape of transcriptional and post-transcriptional silencing, we examine TE expression from the testes and ovaries in five species. Cross-species comparisons revealed that TEs frequently show elevated expression accompanying their expansion (**Figure 6E**). Out of those that have expanded, 46 TE families (58.9%) show the highest expression in the species in which the expansion occurred, significantly higher than the random expectation of 24 (**Figure 6E;** $p < 0.00002$, permutation testing, see Materials and Methods). However, this is not always the case; for example, while DINE shows recurrent and recent expansions in *D. pallidifrons* (**Figure 5A**), it is expressed at intermediate levels in this species (**Figure S16**). Interestingly, we also find at least 15 instances where the TE family is the most lowly expressed in the species in which it expanded; we suspect these may reflect successful suppression mechanisms that evolved after expansion.

In Drosophila, the activity of TEs and their silencing systems can both differ between the sexes (Chen et al. 2021). Across all species, TE expression in testes is higher than in ovaries, suggesting weaker silencing in the testes. Curiously, expression of expanded TEs in *D. pallidifrons* are on average 20.70-fold higher in testes compared to ovaries. This is significantly higher than unexpanded TEs which are only 3.16-fold higher in the testes ($p = 0.0464$). This striking difference suggests that the numerous TEs that have expanded in *D. pallidifrons* may be exploiting the male germline for amplification which is consistent with our observation that insertions are found more frequently around genes with higher expression in testes compared to ovaries (**Figure 3C**).

Epigenetic silencing of expanded TEs moderately reduces expression in neighboring genes

Even though expanded TEs are typically highly expressed when compared to other species, several expanded TEs show low to no expression. We hypothesized that the lowly expressed expanded TEs may have been historically active elements that are now silenced. To evaluate this possibility, we looked at expression of genes neighboring these expanded TEs, reasoning that silencing of TEs will likely lead to reduced expression of neighboring genes.

We identified genes with nearby TE insertions (internal or +/- 1kb up- and downstream), and subdivided them into those with insertions of highly vs. lowly expressed expanded TEs. We focused on *D. pallidifrons* as it has the highest number of expanded TEs, and identified 182 and 552 genes with expanded lowly expressed and expanded highly expressed nearby TEs, respectively. Interestingly, the expression of the former set (genes nearby highly expressed expanded TEs) are significantly higher than those of the latter (genes nearby lowly expressed expanded TEs; **Figure 7A**, p-value $< 3.5392 \times 10^{-16}$, Wilcoxon Rank Sum Test). This is consistent with the notion that silencing of expanded TEs is associated with lower expression of nearby genes.

To differentiate between insertions/survival bias near lowly expressed genes versus bona fide spreading of epigenetic silencing into neighboring genes, we again compared the expression of the orthologs of these genes between species. To sensitively detect potential down regulation, for each gene, we scaled the expression of the *D. pallidifrons* ortholog relative to the most highly expressed ortholog. For genes with no expanded TEs around them (**Figure 7B**, gray), the *D. pallidifrons* orthologs, expectedly, have a median relative expression of 0.50. Although not significantly different, genes with highly expressed expanded TEs nearby show a slightly higher median expression and are slightly skewed towards higher expression (**Figure 7B**, dark yellow). On the other hand, genes near lowly expressed expanded TEs (i.e. near those TEs that are putatively silenced) show a low relative expression of 0.37 (**Figure 7B**, light yellow). These genes show a clear skew towards low to no expression, and are significantly lower ranked than both the control set of genes (no expanded TEs nearby) and genes near highly expressed TEs (**Figure 7B**, light yellow; p= 3.70e-12 and 5.17e-05, Wilcoxon's Rank Sum Test). These results reveal that insertions of recently expanded TEs can cause a subtle but significant decrease in gene expression if inserted nearby, but only if the TEs are targeted for (presumably epigenetic) silencing. However, if a recently expanded TE is not being targeted for silencing, it may potentially induce higher expression of neighboring genes.

We used our H3K9me3 ChIP data in *D. albomicans* to further evaluate whether this effect is due to epigenetic silencing. We plotted H3K9me3 enrichment around TEs with elevated expression and TEs with low expression in *D. albomicans*, removing insertions in the pericentric regions (**Figure 7C**). Consistent with epigenetic spreading at putatively silenced TEs, we find that TEs with low expression show substantially higher H3K9me3 enrichment in surrounding regions, with both elevated and wider spreading of heterochromatin. More highly expressed TEs, in contrast, show substantially less enrichment and spreading of H3K9me3. Therefore, lowly expressed TEs are likely under stronger epigenetic silencing which leads to broader spreading of H3K9me3.

Discussion

Here, we generated repeat-rich genomes of seven closely related Drosophila species, taking advantage of long read sequencing technologies. Enabled by these high quality genome assemblies, we systematically characterized the landscape of TE insertions and evaluated how their activities and regulation influence genome evolution. Specifically, we focused on two questions: how often do TEs influence gene regulation and how common do TEs escape silencing and expand in copy number?

The regulatory impact of TE insertions on gene expression

There are numerous examples of TE insertions affecting expression of neighboring genes, some of which even confer adaptive phenotypes (Casacuberta and González 2013; Mateo, Ullastres, and González 2014; Merenciano et al. 2016; Villanueva-Cañas et al. 2019). However, insertions around genes are primarily thought to be deleterious as they can induce epigenetic silencing of neighboring genes through heterochromatin spreading (Choi and Lee 2020). Here we comprehensively evaluate such an effect in a comparative genomics framework by combining high confidence TE insertion calls from de novo genome assemblies with gene expression data across a group of recently diverged species, the *nasuta* species group. While TE insertions are found more frequently near lowly expressed genes, TE-induced silencing does not appear to be a major cause of this negative association. The vast majority of genes with insertions around them do not show lower expression compared to other species. Therefore, instead of TEs causing nearby down-regulation, it appears that they tend to accumulate and repeatedly insert near historically lowly expressed genes. The fact that independent insertion patterns are positively correlated between species suggest that two types of non-mutually exclusive biases could be at play. TEs may preferentially insert into specific regions, chromatin environments, or gene features resulting in similar insertion patterns between species. This alone is unlikely to fully account for the negative association between gene expression and insertion counts. We, therefore, suspect that the observed insertion landscape also reflects a survivorship bias; insertions with high fitness costs are unlikely to reach high population frequency, therefore most of the observable insertions in the genome will be those with low fitness impacts. Unlike highly expressed genes, such as housekeeping genes that are under strong negative selection, lowly expressed genes may be more permissive to fluctuations in gene expression.

TEs are underrepresented near highly expressed genes, yet most TE insertions identified in our genomes do not appear to alter gene expression (**Figures 2F** and **3G**). If the observed TE insertions rarely influence gene expression, then how could they be more deleterious when inserted near highly expressed genes? One possible solution to this apparent paradox may be that the regulatory effects of TEs become more substantial upon environmental perturbations (Capy et al. 2000). In plants, multiple classes of retrotransposons are activated upon stresses (Wessler 1996; Grandbastien et al. 1997), and in flies and worms, TEs increase in activity during elevated temperatures (M. G. Kidwell, Kidwell, and Sved 1977; Garza et al. 1991; Ratner et al. 1992; Kurhanewicz et al. 2020). The lack of expression change in genes with TEs inserted nearby may therefore be the product of maintaining stocks in stable lab conditions. But upon environmental perturbation, these genes might begin to show more drastic regulatory changes as TEs become active. In changing environmental conditions, insertions around highly expressed and functionally important genes may therefore be under strong negative selection, accounting for the negative association between gene expression and insertions.

Context dependent heterochromatin spreading and epigenetic silencing

Despite no systematic support for widespread downregulation of genes with TEs inserted nearby, we were able to find evidence of epigenetic silencing of genes due to insertions in some cases. In *D. pallidifrons*, insertions of recently expanded TEs can cause moderate down regulation of gene expression, but only if the TEs have low expression - presumably due to epigenetic silencing (**Figure 7A** and **B**). Moreover, in every species a few dozens of species-specific TE insertions appear to be associated with down-regulation of nearby genes (**Figure 3G** and **H, Table S4**). Notably, we also find cases where insertions are associated with large up-regulation in gene expression (**Figure 3G** and **H, Table S4**), but these cases are much rarer than those associated with down-regulation of nearby genes.

While we do not have direct evidence of transcriptional or post-transcriptional silencing in most species, clear spreading of heterochromatin from TE insertions is observed in *D. albomicans* (**Figure 4**). TEs far from genes show the highest and broadest H3K9me3 enrichment, and TE insertions near lowly expressed genes also show more heterochromatin spreading. While consistent with epigenetic silencing of neighboring genes, these results are also consistent with the notion that active transcription antagonizes heterochromatin formation, and vice versa. Lower expression of genes near TEs that show higher levels of heterochromatin spreading could indicate that H3K9me3- inducing TEs are more tolerated near lowly expressed genes. Further, we find that insertions of TE families with low expression are associated with broader and stronger heterochromatin spreading to their surroundings. Indeed, lowly expressed and high copy number TEs are typically recently active and have robust small RNA targeting for post-transcriptional degradation and transcriptional silencing (Wei, Chan, and Bachtrog 2021). Altogether, these results suggest that TE insertions can have multiple effects on gene expression and calls into question how pervasive TE-induced epigenetic silencing of neighboring genes is. The epigenetic effects TEs have on neighboring genes, if any, is likely dependent on multiple factors, such as the transcription rate of the gene, the local repeat density and the 3D architecture of the genome.

Frequent and recurrent TE expansions and silencing

Using a phylogenetic approach to understand the relationship of TE insertions, we revealed that >50% of the TE families show lineage and species-specific amplification. The most striking expansion is the DINE, which has exploded to thousands of copies across the *nasuta* species group. This expansion occurred once prior to the species radiation, and at least twice since, one in *D. pallidifrons*, and one in the related outgroup species *D. immigrans* (~20 million years diverged; Izumitani et al. 2016; O'Grady and DeSalle 2018)(**Figure 5A**). The repeated expansions suggest multiple bouts of suppression escape. Interestingly, we were unable to find unique mutations private to the *D. pallidifrons* expansion clade, which may be causal mutations allowing to avoid suppression. One possible explanation for the absence of such mutations is that gene conversion events have converted some of such nucleotides in insertions within the clade to nucleotides from other variants and vice versa, causing more polymorphic distribution of the nucleotides (Fawcett and Innan 2019). Such events have previously been shown to allow rapid adaptive changes at TE sequences co-opted for X-chromosome dosage compensation (Ellison and Bachtrog 2015). Consistent with gene conversion, there are multiple smaller scale clusters of *D. pallidifrons* DINEs all across the tree which may represent elements that acquired the causal mutations allowing for their own, albeit limited, suppression escapes. Previous analyses of DINEs across *Drosophila* have found their sequences to be species-specific (Yang and Barbash 2008), even for recently diverged species. This may be due to rapid homogenization of copies due to gene conversion events similar to what we are observing in *D. pallidifrons*.

Beyond DINEs, large fractions of TEs also show lineage specific expansions, though at much more limited scales. Most of these expanding TE families show elevated expression only in the species with the expansion, consistent with species-specific suppression escape and derepression allowing for expansion. Most strikingly, 32 families are or have been recently expanding in *D. pallidifrons*. This may in part reflect the fact that it is the least derived of our species and therefore has the longest terminal branches. However, we still find high expression for many of these expanding TEs indicating recent, and perhaps, on-going mobilizations. Why are so many TEs concurrently expanding in *D. pallidifrons*? P-element dysgenesis is caused by the absence of maternally deposited piRNAs against the P-elements, yet derepression and mobilization of TEs is not limited to P-elements (Khurana et al. 2011). Therefore, the large numbers of expanding and highly expressed TEs may be reflecting an on-going sweep of a novel TE in the species. Interestingly, we also find that a fraction of these recently expanded TEs, paradoxically, have low expression, and genes around them show reduced expression. We suspect that these are recently active TEs that are now epigenetically silenced.

The importance of horizontal transfer to the long-term survival and expansion of TEs has been pointed out multiple times in the literature (M. G. Kidwell 1992; Silva et al. 2020; Loreto, Carareto, and Capy 2008; Schaack, Gilbert, and Feschotte 2010; H.-H. Zhang et al. 2020). Horizontal transfer can allow TEs to cross species-boundaries and invade a naive genome that lacks suppressive mechanisms against this TE, where it can proliferate (Le Rouzic and Capy 2005). Once silencing mechanisms against a TE are in place, for example targeting by small RNAs, mobilization of that TE is prevented (Khurana et al. 2011). Inactive TEs will accumulate mutations, and eventually all functional copies may die, and horizontal transfer to a new lineage would allow that TE to escape extinction. Our finding of species-specific escape from TE repression for a large fraction of TE families suggests a very dynamic evolution of host genomes and their TEs. Active TEs are temporarily silenced within a lineage, but over evolutionary timescales, some copies will escape silencing in different lineages, leading to species-specific bursts in TE activity. Thus, in addition to horizontal transfer, our data suggest that escape from host suppression seems to be an important strategy allowing for the long-term survival of TEs.

Long-read genome assemblies open new doors for studying TEs

In our study, high quality genomes assembled via long reads have circumvented many of the previous challenges associated with studying TEs and repeats (Khost, Eickbush, and Larracuente 2017), and enabled high confidence annotation of TE insertions. Further, our approach of integrating phylogenetics, functional genomics, and comparative genomics have revealed a comprehensive picture of the dynamics of TE suppression escape and subsequent re-established silencing and their effects on the rest of the genome. These high quality genome assemblies will further facilitate the molecular dissection of the nucleotide changes in TEs causing suppression escape in future studies. With the rapidly decreasing cost and input material in generating these assemblies (Adams et al. 2020), it will become easier and cheaper to identify de novo insertions. But even with the rapid adoption of these technologies, TEs and repeats remain under-studied and often avoided. Instead, here we show that assembling repeats is among one of the greatest advantages to long read sequencing.

Material and Methods

Fly strains and nanopore sequencing

We extracted high molecular weight DNA from approximately 50 females from *D. nasuta* 15112-1781.00, *D. kepulauana* 15112-1761.03*, D. s. albostrigata* 15112-1771.04*, D. s. bilimbata* 15112-1821.10*, D. s. sulfurigaster* 15112-1831.01*,* and *D. pallidifrons* PN175_E-19901 using the QIAGEN Gentra Puregene Tissue Kit. The *D. kepulauana* high molecular weight DNA was sequenced on PacBio RS II platform at UC Berkley QB3 genome sequencing center.The high molecular weight DNA of other species were sequenced on Nanopore MinIOn.

Genome assemblies

Drosophila albomicans

The *D. albomicans* genome assembly has been previously published, having been generated with DNA sequenced on the PacBio RSII platform resulting in an N50 of 33.4 Mb and BUSCO score of 98%, indicating high contiguity and completeness. Here, Nanopore reads from strain 15112-1751.03 were error corrected with canu and an initial assembly was generated using wtdbg2. The assembly was then polished 3 times using 35.7x coverage Illumina paired end reads from *Mai et al.* 2019 with minimap2 and Racon followed by 1 round of Pilon. Afterwards, we BLAST the assembly against the NCBI BLAST database for potential contamination . We remove 65 contigs making up 1.4 Mb—mostly comprising *Acetobacter*. This filtered genome is then organized using HiC data and the Juicer and 3d-dna pipeline. We stitch adjacent contigs within a scaffold with a string of 50 N's. This stitched genome assembly has an N50 of 33,438,794 bp and a BUSCO score of 91.3%, notably lower than the previously published assembly. To improve upon this assembly, we use quickmerge twice with these two assemblies and twice more with the results, taking complementary information between them to improve contiguity and completeness. Due to the reference dependent asymmetry of quickmerge results, we ran the program using both the old and newly stitched genome as the reference and repeated this with the resulting genomes; we took the one with the highest contiguity and BUSCO score. We generated an even more complete assembly (BUSCO score of 99.6%), an increase in assembly size (167,541,436 bp) and improved contiguity (N50 of 35,291,776 bp).

Drosophila nasuta

The *D. nasuta* genome assembly was generated using Nanopore long read data from strain 15112-1781.00, which were error corrected with canu. The initial genome assembly was generated using wtdbg2 and polished—using 38.3x coverage Illumina paired end reads from *Mai et al.* 2019 3 times with Racon and minimap2 followed by 1 round of Pilon. We BLAST the assembly against the NCBI database for contamination and find 140 contigs making up approximately 5.57 Mb, mostly from *Acetobacter*. This filtered genome is organized with HiC data using the Juicer and 3d-dna pipeline, where we stitched adjacent contigs with a string of 50 N's. The final resulting assembly has a BUSCO score of 99.2%, assembly size of 171,781,232 bp, and an N50 of 33,885,645 bp.

Drosophila kepulauana

The *D. kepulauana* genome assembly was generated with DNA sequenced on the PacBio RSII platform data from strain 15112-1761.03—the reads were error corrected with canu. Similar to the *D.* albomicans assembly, we generated two genomes and used quickmerge to generate the final assembly. The first assembly was initially generated using wtdbg2 and polished with 38x coverage Illumina paired end reads from *Mai et al.* 2019 thrice with Racon and minimap2 followed by 1 round of Pilon. We BLAST the assembly against the NCBI database for contamination and find 87 contigs making up approximately 14.03 Mb, mostly from *Acetobacter*. This filtered genome is organized with HiC data using the Juicer and 3d-dna pipeline, where we stitched adjacent contigs with a string of 50 N's. The second assembly was initially generated with Flye. This assembly was polished and filtered for contamination (58 contigs totaling 13.29 Mb) in the same way as the first assembly. We ran quickmerge twice on the two assemblies, using each one as the reference, and twice more on the resulting assemblies. The assembly deemed as the final assembly is the most contiguous and complete one and has a BUSCO score of 99.7%, assembly size of 163,769,021 bp, and an N50 of 34,564,094 bp.

Drosophila sulfurigaster albostrigata

The *D. s. albostrigata* genome assembly was generated using nanopore sequencing data from strain 15112-1771.04, which were error corrected with canu. Just like the treatment of the *D. kepulauana* assembly, we generated two genomes and used quickmerge to generate the final assembly. The first assembly was initially generated using wtdbg2 and polished with 39.4x coverage Illumina paired end reads from *Mai et al.* 2020 Illumina paired end short reads three times with Racon and minimap2 followed by 1 round of Pilon. We BLAST the assembly against the NCBI database for contamination and find 8 contigs making up approximately 7.1 Mb, mainly from *Acetobacter*. This filtered genome is ordered with HiC data using the Juicer and 3d-dna pipeline and adjacent contigs were stitched with a string of 50 N's. The second assembly was initially generated with Flye. This assembly was polished and filtered for contamination (25 contigs totaling 7.71 Mb) in the same way as the first assembly. We ran quickmerge twice on the two assemblies, using each one as the reference, and twice more on the resulting assemblies. The assembly deemed as the final assembly is the most contiguous and complete one and has a BUSCO score of 98.5%, assembly size of 168,284,230 bp, and an N50 of 37,627,869 bp.

Drosophila sulfurigaster bilimbata

The *D. s. bilimbata* genome assembly was generated using Nanopore long read data from strain 15112-1821.10, which were error corrected with canu. The initial genome assembly was generated using wtdbg2 and polished using 17.1x coverage Illumina single end reads from *Mai et al.* 2020 Illumina single end short reads 3 times with Racon and minimap2 followed by 1 round of Pilon. We BLAST the assembly against the NCBI database for contamination and find 186 contigs totaling around 7.41 Mb, mostly from *Acetobacter*. This filtered genome is organized with HiC data using the Juicer and 3d-dna pipeline, where we stitched adjacent contigs with a string of 50 N's. The final resulting assembly has a BUSCO score of 99.7%, assembly size of 164,595,183 bp, and an N50 of 36,279,119 bp.

Drosophila sulfurigaster sulfurigaster

The *D. s. sulfurigaster* genome assembly was generated using Nanopore long read data from strain 15112-1831.01, which were error corrected with canu. An initial genome assembly was generated using Flye and polished—using 26.7x coverage Illumina paired end reads from *Mai et al.* 2019--3 times with Racon and minimap2 followed by 1 round of Pilon. We BLAST the assembly against the NCBI database for contamination and find 11 contigs totaling around 3.3 Mb, most of which were from *Acetobacter*. This filtered genome is ordered with HiC data using the Juicer and

3d-dna pipeline, where we stitched adjacent contigs with a string of 50 N's. The final resulting assembly has a BUSCO score of 99.9%, assembly size of 165,884,258 bp, and an N50 of 35,818,991 bp.

Drosophila pallidifrons

The *D. pallidifrons* genome assembly was generated using Nanopore long read data from strain PN175_E-19901, which were error corrected with canu. The initial genome assembly was generated using wtdbg2 and polished—using 17x coverage Illumina paired end reads from *Mai et al.* 2020 3 times with Racon and minimap2 followed by 1 round of Pilon. We BLAST the assembly against the NCBI database for contamination and find 59 contigs making up approximately 4.9 Mb, mostly from *Acetobacter*. This filtered genome is organized with HiC data using the Juicer and 3d-dna pipeline, where we stitched adjacent contigs with a string of 50 N's. The final resulting assembly has a BUSCO score of 99.3%, assembly size of 164,659,715 bp, and an N50 of 37,973,042 bp.

Gene annotation and clustering

We used MAKER to annotate genes in each species' genome assembly (Campbell et al. 2014). To train MAKER's gene inference model, we generated a transcriptome from *D. albomicans* and *D. nasuta* RNA seq data from Zhou and Bachtrog 2012 (Zhou and Bachtrog 2012). RNA seq data from *D. albomicans* and *D. nasuta* were aligned to the corresponding genome assemblies with HISAT2 under default settings (Kim, Langmead, and Salzberg 2015). The alignments were then used to create transcriptomes using StringTie (Pertea et al. 2015). Additionally, satellite repeats in the genome assemblies for each species were masked using RepeatMasker in preparation for gene annotations (*RepeatMasker Open-4.0* 2013). Then, using both the *D. albomicans* and *D. nasuta* transcriptomes, we ran MAKER with default settings. We then took the annotations and determined gene homology between species with OrthoDB (Kriventseva et al. 2019).

TE library generation, annotation and analyses

In order to lower the occurrence of nested TE structures, pericentromeric regions were removed from each and the resulting sequences were separately used as input for RepeatModeler2 and the accompanying LTRharvest software with default options (Ellinghaus, Kurtz, and Willhoeft 2008; Flynn et al. 2020). The resulting species-specific TE libraries were merged together. To remove redundancy from the merged library, we used CD-hit2 to cluster TE entries with each. However, instead of allowing CD-hit2 to select the representative sequence of the cluster (which is usually the longest sequence), we evaluated the TEs within clusters based on three criteria: entry sequence length, self-identity, and probability of full length insertions. For self-identity, we blasted each TE entry to itself and calculated the self-blast score as the proportion of the sequence showing alignment to another region of itself. For probability of full length insertions, we blasted each entry to the genome and proportion of near full length blast hits. We then weighed the three criteria to maximize length and probability of full length insertion, while minimizing self-identity, in order to select the representative sequence per CD-hit2 cluster. This procedure is done twice. We used both the Repeatmodeler2 TEs categorization as well as the program ClassifyTE (Panta et al. 2021). When the two disagreed with the TE classification, we used the assignment from RepeatModeler2. Note, even after two round of CD-hit2 we found 10 redundant entries corresponding to variants of the DINE in the genome through manual NCBI BLASTn (Altschul et al. 1990). We removed entries with unique sequences flanking the DINEs and kept the longest entry.

TE insertions in genomes are annotated by RepeatMasking the final nasuta group-specific TE index to the respective species genomes. Because RepeatMasker can provide overlapping annotations, we used bedtools merge to merge overlapping annotations first, generating chimerics. We then blasted all the chimeric annotations to the repeat library and recategorized those where 90% of the sequence blasts to a specific TE. Full length and truncated elements are defined as annotations that are >80% length of the TE entries, or <80% length but >200 bp, respectively. Distances between full length TE and the closest gene in each species were calculated using bedtools closest, (Quinlan and Hall 2010) with species-specific TE and gene annotations as inputs.

Phylogenetic analyses of TEs

We ran BLAST using the TE libraries as the query and the genome assemblies of each species as the database (Altschul et al. 1990). TE sequences from full length BLAST alignments--defined as those in which the alignment length is at least 80% of the TE length--are extracted. We used Clustal Omega under default settings to perform a multiple sequence alignment for all sequences for each TE (Sievers et al. 2011); those with over 200 full length copies across all species were subsampled down to 200 sequences. In order to maintain the different copy number in the different species, the subsampling procedure maintained the proportional difference of insertion counts across the species.

Phylogenies for TEs were then generated with RAxML using the command: raxmlHPC-PTHREADS-AVX -T 24 -f a -x 1255 -p 555 -# 100 -m GTRGAMMA -s input.MSA.fa -n input.MSA.tree > input.MSA.tree.stderr (Stamatakis 2014).

We tested for the presence of species specific expansion of each TE by measuring the extent of clustering using the RRphylo R package (Serio et al. 2019). Tests were carried out for species where there were at least 5 sequences or 5% of the total sequences in the phylogeny. The resulting p-values from the analyses were adjusted for multiple testing using the Benjamini-Hochberg procedure. TEs from a particular species with p-values < 0.05 are considered to be expanded. We note that this program does not take into account the species relationships and therefore cannot capture lineage-specific expansion. Thus this approach under-estimates the number of TEs that have recently expanded.

RNA sample collection and sequencing

Two replicates of RNA sequencing libraries created from males and females of each species were generated and sequenced. Testes from five to eight males from live *D. albomicans, D. nasuta, D. kepulauana,* and *D. s. sulfurigaster* as well as frozen *D. pallidifrons* were dissected for each RNA sequencing library. Ovaries from three to five females from live *D. albomicans, D. nasuta, D. kepulauana,* and *D. s. sulfurigaster* as well as frozen *D. pallidifrons* were dissected for each RNA sequencing library. For each species, tissue samples were placed in Trizol for RNA extraction. RNA was extracted using the Trizol extraction method and enriched for ploytA RNA using NEBNext Poly(A) mRNA Magnetic Isolation Module (E7490) as per manufacturer protocol. The RNA libraries prepared as per NEBNext Ultra II Directional RNA kit (E7760S) and sequenced on illumina NovSeq 6000 on SP flow cell for 150 PE reads.

RNA transcript abundance

Genes

Generated RNA sequencing data for *D. albomicans, D. nasuta, D. kepulauana, D. s. sulfurigaster,* and *D. pallidifrons* were aligned to their corresponding genome assembly. Using the alignment data and gene annotations, we used the featureCounts program from the Subread package to calculate the number of reads mapping to each gene. We then calculated gene transcripts per million with the following formula:

$$TPM = \frac{GeneReadCount/GeneLength \times 1000}{\sum \quad (GeneReadCount/GeneLength * 1000 \div 1000000)}$$

Transposable Elements

Generated RNA sequencing data for *D. albomicans, D. nasuta, D. kepulauana, D. s. sulfurigaster,* and *D. pallidifrons* were aligned to the TE library. A custom script was used to count the number of reads mapping to each transposable element. The number of reads was then normalized by the TE length and then divided by the median of gene read counts that are normalized in the same way from the corresponding species.

Permutation testing of TE expression

The test statistic used for the permutation test is the number of times the highest expression for a particular TE comes from a species where that TE has expanded. We first calculate this test statistic from our data. We then randomly shuffle the species associated with each expansion event and calculate the test statistic 50,000 times. The p-value obtained is the proportion of tests with test statistics less than or equal to our original test statistic.

TE expression comparisons

We categorize whether TEs are highly expressed or lowly expressed upon obtaining normalized TE expression level for testes and ovaries across species. A TE is considered to be highly expressed in a species for a particular tissue if the expression level of the TE is in the top two most highly expressed TE across species in that tissue. A TE is considered to be lowly expressed if it, instead, is in the bottom two most lowly expressed TE across species.

To compare expression between highly expressed TEs and lowly expressed TEs within a species, we first scale a TE's $\log_2$ expression to its maximum $\log_2$ expression across species:

$$scaled\ expression\ = \frac{log_2((species\_expression\ +\ 1)/(minimum\_expression\_of\_te\ +\ 1))}{log_2((maximum\_expression\_of\_te\ +\ 1)/(minimum\_expression\_of\_te\ +\ 1))}$$

The addition of 1 to the values are done to prevent potential division by zero. We then perform the Wilcoxon rank sum test between scaled expression of highly expressed TEs and scaled expression of lowly expressed TE to determine if scaled expression between TEs from different categories are statistically significant.

ChIP-seq analyses

ChIP-seq analyses were slightly modified from methods in (Wei et al. 2021). Briefly, larval H3K9me3 ChIP and input data (Wei and Bachtrog 2019) were aligned to the genome using bwa mem. The per base pair coverage was determined using bedtools coverageBed -d -ibam. Median autosomal coverage was estimated in from 50kb non-overlapping sliding windows. We then inferred enrichment at every position as:

$$enrichment\ = \frac{ChIP\ coverage\ /\ median\ autosomal\ ChIP\ coverage\ +\ 0.01}{Input\ coverage\ /\ median\ autosomal\ input\ coverage\ +\ 0.01}$$

We averaged the enrichment across the three replicates. For H3K9me3 spreading around TE insertions, we lined up annotated TE insertions at either the 5' or 3', and averaged enrichment 5kb upstream and downstream of the insertions, respectively.

Table 1. Size of genome assemblies (and their chromosomes) for each species and their associated summary statistics.

	D. albomicans	*D. nasuta*	*D. kepulauana*	*D. s. albostrigata*	*D. s. bilimbata*	*D. s. sulfurigaster*	*D. pallidifrons*
Chromosome	size (bp)	size (bp)	size (bp)	size (bp)	size (bp)	size (bp)	size (bp)
Muller A	33,597,023	33,189,490	33,291,615	33,386,403	33,007,321	33,493,557	32,839,655
Muller B	30,469,903	28,690,150	31,604,248	30,983,057	30,102,783	29,954,194	29,503,240
Muller CD	55,495,487	55,283,848	55,283,860	56,209,854	54,959,191	55,186,638	55,584,481
Muller E	35,291,776	33,885,645	34,564,094	37,627,869	36,279,119	35,818,991	37,973,042
Muller F	1,839,965	2,061,818	1,552,407	2,495,194	2,423,155	2,918,623	3,067,770
Chromosome total	156,694,154	153,110,951	156,296,224	160,702,377	156,771,569	157,372,003	158,968,188
Assembly total	167,541,436	171,781,232	163,769,021	168,284,230	164,595,183	168,070,293	164,659,715
N50	35,291,776	33,885,645	34,564,094	37,627,869	36,279,119	35,818,991	37,973,042
Number of Scaffolds*	220	282	77	95	201	123	104
BUSCO**	99.62%	99.16%	99.72%	98.50%	99.72%	98.87%	99.62%
Repeat content†	0.2070668834	0.2325837493	0.1875276888	0.2127412711	0.1978890658	0.1960036031	0.2009619839
Annotated genesd	12,395	12,492	12,594	12,595	12,432	12,718	12,362

*See Figure S1-7 for Hi-C scaffolding of the chromosome arms in each species

**See Table S2 for detailed BUSCO statistics

†See Table S3 for Repeat content and masking details

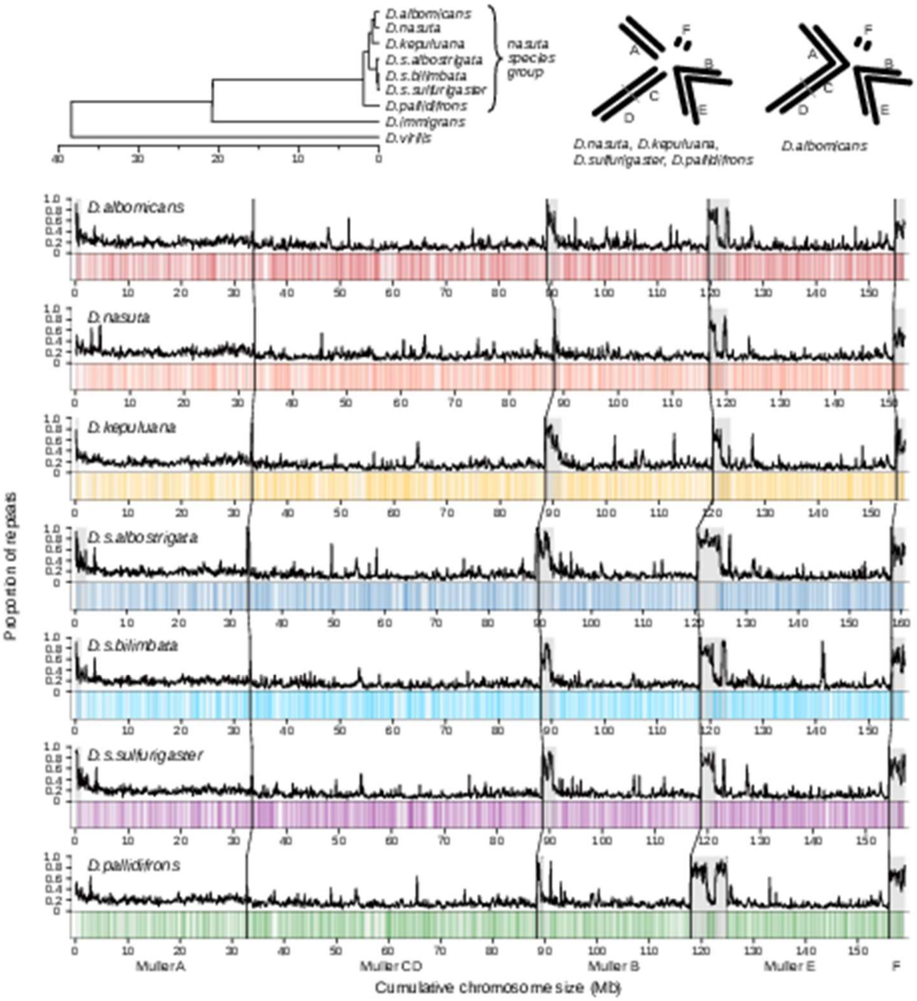


Figure 1. Genomes of the Drosophila nasuta species group. A. Phylogeny of the *nasuta* species radiation within the *Drosophila* subgenus. Tree adapted from (Mai et al. 2020) and (Izumitani et al. 2016) B. Karyotypes of the species group; chromosomes are oriented such that centromeres are pointed towards the center of circle. C. Long read-based genome assemblies of seven species. For each species, the top track depicts the repeat content estimated for 100kb windows. Positions of annotate genes are represented on the bottom track as vertical lines. The centromeric end are on the left side of each chromosome. Regions deemed as pericentromeric are highlighted in gray. Chromosomes are demarcated by black vertical lines. Unless otherwise stated, species are represented by colors used here: red (*D. albomicans*), orange (*D. nasuta*), yellow (*D. kepuluana*), navy (*D. s. albostrigata*), purple (*D. s. bilimbata*), purple (*D. s. sulfurigaster*), and green (*D. pallidifrons*).

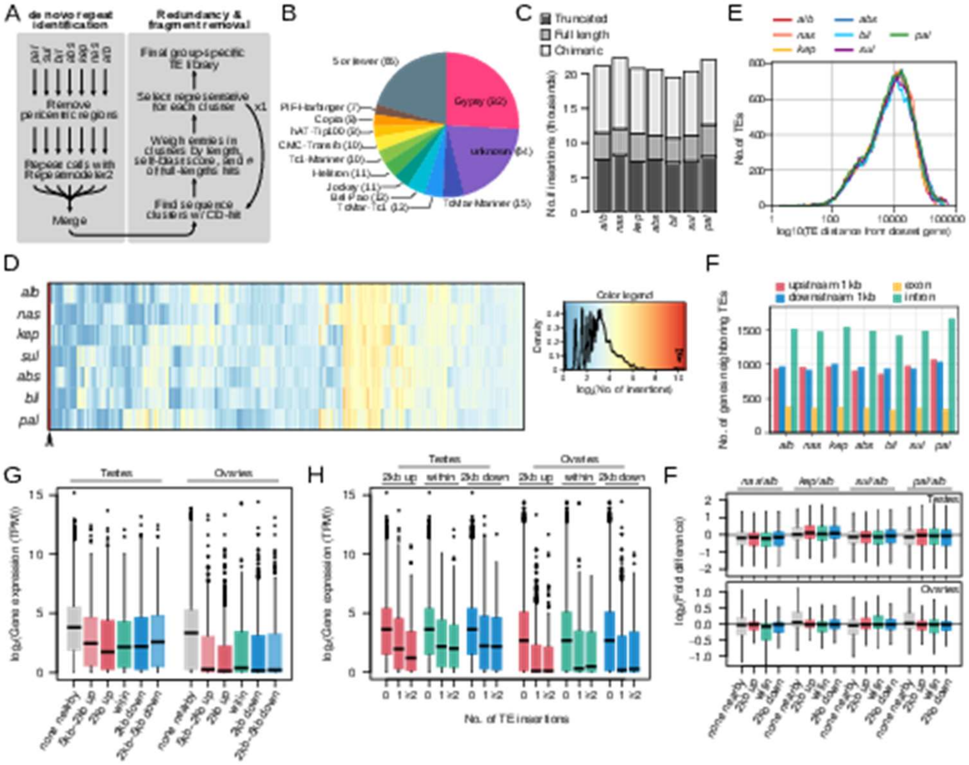


Figure 2. De novo identification and distribution of TE insertions across the genomes. A. Pipeline to construct and refine de novo TE reference from genome assemblies. We used RepeatModeler2 to first identify repeats from the euchromatic regions of each species. The resulting repeat libraries are merged followed by sequence clustering with CD-HIT. Multiple indexes were used to select the full length representative TEs. B. Breakdown of TE classes identified; for breakdown of the grey section see supplementary figure S9. C. Number of full length and truncated insertions found in each genome. The chimeric class represents the merger of annotations that overlap or are contiguous. D. Copy number of full length insertions of 318 TE families across the seven genomes. E. Distribution of the distance between TEs and genes across the species. Intergenic insertions are not counted. See supplementary figure S9 for distribution of intergenic insertions from exons . F. Number of genes with TEs inserted in different regions of genes with and without insertions. G. Transcript abundance of annotated genes in Transcripts per million (TPM), subsetted into different classes depending on where TE insertions are found. H. Transcript abundance of genes with different numbers of TE insertions. I. Fold-difference in transcript abundance of orthologous genes depending on different numbers of insertions in *D. albomicans*. See supplementary figure S10, for comparisons using insertions in other species.

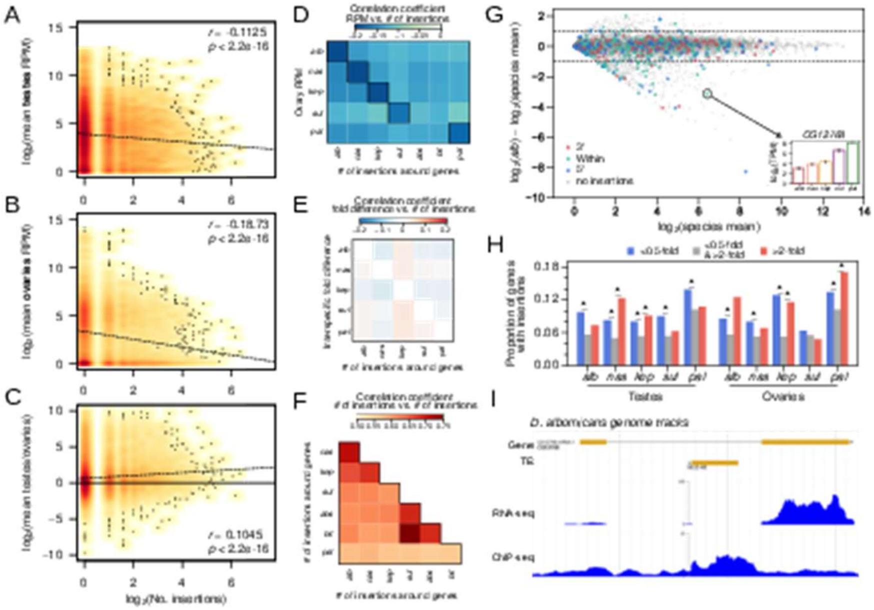


Figure 3. Negative association between TE insertions and genic expression. A-B. Density scatterplots of number of unique (both full length and truncated) TE insertions around genes (±2kb) across all the *nasuta* species genomes plotted against genic transcript abundances (averaged across the species) in the ovaries (A) and testes (B). Increased intensity of warm colors indicate higher density of points. Scattered black dots indicate positions of single points. Regression lines are depicted by dotted lines; the Pearson's correlation coefficients and corresponding p-values are labeled in the top right. C. Same as A and B, but with the fold difference of genic expression between testes and ovaries. D. Pairwise correlation of TE insertion counts around genes in a particular species to the ovarian transcript abundance of the gene orthologs in another species. E. Pairwise correlation of TE insertion counts around orthologous genes across species; genes with no insertions in either species are not used. G. MA-plot of average gene expression (TPM) across species in the testes (x-axis) plotted against fold difference between the D. albomicans expression and the average across species (y-axis). Colored points represent genes with TE insertions in different parts of the gene. Horizontal dotted line demarcates 0.5- and 2-fold differences. Inset shows the testes expression of the CG12768 across all five species. For MA-plots in ovaries and other species, see supplementary figure 13. H. Proportions of genes with TE insertions with low and high gene expression relative the the species average (i.e genes below or above the dotted lines in A), for each species and in ovaries and testes. I. Genome browser shot of CG12768 showing tracks for gene structure, TE insertions, transcript abundance, and H3K9me3 enrichment. For genome browser shot of this gene in other species see supplementary figure.

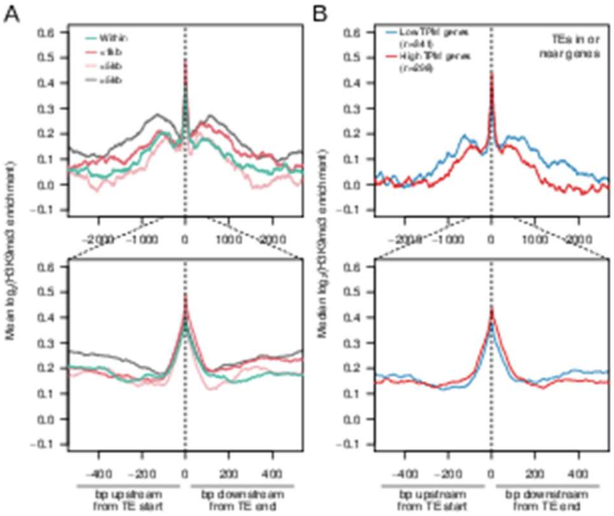


Figure 4. Epigenetic silencing through H3K9me3 spreading around TE insertions. A. Median H3K9me3 enrichment ± 5kb upstream and downstream of TEs inserted at different distances to genes (enrichment across TE insertions not plotted). TE insertions within pericentric regions are removed from analyses. Zoomed in plot (±500 bp) is shown below. B As with A but with TEs inserted within genes or <2kb around (C) genes of different expression levels.

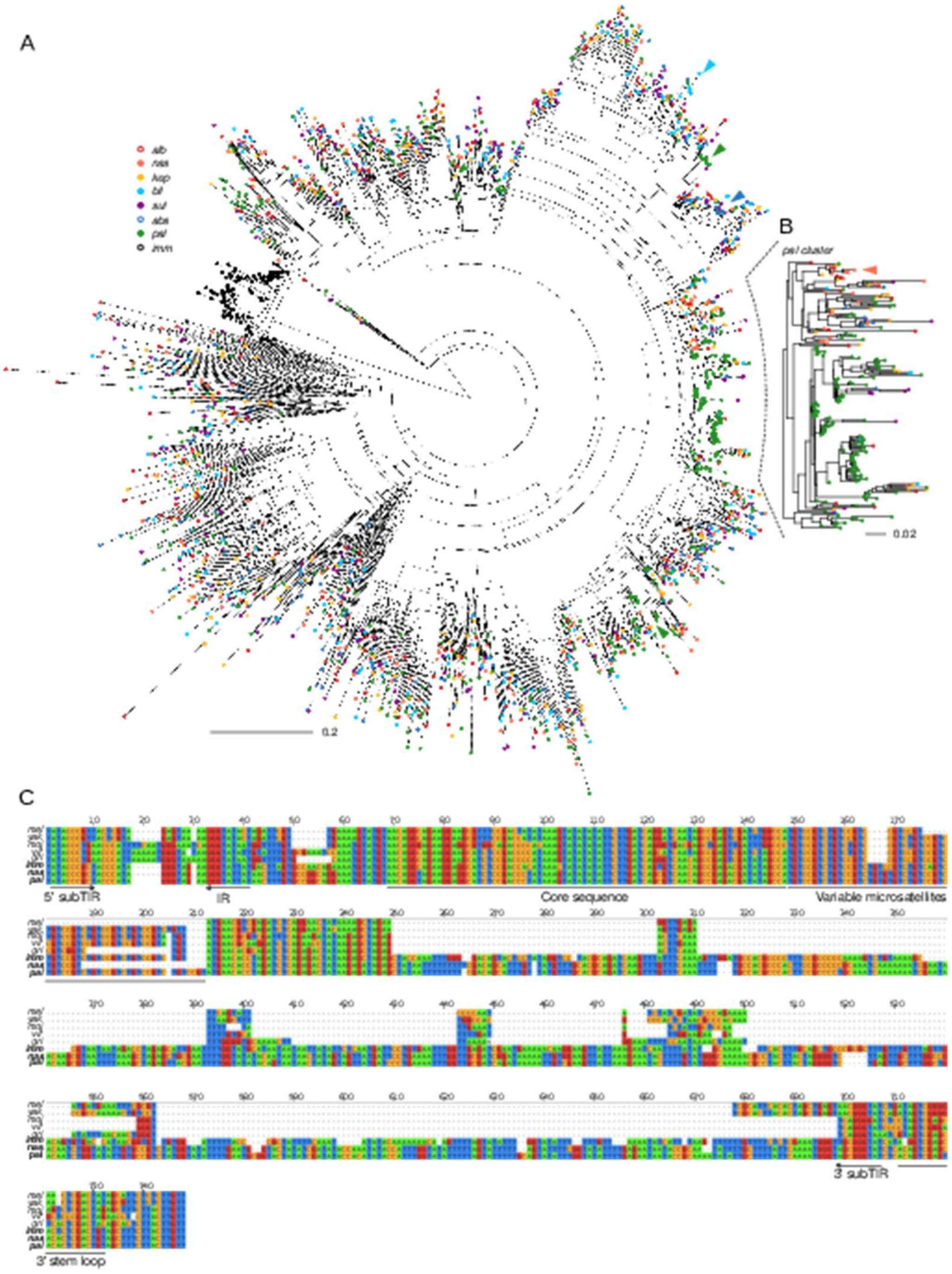


Figure 5. Recurrent DINE expansions. A. Radial tree of subsampled DINE insertions with the addition of *D. immigrans* DINE elements as outgroup. Insertions from the same species have the same colored tips. Colored arrowheads point to small scale species-specific expansions on the tree. B. Large cluster of *D. pallidifrons* DINE insertions indicate recent burst of species-specific activity. C. Multiple sequence alignments of consensus DINE sequences of representative species. DINE-specific sequence features are annotated beneath the tracks.

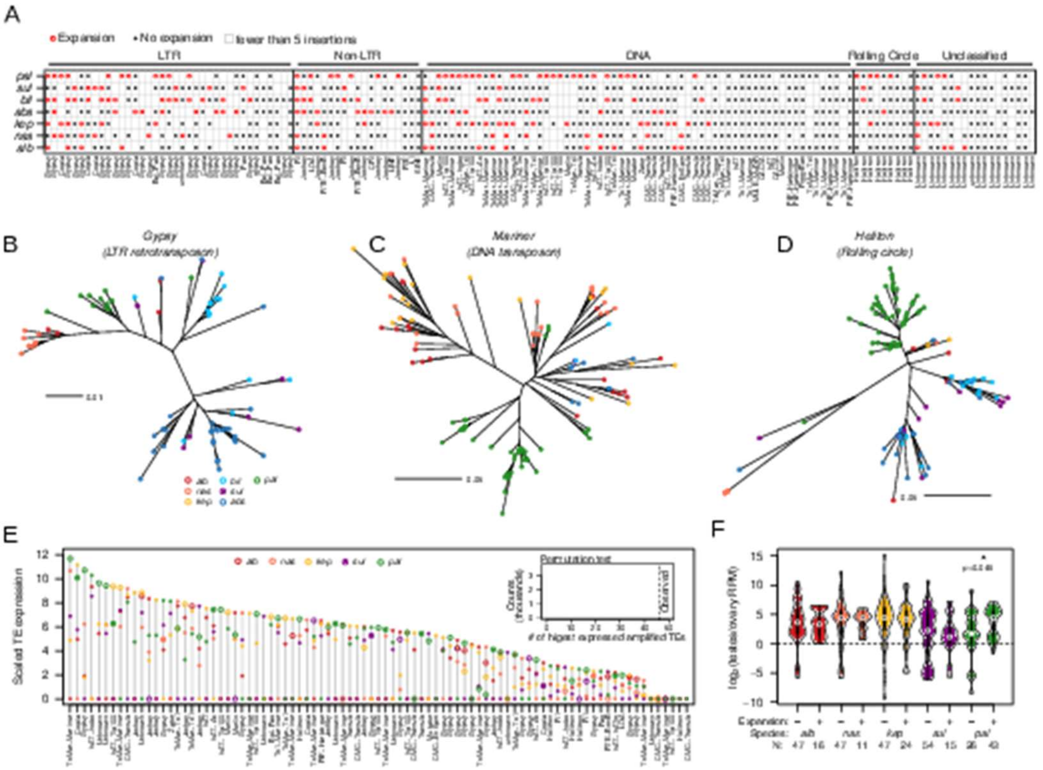


Figure 6. Frequent lineage specific amplifications and suppressions of TE families. A. Species-specific expansion status of different TE families and types based on phylogenies of insertions. Red dots indicate amplification in a *nasuta* species, black dots indicate no amplification, and empty boxes indicate fewer than 5 insertions. B-D. unrooted trees of TE insertions of different types of TEs. Their positions on the table in A are marked by arrowheads. E. Expression of expanded and unexpanded TE families in the testes of different species. For each TE family, the transcript abundance is scaled by the lowest expressed species, and the range of expression across the different species is plotted vertically as demarcated by the gray line. Along this line the expression in the different species are positioned by colored circles. Large circles denote species-specific expansion. The observed positions of the expanded TEs along the expression ranges are tested against the null expectation using randomized permutation testing (top right inset). The null distribution is presented and the observed count is marked by the vertical dotted line. F. Fold-difference in TE transcript abundance between testes and ovarian expression across species. TEs are subdivided into those that have species-specific expansions and those without.

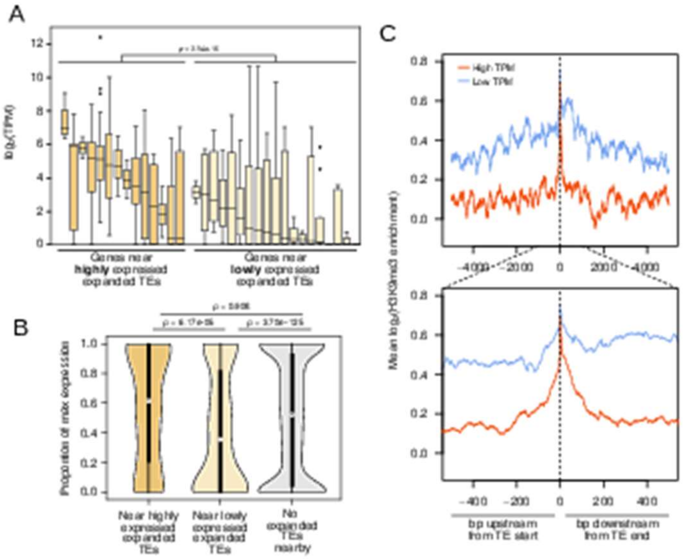


Figure 7. Epigenetic silencing of expanded TEs downregulates nearby genes. A. Expanded TEs are categorized as either highly or lowly expressed depending on expression difference between species. Dark yellow boxes represent genes nearby highly expressed expanded TEs, while light yellow boxes represent genes nearby lowly expressed expanded TEs. Each box represents the distribution of transcript abundances (TPM) of genes with nearby insertions of a given expanded TE family. Genes (n=785) near lowly expressed expanded TEs have significantly lower expression (Wilcoxon's Rank Sum Test, p < 3.54e-16). B. Scaled expression of genes near highly (dark yellow, n=82) and lowly expressed expanded TEs (light yellow, n=552), as well as those with no expanded TEs nearby (gray, n=8705). Genic expression is scaled by the TPM of the highest expressed orthologs across all species. Significance of pairwise comparisons of the three sets are labeled above the figure. C. H3K9me3 enrichment around sull length TE insertions in D. albomicans depending on whether the TE is highly expressed as compared to other species (red) vs lowly expressed (blue). Insertions within the pericentric regions are removed.

Table S1. The median long read coverage used to generate genome assemblies for each species.

species	median coverage			
D. albomicans	35.7433			
D. nasuta	38.2574			
D. kepulauana	38.0462			
D. s. albostrigata	39.4097			
D. s. bilimbata	17.1106			
D. s. sulfurigaster	26.7034			
D. pallidifrons	16.9503			

Table S2. BUSCO scores for each species including the number of genes that fall under single copy, duplicated, fragmented, and missing.

	D. albomicans		*D. nasuta*		*D. kepulauana*		*D. s. albostrigata*		*D. s. bilimbata*		*D. s. sulfurigaster*		*D. pallidifrons*	
BUSCO	Counts	Percent	Counts	Percent	Counts	Percent	Counts	Percent	Counts	Percent	Counts	Percent	Counts	Percent
Complete	1062	99.62%	1057	99.16%	1063	99.72%	1050	98.50%	1063	99.72%	1054	98.87%	1062	99.62%
Complete and single-copy	1052	98.69%	1049	98.41%	1048	98.31%	1040	97.56%	1055	98.97%	1037	97.28%	1055	98.97%
Complete and duplicated	10	0.94%	8	0.75%	15	1.41%	10	0.94%	8	0.75%	17	1.59%	7	0.66%
Fragmented	2	0.19%	6	0.56%	1	0.09%	13	1.22%	2	0.19%	8	0.75%	2	0.19%
Missing	2	0.19%	3	0.28%	2	0.19%	3	0.28%	1	0.09%	4	0.38%	2	0.19%

Table S3. The repeat content for each genome assembly.

species	assembly_size	starting_N*	total_N**	total_repeats	repeat_content
D. albomicans	167541436	400	34692683	34692283	0.2070668834
D. nasuta	171781232	450	39953973	39953523	0.2325837493
D. kepulauana	163769021	301	30711527	30711226	0.1875276888
D. s. albostrigata	168284230	200	35801201	35801001	0.2127412711
D. s. bilimbata	164595183	850	32572437	32571587	0.1978890658
D. s. sulfurigaster	168070293	1300	32943683	32942383	0.1960036031
D. pallidifrons	164659715	350	33090693	33090343	0.2009619839

* the number of masked bases before repeat masking the genome

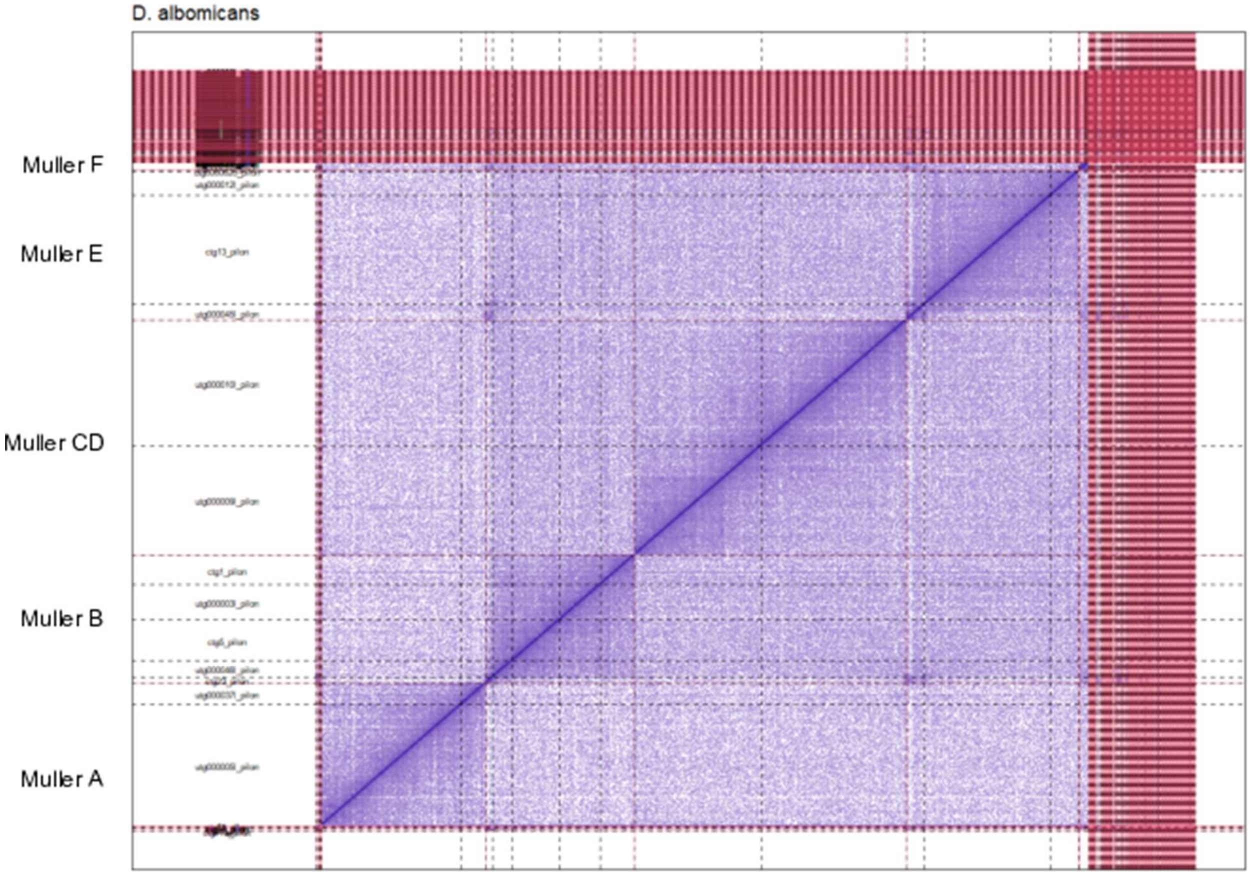
** the number of masked bases after RepeatMasker

Table S4. The number of genes that are up- or down-regulated divided into those with and without TE insertions.

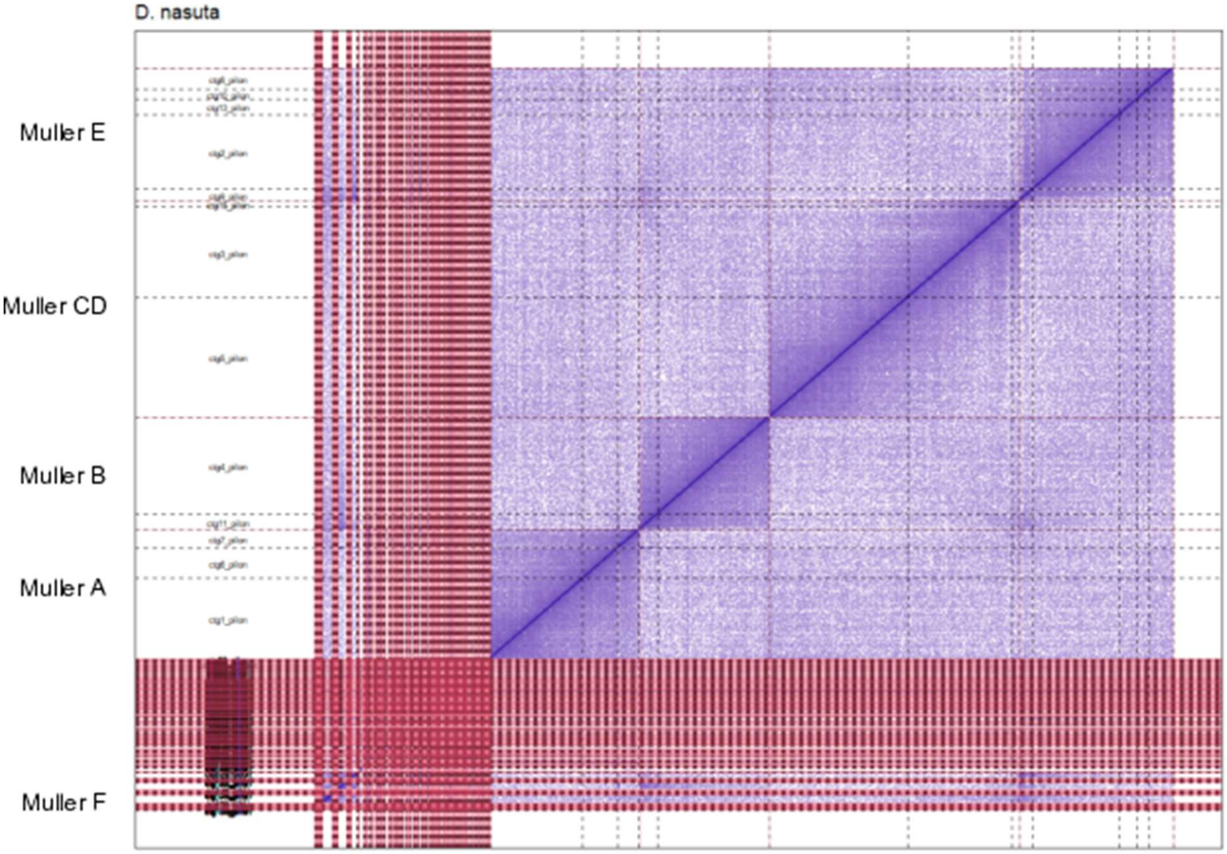
species	Up-regulated genes			Down-regulated genes*		
	Number without insertions	With insertions*		Number without insertions	With insertions*	
		Number	Gene names		Number	Gene names
D. albomicans	175	16	Hr3, Shab, bma, otk, CG17999, Apoltp, CG13086, Vm26Aa, CG14342, CG10431, Trpgamma, CG9287, SpdS, Gyc88E, unknown_102, Vsx2	589	56	CG12502, Cpr50Cb, CG12768, SmydA-2, CG14820, mspo, CG1688, CG33107.1, Hacl, CG7881, CG1299, Cpr76Bd, gsb-n, CG13321, CG44251, unknown_16, CG12341, CG8854, Amph, DCX-EMAP, Ser8, CG13086, CG3769, CG17633.2, Or22c, unknown_734, ninaD, CG15153, IA-2, CG18641, CG8419, Myb.2, CG3999, Csk, CG8369, CG16904, CG31221, e, CG6283.2, CG6283.1, E5, CG15696, ZIPIC, unknown_485, Npc2f, CG6125, OdsH, FBgn0282836, CG6106, CG11162, unknown_331, FBgn0203711, dpr18, sov.1, Tat, unknown_688
D. nasuta	228	22	toy, Ilp8, Oatp74D, ttv.2, CG32333, Ilp3.1, CG44251, CG13510, Corin, Qsox1, unknown_78, Shawl, IFT46, Pbp45, dmrt93B, Dnali1, SP1029, unknown_102, CG42564, FBgn0282836, CG3091, unknown_331	1056	89	Cadps, Sox15, unknown_1004, LRP1, CG14760, Su(var)2-HP2, wrapper, CG14820, FBgn0200328, CG13724, trpl, CG7881, mkg-p.2, Pxn, FBgn0200202, CG13920, Cpr76Bd, gsb-n, CG13058, CG44251, unknown_786, CG8888, Fancm, Mctp, CG30427, CG17999, CG3829, CAH14, CG12970, Daao1, mms4, 5-HT1A, CG42713.2, CG8834, Amph, cv-2, unknown_538, pk, Ser8, CG4480, l(2)k05911, CG44153, Mco1, CG8138.1, CG3769, CG17633.2, psd, gudu, Or22c, Rab5, Tep3, IFT57, CG31663, Oatp26F, Myb.2, nub, ush, Rfx, CG18599, 5-HT2B, AOX1, FBgn0205031, CG18528, sov.2, SP1029, unc80, CG6283.2, CG6283.1, kar, Npc2f, wat, Dora, CG32547, CG3091, Vsx2, sd, FBgn0206680, antdh, CG6106, CG5921, Tsp5D, Ir7c, Npc1b, RhoGAP15B, rg, CG42749, pcx, CG32532, unknown_688
D. kepulauana	380	33	apolpp, Pex1, CG6484, stj, mag, CG33107.1, CG13675, CG5644, gsb-n, CG30427, CG17999, CG2064, CG13203, exex, Ser8, Msp300, Dh31, CG43050, Tep3, Fbw5, CG11453, CG3739.1, CG6296, CG6283.1, CG9988, CG2767.1, Myb.1, CG3842, CG1494, CG3106, CG9672, CG11162, CG14234	504	55	dati, CaMKII, toy, RIC-3, CG42747, CG12502, Ir41a, CG13306, Hsp67Bc, ect, Ets65A, PGRP-LD, CG32271, CG32032, fl(2)d, hng3, Cyp9c1, CG44251, CG3955, CG2736, CG10508, CG12355, mthl8, CG13204, Phlpp, ab, Shawl, unknown_257, Pde1c, CG43394, Wnt4, IFT57, beat-IIIc, Rfx, Zip89B, alpha-Est10, Ace, CG6283.2, E5, Rim, unknown_102, Myb.1, FBgn0282836, pigs, CG3823, antdh, unknown_331, FBgn0203886, shf, Tsp5D, dpr18, sov.1, sgg, CG32572, unknown_688
D. s. sulfurigaster	352	21	dati, PGRP-SC1a.2, CG14082, CG30371, Or45b, Obp56e, CG5687, CG1299, sff, serp, slow, Arc2.2, CG15482, CG31100, CG18528, beat-VII, CG6283.2, FBgn0283295, Ndc80, tty, Nep1	865	72	toy, CG3216, FBgn0207282, CG14082, unknown_184, conv, GstE5, Sema5c, unknown_282, lambdaTry.1, Obp56e, mkg-p.2, CG8543, CG12769, CG10912, CG12038, Dscam4, unc-13-4A, CG12869, CG6329, or, CG17999, CG2064, CG1358, CG12355, CG15879, SCOT, Phlpp, CG34367, DIP-iota, CG31869, Nhe2, Msp300, Ret, CG9287, beat-IIIc, kek3, eIF4B, CG31126, CG6325, Rfx, spn-F, CG12420, dpr5, Adk1, CG31496, CG17571, CG31446, Kdm3, Cad96Cb, CG31221, sov.2, CG6283.1, Pxd, CG5555, unknown_102, CG14395, Lerp, CG5359, Npc2f, unknown_125, FBgn0282836, Gr10a, Or10a, OtopLc, Vsx2, Ndc80, CARPB, CG2990, FBgn0204143, CG6106, CG1304

D. pallidifrons	335	44	rho-5, RIC-3, FBgn0207282, CG34116, CG13476, ste24a, Or63a, CG7881, CG32365, frac, CG13920, CG34386, Cht7, CG44251, unknown_786, Hey, DIP-iota, Or22c, GATAd, Mal-B2, kel, H15, CG31663, Myb.2, beat-IIIc, fipi, CG31690, polybromo, CG3999, Ets96B, nAChRalpha1, Cad96Cb, unknown_3, CG14441, CG9981, sd, CG33253, t, m, FBgn0204048, Alms1a, CG42339, CG1695, unknown_688		1129	167	dati, PMCA, sv, ATPsynbeta, jv, FBgn0198895, Tdc2, CG13157, hui, CG14450, CG4186, CG30383, LanA, Best2, Mlh1, CG14760, Cyt-b5, Hsc70-1, 26-29-p, Toll-9, Alp1, CG6163, Alp7, Dp, trpl, lambdaTry.1, CG14837, Sec63, CG13920, CG1299, Cpr76Bd, GNBP1, NUCB1, CG32206, unknown_820, unc-13-4A, CG13930, Acp65Aa, Pgant9, gsb-n, CG15019, CG13502, CG30076, CG13321, CG33143, otk, Hs3st-A, CG2064, grh, CG2915, mms4, CNMaR, mrn, CG7011, CG7458, FBgn0209094, Lst, Cdk4, Spn43Aa, Arf51F, Gr59f, nemy, Dh31-R, Fmo-2, IMPPP, Ip259, LManII, CG34367, jp, CG34109, CG44153, CG9426, Tep4, Ddc, CG17633.2, NimC2, DIP-theta, CG5177, Hr38, uex, Ir40a, CG18302, Lip4, FBgn0198196.1, unknown_46, CG44008, Mal-B2, Sec24CD, Msp300, CG10431, cad, unknown_848, MFS3, toc.1, spz3, unknown_994, vkg, CG9331, Atac2, Nhe3, Myb.2, CG15254, Kr-h1, SpdS, Calr, RpL10, Syn, BBIP1, CG16904, CG4459, unknown_855, CG34384, Spec2, CG1124, mRpL40, Ufl1, CG45263, NKCC, side-III, Cbs, CG3301.1, TwdlN.1, FipoQ, FBgn0201371, CG11899, CG15522, gammaCOP, CG6283.2, CG11550, trv, unknown_102, kar, CG33108, Hsp68.2, Npc2f, Elovl7, CG3091, CG43163, CG3556, CG3106, Edem1, REG, sd, CG33253, unknown_62, CARPB, rst, CG16700, CG1632, CG2681, trol, CG12576, CG9114, Tbh, Cyp318a1, mkg-p.1, Gr5a, Nep1, dpr18, FBgn0203580, Ir7e, CG2556, dec, Npc1b, Tat, Cdk7, CG32533

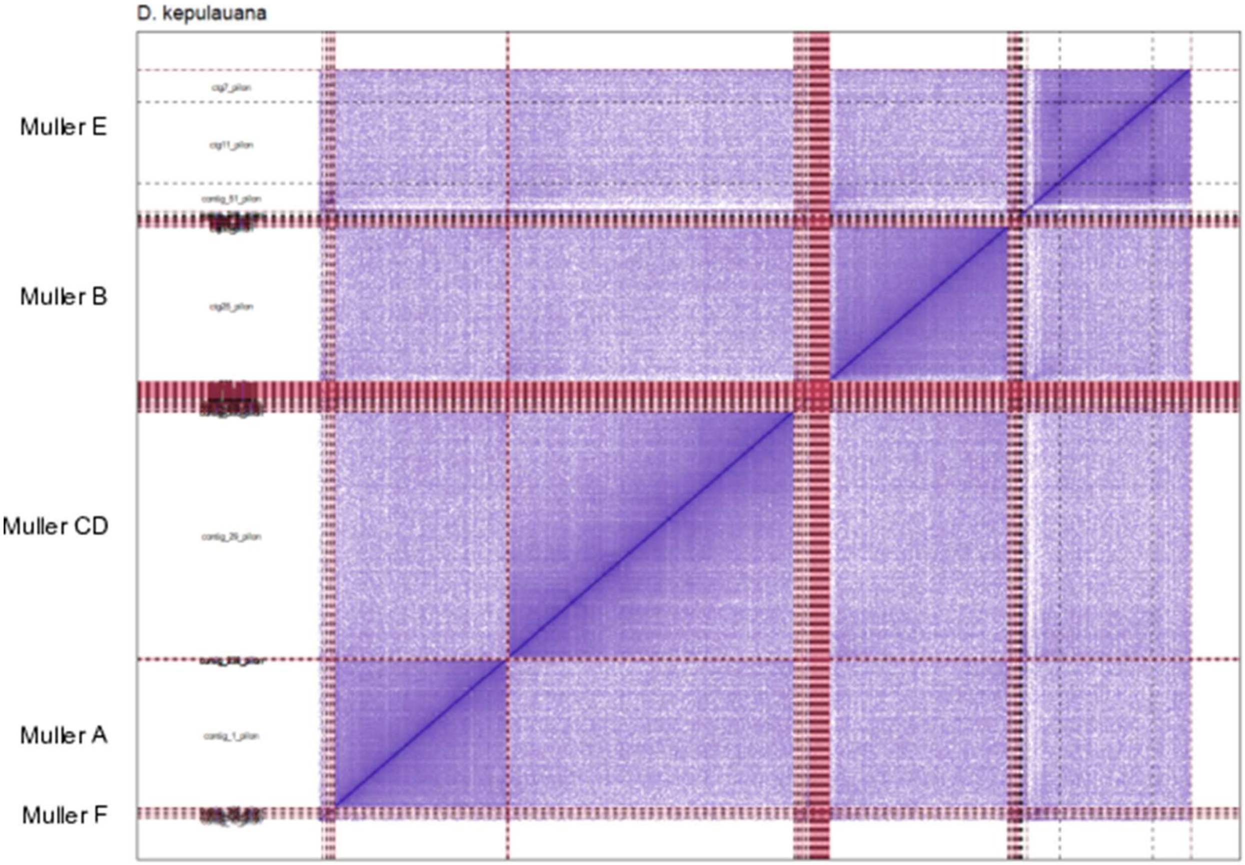
A gene with a TE insertion is one where one or more TEs are found within or +/−1 kb of the gene*



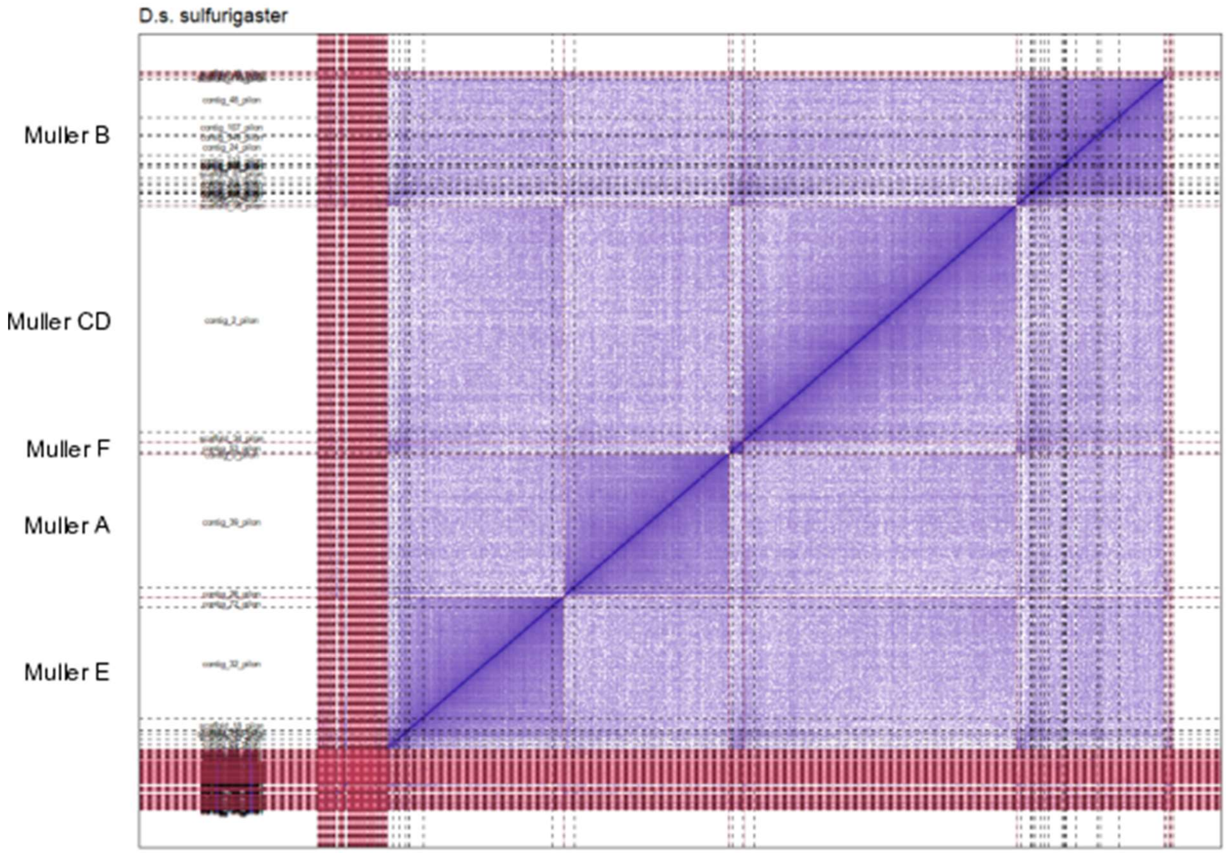
Supplementary Figure S1. Contact heatmap of Hi-C scaffolding of the *D. albomicans* genome. Color intensity represents the intensity of association between positions in the genome. Black dotted lines delimit different contigs while red dotted lines demarcate different scaffolds. Contigs within a scaffold are placed together by Juicer and 3d-dna.



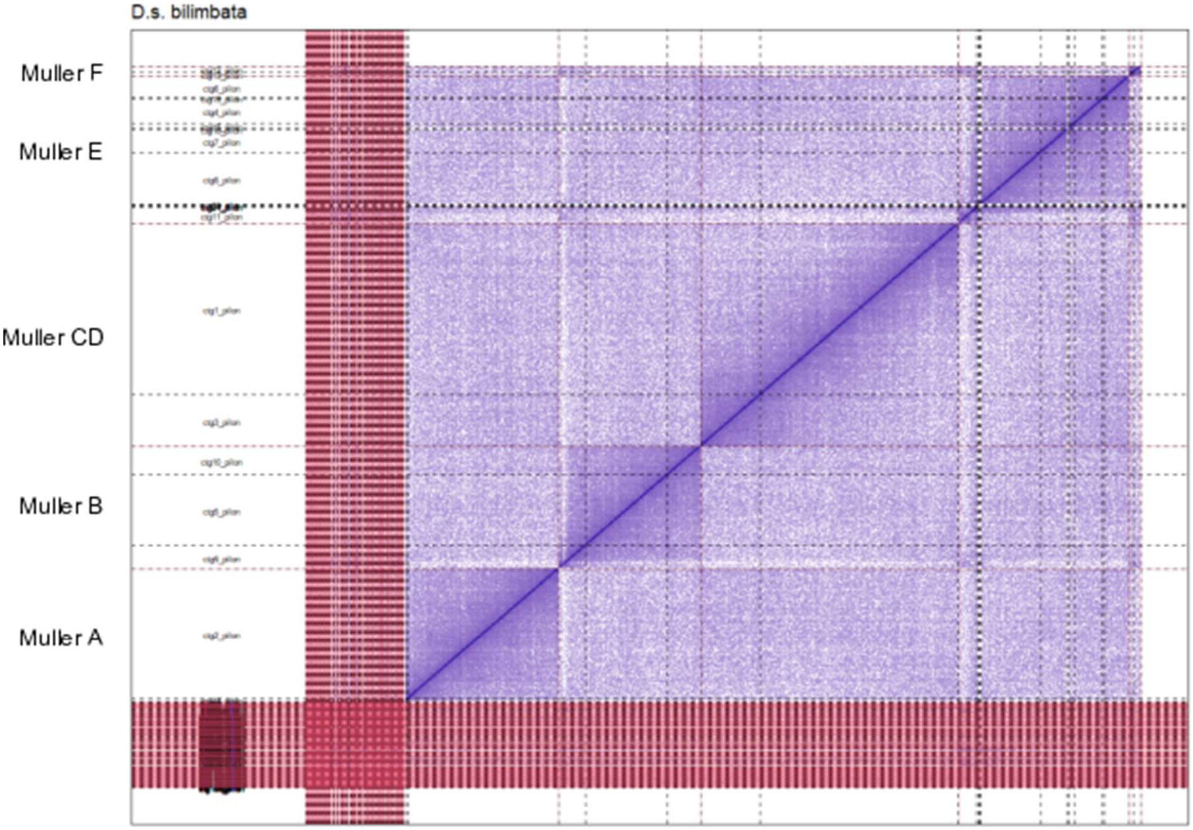
Supplementary Figure S2. Contact heatmap of Hi-C scaffolding of the *D. nasuta* genome. Color intensity represents the intensity of association between positions in the genome. Black dotted lines delimit different contigs while red dotted lines demarcate different scaffolds. Contigs within a scaffold are placed together by Juicer and 3d-dna.



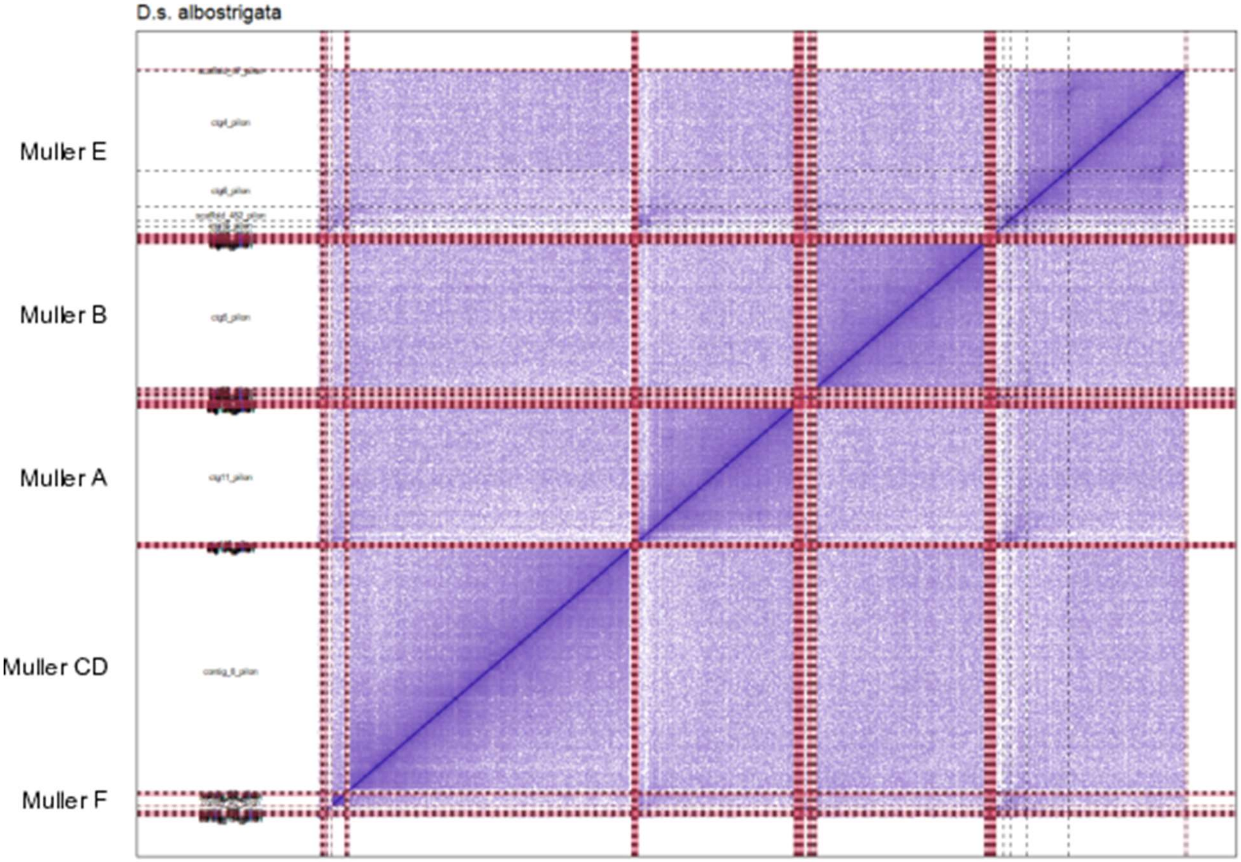
Supplementary Figure S3. Contact heatmap of Hi-C scaffolding of the *D. kepulauana* genome. Color intensity represents the intensity of association between positions in the genome. Black dotted lines delimit different contigs while red dotted lines demarcate different scaffolds. Contigs within a scaffold are placed together by Juicer and 3d-dna.



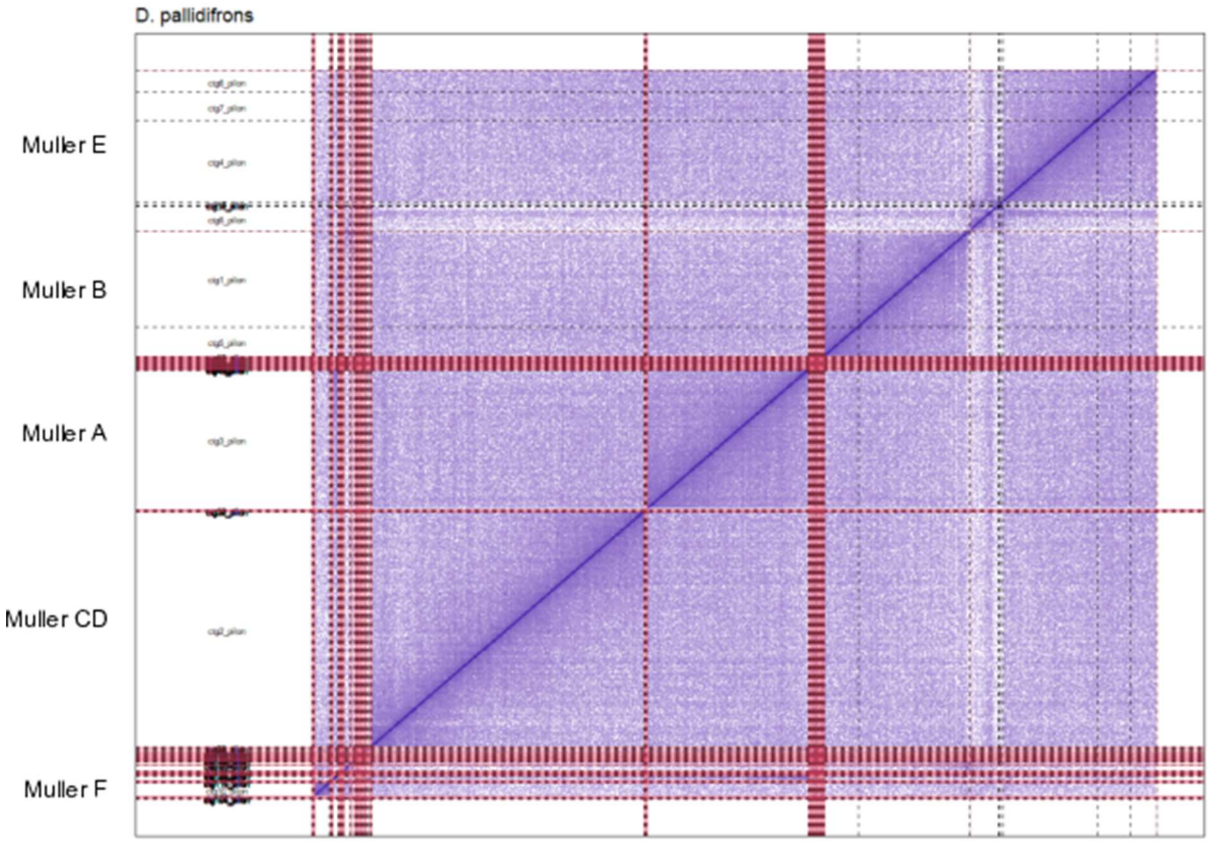
Supplementary Figure S4. Contact heatmap of Hi-C scaffolding of the *D. s. sulfurigaster* genome. Color intensity represents the intensity of association between positions in the genome. Black dotted lines delimit different contigs while red dotted lines demarcate different scaffolds. Contigs within a scaffold are placed together by Juicer and 3d-dna.



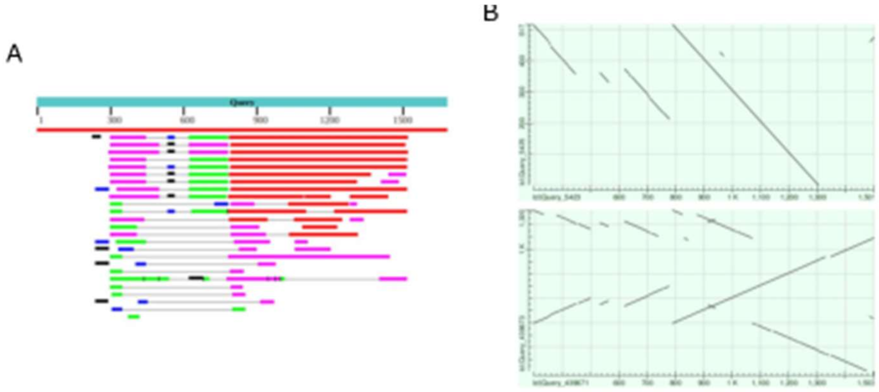
Supplementary Figure S5. Contact heatmap of Hi-C scaffolding of the *D. s. bilimbata* genome. Color intensity represents the intensity of association between positions in the genome. Black dotted lines delimit different contigs while red dotted lines demarcate different scaffolds. Contigs within a scaffold are placed together by Juicer and 3d-dna.



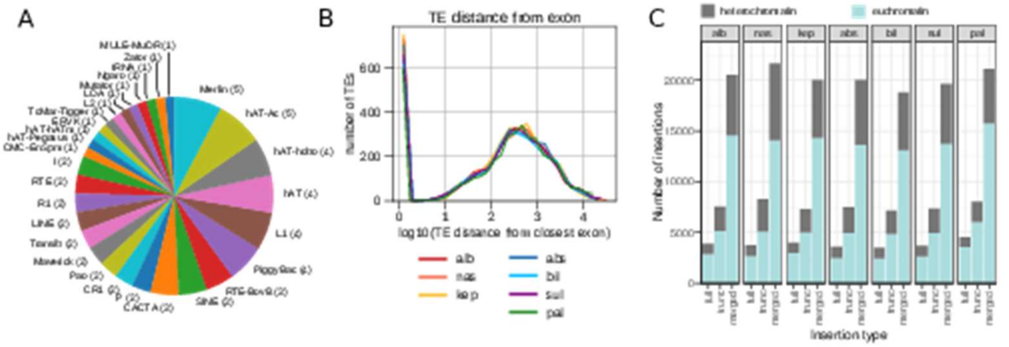
Supplementary Figure S6. Contact heatmap of Hi-C scaffolding of the *D. s. abostrigata* genome. Color intensity represents the intensity of association between positions in the genome. Black dotted lines delimit different contigs while red dotted lines demarcate different scaffolds. Contigs within a scaffold are placed together by Juicer and 3d-dna.



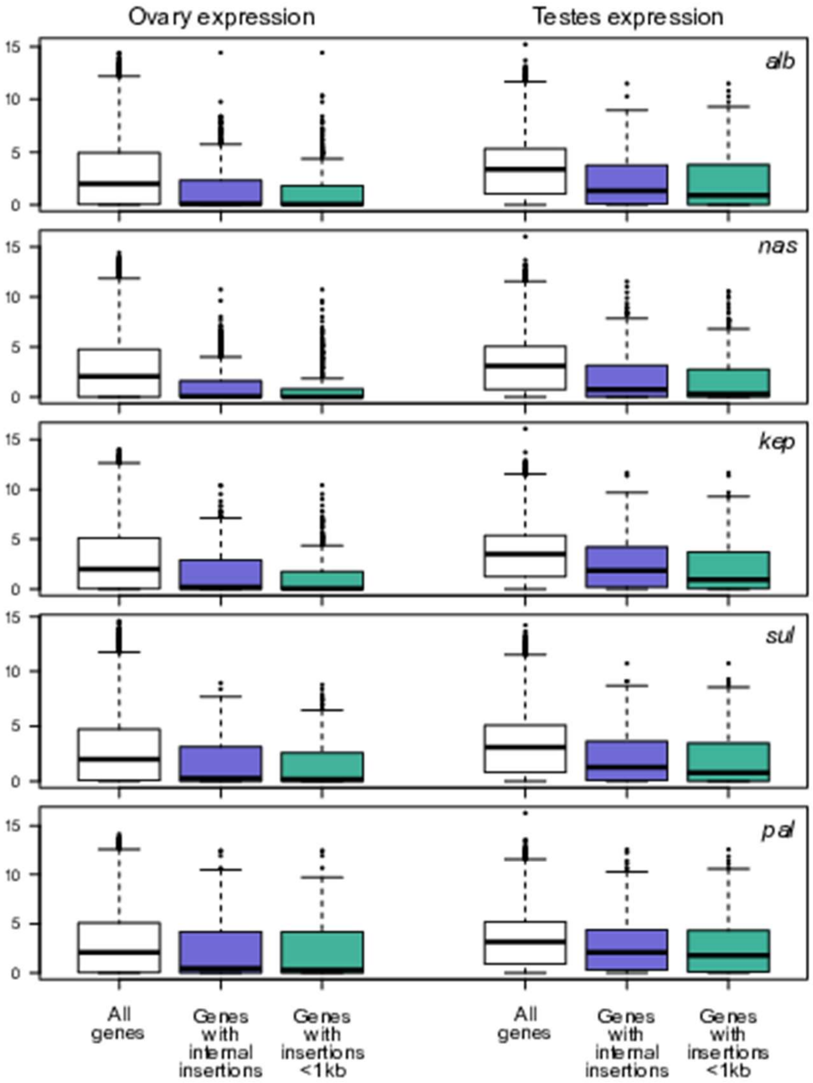
Supplementary Figure S7. Contact heatmap of Hi-C scaffolding of the *D. pallidifrons* genome. Color intensity represents the intensity of association between positions in the genome. Black dotted lines delimit different contigs while red dotted lines demarcate different scaffolds. Contigs within a scaffold are placed together by Juicer and 3d-dna.



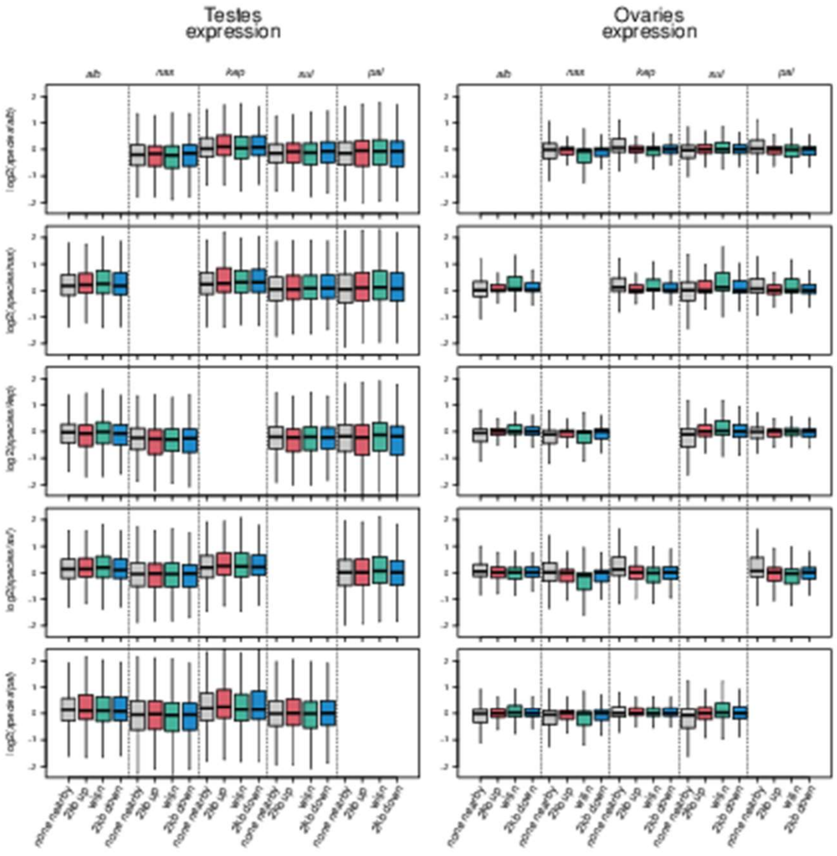
Supplementary Figure S8. Redundancies and fragmentations in the RepeatModeler2 output. **A.** Example of multiple blast hits of one entry to many other entries in the final repeat index. **B.** Examples of alignments between redundant entries showing complex, nested, and redundant structures between entries.



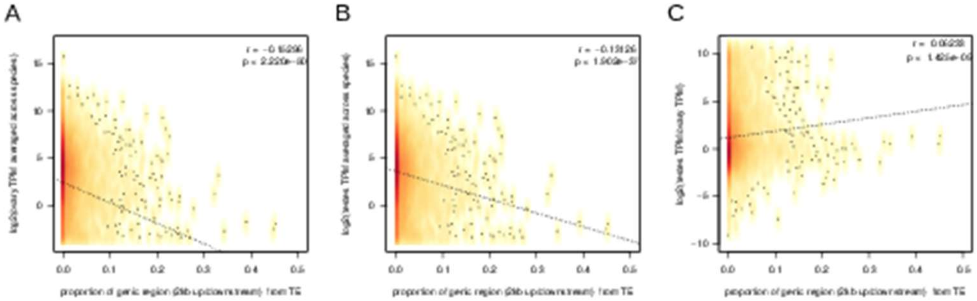
Supplementary Figure S9. Distributions of TE insertions. **A.** Breakdown of the "other" TE classes in the repeat library. **B.** Distribution of the distance of intronic TE insertions to the nearest exon. **C.** Breakdown of the TE insertions falling within euchromatin and pericentric regions.



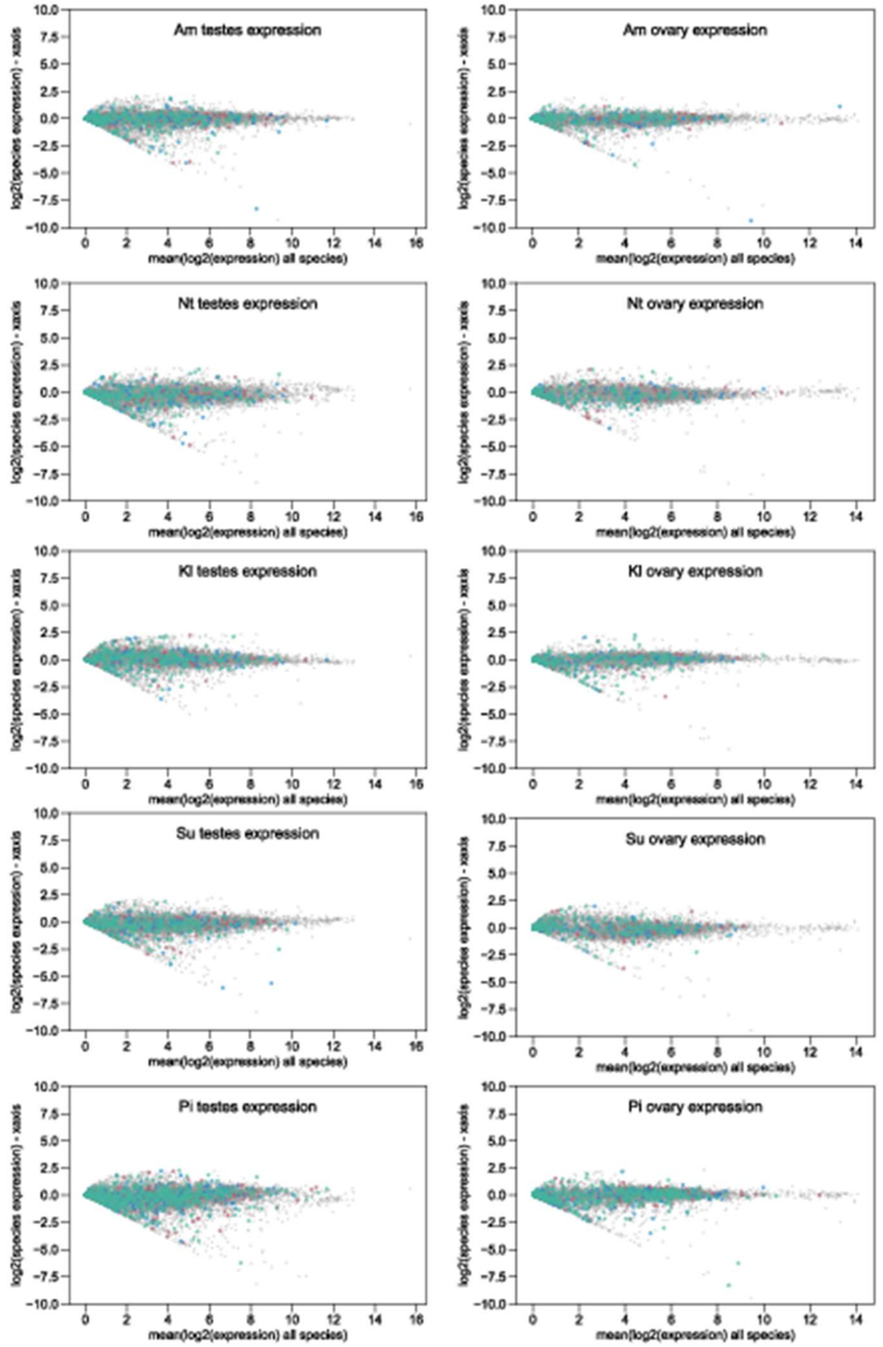
Supplementary Figure S10. TPM of genes with TE insertions nearby/within.



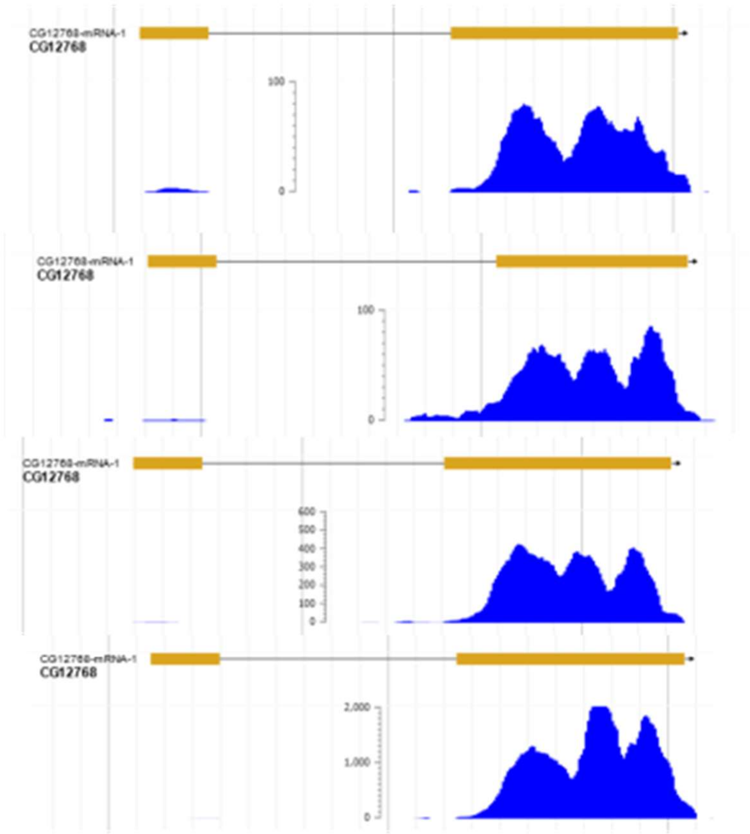
Supplementary Figure S11. Pairwise fold-difference between expression of orthologs in testes and ovaries (left and right in each row, respectively) with TE insertions found in different species. Genes are subsetted as in Figure 2F: those without insertions nearby (grey), those with 5' insertions (red), those with internal insertions (green) and those with 3' insertions (blue) in the species represented by the row. For example, row 1 column 2 shows the $\log_2$(alb TPM / nas TPM) of the genes with insertions in *D. albomicans*.



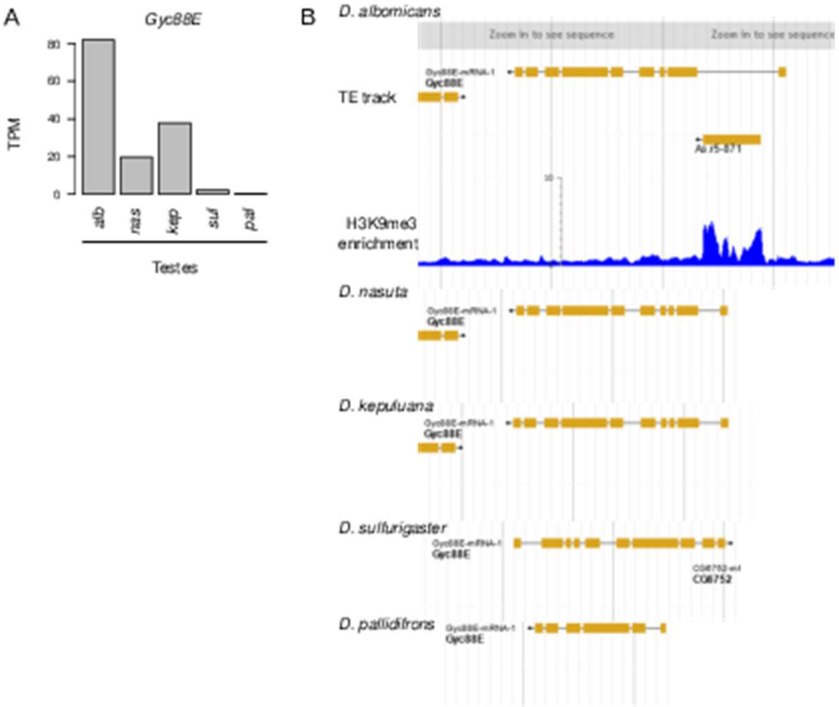
Supplementary Figure S12. A. Correlation between TE content at genes and ovary TPM average across all species. TE content within genes is estimated as the proportion of bases within a window 2kb upstream and down stream of a gene that are TEs. B. Same as A, but with testes TPM. C. Correlation between TE content and fold difference between testes and ovary expression



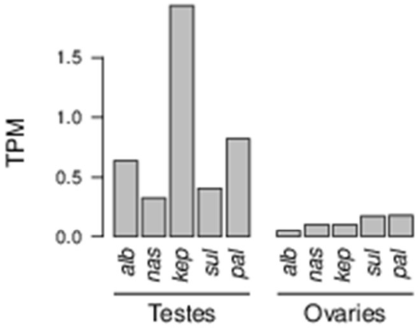
Supplementary Figure S13. A. MA-plots showing fold difference between expression in a species and the cross species average (As is figure 4G).



Supplementary Figure S14. Genome browser shots of CG12768 in different species without insertions. Note the track height diferences for RNA-seq.



Supplementary Figure S15. Example of gene upregulation associated with TE insertion. A. Gene expression across species. B. Genome browser track of the gene in the species genomes.



Supplementary Figure S16. DINE expression in testes and ovaries of the five species.

References

Abraham, Gad, and Michael Inouye. 2014. “Fast Principal Component Analysis of Large-Scale Genome-Wide Data.” *PLOS ONE* 9 (4): e93766. https://doi.org/10.1371/journal.pone.0093766.

Adams, Matthew, Jakob McBroome, Nicholas Maurer, Evan Pepper-Tunick, Nedda F. Saremi, Richard E. Green, Christopher Vollmers, and Russell B. Corbett-Detig. 2020. “One Fly-One Genome: Chromosome-Scale Genome Assembly of a Single Outbred Drosophila Melanogaster.” *Nucleic Acids Research* 48 (13): e75. https://doi.org/10.1093/nar/gkaa450.

Akera, Takashi, Lukáš Chmátal, Emily Trimm, Karren Yang, Chanat Aonbangkhen, David M. Chenoweth, Carsten Janke, Richard M. Schultz, and Michael A. Lampson. 2017. “Spindle Asymmetry Drives Non-Mendelian Chromosome Segregation.” *Science (New York, N.Y.)* 358 (6363): 668–72. https://doi.org/10.1126/science.aan0092.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Anxolabéhère, D., M. G. Kidwell, and G. Periquet. 1988. “Molecular Characteristics of Diverse Populations Are Consistent with the Hypothesis of a Recent Invasion of Drosophila Melanogaster by Mobile P Elements.” *Molecular Biology and Evolution* 5 (3): 252–69. https://doi.org/10.1093/oxfordjournals.molbev.a040491.

Athma, P., and T. Peterson. 1991. “Ac Induces Homologous Recombination at the Maize P Locus.” *Genetics* 128 (1): 163–73.

Bachtrog, Doris. 2006. “The Speciation History of the Drosophila Nasuta Complex.” *Genetical Research* 88 (1): 13–26. https://doi.org/10.1017/S0016672306008330.

Bachtrog, Doris, and Brian Charlesworth. 2002. “Reduced Adaptation of a Non-Recombining Neo-Y Chromosome.” *Nature* 416 (6878): 323–26. https://doi.org/10.1038/416323a.

Bannister, A. J., P. Zegerman, J. F. Partridge, E. A. Miska, J. O. Thomas, R. C. Allshire, and T. Kouzarides. 2001. “Selective Recognition of Methylated Lysine 9 on Histone H3 by the HP1 Chromo Domain.” *Nature* 410 (6824): 120–24. https://doi.org/10.1038/35065138.

Bartolomé, Carolina, and Brian Charlesworth. 2006. “Rates and Patterns of Chromosomal Evolution in Drosophila Pseudoobscura and D. Miranda.” *Genetics* 173 (2): 779–91. https://doi.org/10.1534/genetics.105.054585.

Bayes, Joshua J., and Harmit S. Malik. 2009. “Altered Heterochromatin Binding by a Hybrid Sterility Protein in Drosophila Sibling Species.” *Science (New York, N.Y.)* 326 (5959): 1538–41. https://doi.org/10.1126/science.1181756.

Bhutkar, Arjun, Stephen W. Schaeffer, Susan M. Russo, Mu Xu, Temple F. Smith, and William M. Gelbart. 2008. “Chromosomal Rearrangement Inferred from Comparisons of 12 Drosophila Genomes.” *Genetics* 179 (3): 1657–80. https://doi.org/10.1534/genetics.107.086108.

Bickhart, Derek M., Benjamin D. Rosen, Sergey Koren, Brian L. Sayre, Alex R. Hastie, Saki Chan, Joyce Lee, et al. 2017. “Single-Molecule Sequencing and Chromatin Conformation Capture Enable de Novo Reference Assembly of the Domestic Goat Genome.” *Nature Genetics* 49 (4): 643–50. https://doi.org/10.1038/ng.3802.

Biémont, C., and C. Vieira. 2005. “What Transposable Elements Tell Us about Genome Organization and Evolution: The Case of Drosophila.” *Cytogenetic and Genome Research* 110 (1–4): 25–34. https://doi.org/10.1159/000084935.

Biessmann, H., J. M. Mason, K. Ferry, M. d’Hulst, K. Valgeirsdottir, K. L. Traverse, and M. L. Pardue. 1990. “Addition of Telomere-Associated HeT DNA Sequences ‘Heals’ Broken Chromosome Ends in Drosophila.” *Cell* 61 (4): 663–73. https://doi.org/10.1016/0092-

8674(90)90478-w.

Bourque, Guillaume, Kathleen H. Burns, Mary Gehring, Vera Gorbunova, Andrei Seluanov, Molly Hammell, Michaël Imbeault, et al. 2018. “Ten Things You Should Know about Transposable Elements.” *Genome Biology* 19 (1): 199. https://doi.org/10.1186/s13059-018-1577-z.

Bracewell, Ryan, Kamalakar Chatla, Matthew J. Nalley, and Doris Bachtrog. 2019. “Dynamic Turnover of Centromeres Drives Karyotype Evolution in Drosophila.” *ELife* 8 (September): e49002. https://doi.org/10.7554/eLife.49002.

Bracewell, Ryan R., Barbara J. Bentz, Brian T. Sullivan, and Jeffrey M. Good. 2017. “Rapid Neo-Sex Chromosome Evolution and Incipient Speciation in a Major Forest Pest.” *Nature Communications* 8 (1): 1593. https://doi.org/10.1038/s41467-017-01761-4.

Brand, Cara L., Sarah B. Kingan, Longjun Wu, and Daniel Garrigan. 2013. “A Selective Sweep across Species Boundaries in Drosophila.” *Molecular Biology and Evolution* 30 (9): 2177–86. https://doi.org/10.1093/molbev/mst123.

Brawand, David, Catherine E. Wagner, Yang I. Li, Milan Malinsky, Irene Keller, Shaohua Fan, Oleg Simakov, et al. 2014. “The Genomic Substrate for Adaptive Radiation in African Cichlid Fish.” *Nature* 513 (7518): 375–81. https://doi.org/10.1038/nature13726.

Brideau, Nicholas J., Heather A. Flores, Jun Wang, Shamoni Maheshwari, Xu Wang, and Daniel A. Barbash. 2006. “Two Dobzhansky-Muller Genes Interact to Cause Hybrid Lethality in Drosophila.” *Science (New York, N.Y.)* 314 (5803): 1292–95. https://doi.org/10.1126/science.1133953.

Brown, Judith D., and Rachel J. O’Neill. 2010. “Chromosomes, Conflict, and Epigenetics: Chromosomal Speciation Revisited.” *Annual Review of Genomics and Human Genetics* 11: 291–316. https://doi.org/10.1146/annurev-genom-082509-141554.

Buzdin, A. A. 2004. “Retroelements and Formation of Chimeric Retrogenes.” *Cellular and Molecular Life Sciences: CMLS* 61 (16): 2046–59. https://doi.org/10.1007/s00018-004-4041-z.

Cáceres, M., J. M. Ranz, A. Barbadilla, M. Long, and A. Ruiz. 1999. “Generation of a Widespread Drosophila Inversion by a Transposable Element.” *Science (New York, N.Y.)* 285 (5426): 415–18. https://doi.org/10.1126/science.285.5426.415.

Calvete, Oriol, Josefa González, Esther Betrán, and Alfredo Ruiz. 2012. “Segmental Duplication, Microinversion, and Gene Loss Associated with a Complex Inversion Breakpoint Region in Drosophila.” *Molecular Biology and Evolution* 29 (7): 1875–89. https://doi.org/10.1093/molbev/mss067.

Campbell, Michael S., Carson Holt, Barry Moore, and Mark Yandell. 2014. “Genome Annotation and Curation Using MAKER and MAKER-P.” *Current Protocols in Bioinformatics* 48 (December): 4.11.1-39. https://doi.org/10.1002/0471250953.bi0411s48.

Capy, P., G. Gasperi, C. Biémont, and C. Bazin. 2000. “Stress and Transposable Elements: Co-Evolution or Useful Parasites?” *Heredity* 85 (Pt 2) (August): 101–6. https://doi.org/10.1046/j.1365-2540.2000.00751.x.

Carneiro, Miguel, Frank W. Albert, Sandra Afonso, Ricardo J. Pereira, Hernan Burbano, Rita Campos, José Melo-Ferreira, et al. 2014. “The Genomic Architecture of Population Divergence between Subspecies of the European Rabbit.” *PLOS Genetics* 10 (8): e1003519. https://doi.org/10.1371/journal.pgen.1003519.

Casacuberta, Elena, and Josefa González. 2013. “The Impact of Transposable Elements in Environmental Adaptation.” *Molecular Ecology* 22 (6): 1503–17. https://doi.org/10.1111/mec.12170.

Castillo, Dean M., and Daniel A. Barbash. 2017. “Moving Speciation Genetics Forward: Modern Techniques Build on Foundational Studies in Drosophila.” *Genetics* 207 (3): 825–42. https://doi.org/10.1534/genetics.116.187120.

Casu, R. E. 1990. “Inversion Polymorphism in Populations of Drosophila Sulphurigaster

Albostrigata and Drosophila Nasuta Albomicans from Phuket, Thailand." *Genetica* 81 (3): 157–69. https://doi.org/10.1007/BF00360861.

Chakraborty, Mahul, Ching-Ho Chang, Danielle E. Khost, Jeffrey Vedanayagam, Jeffrey R. Adrion, Yi Liao, Kristi L. Montooth, Colin D. Meiklejohn, Amanda M. Larracuente, and J. J. Emerson. 2021. "Evolution of Genome Structure in the Drosophila Simulans Species Complex." *Genome Research* 31 (3): 380–96. https://doi.org/10.1101/gr.263442.120.

Chang, Christopher C, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. 2015. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." *GigaScience* 4 (1). https://doi.org/10.1186/s13742-015-0047-8.

Chen, Peiwei, Alexei A. Kotov, Baira K. Godneeva, Sergei S. Bazylev, Ludmila V. Olenina, and Alexei A. Aravin. 2021. "PiRNA-Mediated Gene Regulation and Adaptation to Sex-Specific Transposon Expression in D. Melanogaster Male Germline." *Genes & Development* 35 (11–12): 914–35. https://doi.org/10.1101/gad.345041.120.

Cheng, Changde, and Mark Kirkpatrick. 2019. "Inversions Are Bigger on the X Chromosome." *Molecular Ecology* 28 (6): 1238–45. https://doi.org/10.1111/mec.14819.

Cheng, Jade Yu, Thomas Mailund, and Rasmus Nielsen. 2017. "Fast Admixture Analysis and Population Tree Estimation for SNP and NGS Data." *Bioinformatics (Oxford, England)* 33 (14): 2148–55. https://doi.org/10.1093/bioinformatics/btx098.

Choi, Jae Young, and Yuh Chwen G. Lee. 2020. "Double-Edged Sword: The Evolutionary Consequences of the Epigenetic Silencing of Transposable Elements." *PLoS Genetics* 16 (7): e1008872. https://doi.org/10.1371/journal.pgen.1008872.

Chu, Chong, Rasmus Nielsen, and Yufeng Wu. 2016. "REPdenovo: Inferring De Novo Repeat Motifs from Short Sequence Reads." *PLOS ONE* 11 (3): e0150719. https://doi.org/10.1371/journal.pone.0150719.

Coluzzi, M., A. Sabatini, V. Petrarca, and M. A. Di Deco. 1979. "Chromosomal Differentiation and Adaptation to Human Environments in the Anopheles Gambiae Complex." *Transactions of the Royal Society of Tropical Medicine and Hygiene* 73 (5): 483–97. https://doi.org/10.1016/0035-9203(79)90036-1.

Corbett-Detig, Russell B., Charis Cardeno, and Charles H. Langley. 2012. "Sequence-Based Detection and Breakpoint Assembly of Polymorphic Inversions." *Genetics* 192 (1): 131–37. https://doi.org/10.1534/genetics.112.141622.

Cosby, Rachel L., Julius Judd, Ruiling Zhang, Alan Zhong, Nathaniel Garry, Ellen J. Pritham, and Cédric Feschotte. 2021. "Recurrent Evolution of Vertebrate Transcription Factors by Transposase Capture." *Science (New York, N.Y.)* 371 (6531): eabc6405. https://doi.org/10.1126/science.abc6405.

Cridland, Julie M., Stuart J. Macdonald, Anthony D. Long, and Kevin R. Thornton. 2013. "Abundance and Distribution of Transposable Elements in Two Drosophila QTL Mapping Resources." *Molecular Biology and Evolution* 30 (10): 2311–27. https://doi.org/10.1093/molbev/mst129.

Cutter, Asher D. 2008. "Divergence Times in Caenorhabditis and Drosophila Inferred from Direct Estimates of the Neutral Mutation Rate." *Molecular Biology and Evolution* 25 (4): 778–86. https://doi.org/10.1093/molbev/msn024.

Czech, Benjamin, Marzia Munafò, Filippo Ciabrelli, Evelyn L. Eastwood, Martin H. Fabry, Emma Kneuss, and Gregory J. Hannon. 2018. "PiRNA-Guided Genome Defense: From Biogenesis to Silencing." *Annual Review of Genetics* 52 (November): 131–57. https://doi.org/10.1146/annurev-genet-120417-031441.

Daniels, S. B., K. R. Peterson, L. D. Strausbaugh, M. G. Kidwell, and A. Chovnick. 1990. "Evidence for Horizontal Transmission of the P Transposable Element between Drosophila Species." *Genetics* 124 (2): 339–55.

Dannemann, Michael, and Fernando Racimo. 2018. "Something Old, Something Borrowed:

Admixture and Adaptation in Human Evolution.” *Current Opinion in Genetics & Development* 53 (December): 1–8. https://doi.org/10.1016/j.gde.2018.05.009.

DePristo, Mark A., Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, et al. 2011. “A Framework for Variation Discovery and Genotyping Using Next-Generation DNA Sequencing Data.” *Nature Genetics* 43 (5): 491–98. https://doi.org/10.1038/ng.806.

Dion-Côté, Anne-Marie, Sébastien Renaut, Eric Normandeau, and Louis Bernatchez. 2014. “RNA-Seq Reveals Transcriptomic Shock Involving Transposable Elements Reactivation in Hybrids of Young Lake Whitefish Species.” *Molecular Biology and Evolution* 31 (5): 1188–99. https://doi.org/10.1093/molbev/msu069.

Dobzhansky, T. 1948. “Genetics of Natural Populations; Altitudinal and Seasonal Changes Produced by Natural Selection in Certain Populations of Drosophila Persimilis.” *Genetics* 33 (2): 158–76.

Dobzhansky, Theodosius. 1937. *Genetics and the Origin of Species.* New York: Columbia Univ. Press.

———. 1944. “Chromosomal Races in Drosophila Pseudoobscura and Drosophila Persimilis.” *Carnegie Inst. Washington Publ*, no. 554: 47–144.

Dobzhansky, Theodosius, and D. Socolov. 1939. “Structure and Variation of the Chromosomes in Drosophila Azteca.” *Journal of Heredity* 30 (1): 3–19. https://doi.org/10.1093/oxfordjournals.jhered.a104629.

Drosophila 12 Genomes Consortium, Andrew G. Clark, Michael B. Eisen, Douglas R. Smith, Casey M. Bergman, Brian Oliver, Therese A. Markow, et al. 2007. “Evolution of Genes and Genomes on the Drosophila Phylogeny.” *Nature* 450 (7167): 203–18. https://doi.org/10.1038/nature06341.

Dudchenko, Olga, Sanjit S. Batra, Arina D. Omer, Sarah K. Nyquist, Marie Hoeger, Neva C. Durand, Muhammad S. Shamim, et al. 2017. “De Novo Assembly of the Aedes Aegypti Genome Using Hi-C Yields Chromosome-Length Scaffolds.” *Science (New York, N.Y.)* 356 (6333): 92–95. https://doi.org/10.1126/science.aal3327.

Durand, Eric Y., Nick Patterson, David Reich, and Montgomery Slatkin. 2011. “Testing for Ancient Admixture between Closely Related Populations.” *Molecular Biology and Evolution* 28 (8): 2239–52. https://doi.org/10.1093/molbev/msr048.

Durand, Neva C., Muhammad S. Shamim, Ido Machol, Suhas S. P. Rao, Miriam H. Huntley, Eric S. Lander, and Erez Lieberman Aiden. 2016. “Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments.” *Cell Systems* 3 (1): 95–98. https://doi.org/10.1016/j.cels.2016.07.002.

Elgin, Sarah C. R., and Gunter Reuter. 2013. “Position-Effect Variegation, Heterochromatin Formation, and Gene Silencing in Drosophila.” *Cold Spring Harbor Perspectives in Biology* 5 (8): a017780. https://doi.org/10.1101/cshperspect.a017780.

Ellinghaus, David, Stefan Kurtz, and Ute Willhoeft. 2008. “LTRharvest, an Efficient and Flexible Software for de Novo Detection of LTR Retrotransposons.” *BMC Bioinformatics* 9 (January): 18. https://doi.org/10.1186/1471-2105-9-18.

Ellison, Christopher E., and Doris Bachtrog. 2013. “Dosage Compensation via Transposable Element Mediated Rewiring of a Regulatory Network.” *Science (New York, N.Y.)* 342 (6160): 846–50. https://doi.org/10.1126/science.1239552.

———. 2015. “Non-Allelic Gene Conversion Enables Rapid Evolutionary Change at Multiple Regulatory Sites Encoded by Transposable Elements.” *ELife* 4 (February). https://doi.org/10.7554/eLife.05899.

Fawcett, Jeffrey A., and Hideki Innan. 2019. “The Role of Gene Conversion between Transposable Elements in Rewiring Regulatory Networks.” *Genome Biology and Evolution* 11 (7): 1723–29. https://doi.org/10.1093/gbe/evz124.

Ferree, Patrick M., and Daniel A. Barbash. 2009. “Species-Specific Heterochromatin Prevents

Mitotic Chromosome Segregation to Cause Hybrid Lethality in Drosophila." *PLoS Biology* 7 (10): e1000234. https://doi.org/10.1371/journal.pbio.1000234.

Feuk, Lars, Jeffrey R. MacDonald, Terence Tang, Andrew R. Carson, Martin Li, Girish Rao, Razi Khaja, and Stephen W. Scherer. 2005. "Discovery of Human Inversion Polymorphisms by Comparative Analysis of Human and Chimpanzee DNA Sequence Assemblies." *PLoS Genetics* 1 (4): e56. https://doi.org/10.1371/journal.pgen.0010056.

Fishman, Lila, and Arpiar Saunders. 2008. "Centromere-Associated Female Meiotic Drive Entails Male Fitness Costs in Monkeyflowers." *Science (New York, N.Y.)* 322 (5907): 1559–62. https://doi.org/10.1126/science.1161406.

Flynn, Jullien M., Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G. Clark, Cédric Feschotte, and Arian F. Smit. 2020. "RepeatModeler2 for Automated Genomic Discovery of Transposable Element Families." *Proceedings of the National Academy of Sciences of the United States of America* 117 (17): 9451–57. https://doi.org/10.1073/pnas.1921046117.

Fonseca, Nuno A., Cristina P. Vieira, Christian Schlötterer, and Jorge Vieira. 2012. "The DAIBAM MITE Element Is Involved in the Origin of One Fixed and Two Polymorphic Drosophila Virilis Phylad Inversions." *Fly* 6 (2): 71–74. https://doi.org/10.4161/fly.19423.

Fontaine, Michael C., James B. Pease, Aaron Steele, Robert M. Waterhouse, Daniel E. Neafsey, Igor V. Sharakhov, Xiaofang Jiang, et al. 2015. "Extensive Introgression in a Malaria Vector Species Complex Revealed by Phylogenomics." *Science (New York, N.Y.)* 347 (6217): 1258524. https://doi.org/10.1126/science.1258524.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. "CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data." *Bioinformatics (Oxford, England)* 28 (23): 3150–52. https://doi.org/10.1093/bioinformatics/bts565.

Fuller, Zachary L., Gwilym D. Haynes, Stephen Richards, and Stephen W. Schaeffer. 2017. "Genomics of Natural Populations: Evolutionary Forces That Establish and Maintain Gene Arrangements in Drosophila Pseudoobscura." *Molecular Ecology* 26 (23): 6539–62. https://doi.org/10.1111/mec.14381.

Fuller, Zachary L., Christopher J. Leonard, Randee E. Young, Stephen W. Schaeffer, and Nitin Phadnis. 2018. "Ancestral Polymorphisms Explain the Role of Chromosomal Inversions in Speciation." *PLoS Genetics* 14 (7): e1007526. https://doi.org/10.1371/journal.pgen.1007526.

Garrigan, Daniel. 2013. "POPBAM: Tools for Evolutionary Analysis of Short Read Sequence Alignments." *Evolutionary Bioinformatics Online* 9: 343–53. https://doi.org/10.4137/EBO.S12751.

Garrigan, Daniel, Sarah B. Kingan, Anthony J. Geneva, Peter Andolfatto, Andrew G. Clark, Kevin R. Thornton, and Daven C. Presgraves. 2012. "Genome Sequencing Reveals Complex Speciation in the Drosophila Simulans Clade." *Genome Research* 22 (8): 1499–1511. https://doi.org/10.1101/gr.130922.111.

Garza, D., M. Medhora, A. Koga, and D. L. Hartl. 1991. "Introduction of the Transposable Element Mariner into the Germline of Drosophila Melanogaster." *Genetics* 128 (2): 303–10.

Geneva, Anthony J., Christina A. Muirhead, Sarah B. Kingan, and Daniel Garrigan. 2015. "A New Method to Scan Genomes for Introgression in a Secondary Contact Model." *PloS One* 10 (4): e0118621. https://doi.org/10.1371/journal.pone.0118621.

Gleason, Jennifer M., and Michael G. Ritchie. 1998. "Evolution of Courtship Song and Reproductive Isolation in the Drosophila Willistoni Species Complex: Do Sexual Signals Diverge the Most Quickly?" *Evolution; International Journal of Organic Evolution* 52 (5): 1493–1500. https://doi.org/10.1111/j.1558-5646.1998.tb02031.x.

Goidts, Violaine, Justyna M. Szamalek, Pieter J. de Jong, David N. Cooper, Nadia Chuzhanova, Horst Hameister, and Hildegard Kehrer-Sawatzki. 2005. "Independent Intrachromosomal

Recombination Events Underlie the Pericentric Inversions of Chimpanzee and Gorilla Chromosomes Homologous to Human Chromosome 16." *Genome Research* 15 (9): 1232–42. https://doi.org/10.1101/gr.3732505.

González, Josefa, Ferran Casals, and Alfredo Ruiz. 2007. "Testing Chromosomal Phylogenies and Inversion Breakpoint Reuse in Drosophila." *Genetics* 175 (1): 167–77. https://doi.org/10.1534/genetics.106.062612.

Grandbastien, M. A., H. Lucas, J. B. Morel, C. Mhiri, S. Vernhettes, and J. M. Casacuberta. 1997. "The Expression of the Tobacco Tnt1 Retrotransposon Is Linked to Plant Defense Responses." *Genetica* 100 (1–3): 241–52.

Green, Richard E., Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, et al. 2010. "A Draft Sequence of the Neandertal Genome." *Science (New York, N.Y.)* 328 (5979): 710–22. https://doi.org/10.1126/science.1188021.

Guillén, Yolanda, and Alfredo Ruiz. 2012. "Gene Alterations at Drosophila Inversion Breakpoints Provide Prima Facie Evidence for Natural Selection as an Explanation for Rapid Chromosomal Evolution." *BMC Genomics* 13 (February): 53. https://doi.org/10.1186/1471-2164-13-53.

Hall, Ira M., Gurumurthy D. Shankaranarayana, Ken-Ichi Noma, Nabieh Ayoub, Amikam Cohen, and Shiv I. S. Grewal. 2002. "Establishment and Maintenance of a Heterochromatin Domain." *Science (New York, N.Y.)* 297 (5590): 2232–37. https://doi.org/10.1126/science.1076466.

Hedges, D. J., and P. L. Deininger. 2007. "Inviting Instability: Transposable Elements, Double-Strand Breaks, and the Maintenance of Genome Integrity." *Mutation Research* 616 (1–2): 46–59. https://doi.org/10.1016/j.mrfmmm.2006.11.021.

Henikoff, S., K. Ahmad, and H. S. Malik. 2001. "The Centromere Paradox: Stable Inheritance with Rapidly Evolving DNA." *Science (New York, N.Y.)* 293 (5532): 1098–1102. https://doi.org/10.1126/science.1062939.

Hoffmann, Ary A., and Loren H. Rieseberg. 2008. "Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation?" *Annual Review of Ecology, Evolution, and Systematics* 39 (December): 21–42. https://doi.org/10.1146/annurev.ecolsys.39.110707.173532.

Hoffmann, Ary A., Carla M. Sgrò, and Andrew R. Weeks. 2004. "Chromosomal Inversion Polymorphisms and Adaptation." *Trends in Ecology & Evolution* 19 (9): 482–88. https://doi.org/10.1016/j.tree.2004.06.013.

Hollister, Jesse D., and Brandon S. Gaut. 2009. "Epigenetic Silencing of Transposable Elements: A Trade-off between Reduced Transposition and Deleterious Effects on Neighboring Gene Expression." *Genome Research* 19 (8): 1419–28. https://doi.org/10.1101/gr.091678.109.

Hotaling, Scott, John S. Sproul, Jacqueline Heckenhauer, Ashlyn Powell, Amanda M. Larracuente, Steffen U. Pauls, Joanna L. Kelley, and Paul B. Frandsen. 2021. "Long-Reads Are Revolutionizing 20 Years of Insect Genome Sequencing." *Genome Biology and Evolution*, June, evab138. https://doi.org/10.1093/gbe/evab138.

Iwata-Otsubo, Aiko, Jennine M. Dawicki-McKenna, Takashi Akera, Samantha J. Falk, Lukáš Chmátal, Karren Yang, Beth A. Sullivan, Richard M. Schultz, Michael A. Lampson, and Ben E. Black. 2017. "Expanded Satellite Repeats Amplify a Discrete CENP-A Nucleosome Assembly Site on Chromosomes That Drive in Female Meiosis." *Current Biology: CB* 27 (15): 2365-2373.e8. https://doi.org/10.1016/j.cub.2017.06.069.

Izumitani, Hiroyuki F., Yohei Kusaka, Shigeyuki Koshikawa, Masanori J. Toda, and Toru Katoh. 2016. "Phylogeography of the Subgenus Drosophila (Diptera: Drosophilidae): Evolutionary History of Faunal Divergence between the Old and the New Worlds." *PloS One* 11 (7): e0160051. https://doi.org/10.1371/journal.pone.0160051.

Jacques, Pierre-Étienne, Justin Jeyakani, and Guillaume Bourque. 2013. "The Majority of

Primate-Specific Regulatory Sequences Are Derived from Transposable Elements." *PLoS Genetics* 9 (5): e1003504. https://doi.org/10.1371/journal.pgen.1003504.

Keightley, Peter D., Urmi Trivedi, Marian Thomson, Fiona Oliver, Sujai Kumar, and Mark L. Blaxter. 2009a. "Analysis of the Genome Sequences of Three Drosophila Melanogaster Spontaneous Mutation Accumulation Lines." *Genome Research* 19 (7): 1195–1201. https://doi.org/10.1101/gr.091231.109.

———. 2009b. "Analysis of the Genome Sequences of Three Drosophila Melanogaster Spontaneous Mutation Accumulation Lines." *Genome Research* 19 (7): 1195–1201. https://doi.org/10.1101/gr.091231.109.

Kelleher, Erin S., and Daniel A. Barbash. 2013. "Analysis of PiRNA-Mediated Silencing of Active TEs in Drosophila Melanogaster Suggests Limits on the Evolution of Host Genome Defense." *Molecular Biology and Evolution* 30 (8): 1816–29. https://doi.org/10.1093/molbev/mst081.

Khost, Daniel E., Danna G. Eickbush, and Amanda M. Larracuente. 2017. "Single-Molecule Sequencing Resolves the Detailed Structure of Complex Satellite DNA Loci in Drosophila Melanogaster." *Genome Research* 27 (5): 709–21. https://doi.org/10.1101/gr.213512.116.

Khurana, Jaspreet S., Jie Wang, Jia Xu, Birgit S. Koppetsch, Travis C. Thomson, Anetta Nowosielska, Chengjian Li, Phillip D. Zamore, Zhiping Weng, and William E. Theurkauf. 2011. "Adaptation to P Element Transposon Invasion in Drosophila Melanogaster." *Cell* 147 (7): 1551–63. https://doi.org/10.1016/j.cell.2011.11.042.

Kidwell, M. G. 1992. "Horizontal Transfer of P Elements and Other Short Inverted Repeat Transposons." *Genetica* 86 (1–3): 275–86. https://doi.org/10.1007/BF00133726.

Kidwell, M. G., and A. J. Holyoake. 2001. "Transposon-Induced Hotspots for Genomic Instability." *Genome Research* 11 (8): 1321–22. https://doi.org/10.1101/gr.201201.

Kidwell, M. G., J. F. Kidwell, and J. A. Sved. 1977. "Hybrid Dysgenesis in DROSOPHILA MELANOGASTER: A Syndrome of Aberrant Traits Including Mutation, Sterility and Male Recombination." *Genetics* 86 (4): 813–33.

Kidwell, M. G., and D. Lisch. 1997. "Transposable Elements as Sources of Variation in Animals and Plants." *Proceedings of the National Academy of Sciences of the United States of America* 94 (15): 7704–11. https://doi.org/10.1073/pnas.94.15.7704.

Kidwell, M. G., and J. B. Novy. 1979. "Hybrid Dysgenesis in DROSOPHILA MELANOGASTER: Sterility Resulting from Gonadal Dysgenesis in the P-M System." *Genetics* 92 (4): 1127–40.

Kidwell, Margaret G. 2002. "Transposable Elements and the Evolution of Genome Size in Eukaryotes." *Genetica* 115 (1): 49–63. https://doi.org/10.1023/a:1016072014259.

Kim, Daehwan, Ben Langmead, and Steven L. Salzberg. 2015. "HISAT: A Fast Spliced Aligner with Low Memory Requirements." *Nature Methods* 12 (4): 357–60. https://doi.org/10.1038/nmeth.3317.

Kitagawa, Osamu, Ken-Ichi Wakahama, Yoshiaki Fuyama, Yoko Shimada, Etsuko Takanashi, Machiko Hatsumi, Momoko Uwabo, and Yoshiko Mita. 1982. "Genetic Studies of the Drosophila Nasuta Subgroup, with Notes on Distribution and Morphology." 遺伝學雑誌 57 (2): 113–41. https://doi.org/10.1266/jjg.57.113.

Kitano, Jun, Joseph A. Ross, Seiichi Mori, Manabu Kume, Felicity C. Jones, Yingguang F. Chan, Devin M. Absher, et al. 2009. "A Role for a Neo-Sex Chromosome in Stickleback Speciation." *Nature* 461 (7267): 1079–83. https://doi.org/10.1038/nature08441.

Kliman, R. M., P. Andolfatto, J. A. Coyne, F. Depaulis, M. Kreitman, A. J. Berry, J. McCarter, J. Wakeley, and J. Hey. 2000. "The Population Genetics of the Origin and Divergence of the Drosophila Simulans Complex Species." *Genetics* 156 (4): 1913–31.

Kofler, Robert, Tom Hill, Viola Nolte, Andrea J. Betancourt, and Christian Schlötterer. 2015. "The Recent Invasion of Natural Drosophila Simulans Populations by the P-Element."

Proceedings of the National Academy of Sciences of the United States of America 112 (21): 6659–63. https://doi.org/10.1073/pnas.1500758112.

Kolmogorov, Mikhail, Jeffrey Yuan, Yu Lin, and Pavel A. Pevzner. 2019. “Assembly of Long, Error-Prone Reads Using Repeat Graphs.” *Nature Biotechnology* 37 (5): 540–46. https://doi.org/10.1038/s41587-019-0072-8.

Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. “Canu: Scalable and Accurate Long-Read Assembly via Adaptive k-Mer Weighting and Repeat Separation.” *Genome Research* 27 (5): 722–36. https://doi.org/10.1101/gr.215087.116.

Krimbas, Kōstas V., and Jeffrey R. Powell, eds. 1992. *Drosophila Inversion Polymorphism*. Boca Raton, Fla: CRC Press.

Kriventseva, Evgenia V., Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosè Manni, Renata Dias, Felipe A. Simão, and Evgeny M. Zdobnov. 2019. “OrthoDB V10: Sampling the Diversity of Animal, Plant, Fungal, Protist, Bacterial and Viral Genomes for Evolutionary and Functional Annotations of Orthologs.” *Nucleic Acids Research* 47 (D1): D807–11. https://doi.org/10.1093/nar/gky1053.

Krivshenko, J. D. 1963. “The Chromosomal Polymorphism of Drosophila Busckii in Natural Populations.” *Genetics* 48 (September): 1239–58.

Kurhanewicz, Nicole A., Devin Dinwiddie, Zachary D. Bush, and Diana E. Libuda. 2020. “Elevated Temperatures Cause Transposon-Associated DNA Damage in C. Elegans Spermatocytes.” *Current Biology: CB* 30 (24): 5007-5017.e4. https://doi.org/10.1016/j.cub.2020.09.050.

Kurtz, Stefan, Adam Phillippy, Arthur L. Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L. Salzberg. 2004. “Versatile and Open Software for Comparing Large Genomes.” *Genome Biology* 5 (2): R12. https://doi.org/10.1186/gb-2004-5-2-r12.

Lachner, M., D. O’Carroll, S. Rea, K. Mechtler, and T. Jenuwein. 2001. “Methylation of Histone H3 Lysine 9 Creates a Binding Site for HP1 Proteins.” *Nature* 410 (6824): 116–20. https://doi.org/10.1038/35065132.

Lambert, D. M. 1982. “Mate Recognition in Members of the Drosophila Nasuta Complex.” *Animal Behaviour* 30 (2): 438–43. https://doi.org/10.1016/S0003-3472(82)80054-7.

Lamichhaney, Sangeet, Jonas Berglund, Markus Sällman Almén, Khurram Maqbool, Manfred Grabherr, Alvaro Martinez-Barrio, Marta Promerová, et al. 2015. “Evolution of Darwin’s Finches and Their Beaks Revealed by Genome Sequencing.” *Nature* 518 (7539): 371–75. https://doi.org/10.1038/nature14181.

Le Rouzic, Arnaud, and Pierre Capy. 2005. “The First Steps of Transposable Elements Invasion: Parasitic Strategy vs. Genetic Drift.” *Genetics* 169 (2): 1033–43. https://doi.org/10.1534/genetics.104.031211.

Lee, Yuh Chwen G. 2015. “The Role of PiRNA-Mediated Epigenetic Silencing in the Population Dynamics of Transposable Elements in Drosophila Melanogaster.” *PLoS Genetics* 11 (6): e1005269. https://doi.org/10.1371/journal.pgen.1005269.

Lee, Yuh Chwen G., and Gary H. Karpen. 2017. “Pervasive Epigenetic Effects of Drosophila Euchromatic Transposable Elements Impact Their Evolution.” *ELife* 6 (July): e25762. https://doi.org/10.7554/eLife.25762.

Lemeunier, F., and M. Ashburner. 1984. “Relationships within the Melanogaster Species Subgroup of the Genus Drosophila (Sophophora).” *Chromosoma* 89 (5): 343–51. https://doi.org/10.1007/BF00331251.

Levis, R. W., R. Ganesan, K. Houtchens, L. A. Tolar, and F. M. Sheen. 1993. “Transposons in Place of Telomeric Repeats at a Drosophila Telomere.” *Cell* 75 (6): 1083–93. https://doi.org/10.1016/0092-8674(93)90318-k.

Li, Heng. 2016. “Minimap and Miniasm: Fast Mapping and de Novo Assembly for Noisy Long Sequences.” *Bioinformatics (Oxford, England)* 32 (14): 2103–10.

https://doi.org/10.1093/bioinformatics/btw152.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics (Oxford, England)* 25 (14): 1754–60. https://doi.org/10.1093/bioinformatics/btp324.

Linheiro, Raquel S., and Casey M. Bergman. 2012. “Whole Genome Resequencing Reveals Natural Target Site Preferences of Transposable Elements in Drosophila Melanogaster.” *PloS One* 7 (2): e30008. https://doi.org/10.1371/journal.pone.0030008.

Lopez-Maestre, Hélène, Elias A. G. Carnelossi, Vincent Lacroix, Nelly Burlet, Bruno Mugat, Séverine Chambeyron, Claudia M. A. Carareto, and Cristina Vieira. 2017. “Identification of Misexpressed Genetic Elements in Hybrids between Drosophila-Related Species.” *Scientific Reports* 7 (January): 40618. https://doi.org/10.1038/srep40618.

Loreto, E. L. S., C. M. A. Carareto, and P. Capy. 2008. “Revisiting Horizontal Transfer of Transposable Elements in Drosophila.” *Heredity* 100 (6): 545–54. https://doi.org/10.1038/sj.hdy.6801094.

Luo, Shiqi, Hong Zhang, Yuange Duan, Xinmin Yao, Andrew G. Clark, and Jian Lu. 2020. “The Evolutionary Arms Race between Transposable Elements and PiRNAs in Drosophila Melanogaster.” *BMC Evolutionary Biology* 20 (1): 14. https://doi.org/10.1186/s12862-020-1580-3.

Mahajan, Shivani, Kevin H.-C. Wei, Matthew J. Nalley, Lauren Gibilisco, and Doris Bachtrog. 2018. “De Novo Assembly of a Young Drosophila Y Chromosome Using Single-Molecule Sequencing and Chromatin Conformation Capture.” *PLoS Biology* 16 (7): e2006348. https://doi.org/10.1371/journal.pbio.2006348.

Mai, Dat, Matthew J. Nalley, and Doris Bachtrog. 2020. “Patterns of Genomic Differentiation in the Drosophila Nasuta Species Complex.” *Molecular Biology and Evolution* 37 (1): 208–20. https://doi.org/10.1093/molbev/msz215.

Martin, Simon H., Kanchon K. Dasmahapatra, Nicola J. Nadeau, Camilo Salazar, James R. Walters, Fraser Simpson, Mark Blaxter, Andrea Manica, James Mallet, and Chris D. Jiggins. 2013. “Genome-Wide Evidence for Speciation with Gene Flow in Heliconius Butterflies.” *Genome Research* 23 (11): 1817–28. https://doi.org/10.1101/gr.159426.113.

Martin, Simon H., John W. Davey, and Chris D. Jiggins. 2015. “Evaluating the Use of ABBA-BABA Statistics to Locate Introgressed Loci.” *Molecular Biology and Evolution* 32 (1): 244–57. https://doi.org/10.1093/molbev/msu269.

Martin, Simon H., and Chris D. Jiggins. 2017. “Interpreting the Genomic Landscape of Introgression.” *Current Opinion in Genetics & Development* 47 (December): 69–74. https://doi.org/10.1016/j.gde.2017.08.007.

Masly, John P., and Daven C. Presgraves. 2007. “High-Resolution Genome-Wide Dissection of the Two Rules of Speciation in Drosophila.” *PLoS Biology* 5 (9): e243. https://doi.org/10.1371/journal.pbio.0050243.

Mateo, Lidia, Anna Ullastres, and Josefa González. 2014. “A Transposable Element Insertion Confers Xenobiotic Resistance in Drosophila.” *PLoS Genetics* 10 (8): e1004560. https://doi.org/10.1371/journal.pgen.1004560.

Matthews, Benjamin J., Olga Dudchenko, Sarah B. Kingan, Sergey Koren, Igor Antoshechkin, Jacob E. Crawford, William J. Glassford, et al. 2018. “Improved Reference Genome of Aedes Aegypti Informs Arbovirus Vector Control.” *Nature* 563 (7732): 501–7. https://doi.org/10.1038/s41586-018-0692-z.

McBroome, Jakob, David Liang, and Russell Corbett-Detig. 2020. “Fine-Scale Position Effects Shape the Distribution of Inversion Breakpoints in Drosophila Melanogaster.” *Genome Biology and Evolution* 12 (8): 1378–91. https://doi.org/10.1093/gbe/evaa103.

McGurk, Michael P., and Daniel A. Barbash. 2018. “Double Insertion of Transposable Elements Provides a Substrate for the Evolution of Satellite DNA.” *Genome Research* 28 (5): 714–25. https://doi.org/10.1101/gr.231472.117.

Meiklejohn, Colin D., Emily L. Landeen, Kathleen E. Gordon, Thomas Rzatkiewicz, Sarah B. Kingan, Anthony J. Geneva, Jeffrey P. Vedanayagam, et al. 2018. "Gene Flow Mediates the Role of Sex Chromosome Meiotic Drive during Complex Speciation." *ELife* 7 (December). https://doi.org/10.7554/eLife.35468.

Mérel, Vincent, Matthieu Boulesteix, Marie Fablet, and Cristina Vieira. 2020. "Transposable Elements in Drosophila." *Mobile DNA* 11: 23. https://doi.org/10.1186/s13100-020-00213-z.

Merenciano, Miriam, Anna Ullastres, M. a. R. de Cara, Maite G. Barrón, and Josefa González. 2016. "Multiple Independent Retroelement Insertions in the Promoter of a Stress Response Gene Have Variable Molecular and Functional Effects in Drosophila." *PLoS Genetics* 12 (8): e1006249. https://doi.org/10.1371/journal.pgen.1006249.

Michael, Todd P., Florian Jupe, Felix Bemm, S. Timothy Motley, Justin P. Sandoval, Christa Lanz, Olivier Loudet, Detlef Weigel, and Joseph R. Ecker. 2018. "High Contiguity Arabidopsis Thaliana Genome Assembly with a Single Nanopore Flow Cell." *Nature Communications* 9 (1): 541. https://doi.org/10.1038/s41467-018-03016-2.

Mirarab, S., R. Reaz, Md S. Bayzid, T. Zimmermann, M. S. Swenson, and T. Warnow. 2014. "ASTRAL: Genome-Scale Coalescent-Based Species Tree Estimation." *Bioinformatics (Oxford, England)* 30 (17): i541-548. https://doi.org/10.1093/bioinformatics/btu462.

Muller, H.J. 1940. "Bearing of the Drosophila Work on Systematics." In *The New Systematics*, 185--268.

———. 1942. "Isolating Mechanisms, Evolution, and Temperature." *Biology Symposium*, no. 6: 71–125.

Murphy, William J., Denis M. Larkin, Annelie Everts-van der Wind, Guillaume Bourque, Glenn Tesler, Loretta Auvil, Jonathan E. Beever, et al. 2005. "Dynamics of Mammalian Chromosome Evolution Inferred from Multispecies Comparative Maps." *Science (New York, N.Y.)* 309 (5734): 613–17. https://doi.org/10.1126/science.1111387.

Nagaraja, null, J. Nagaraju, and H. A. Ranganath. 2004. "Molecular Phylogeny of the Nasuta Subgroup of Drosophila Based on 12S RRNA, 16S RRNA and CoI Mitochondrial Genes, RAPD and ISSR Polymorphisms." *Genes & Genetic Systems* 79 (5): 293–99. https://doi.org/10.1266/ggs.79.293.

Nakayama, J., J. C. Rice, B. D. Strahl, C. D. Allis, and S. I. Grewal. 2001. "Role of Histone H3 Lysine 9 Methylation in Epigenetic Control of Heterochromatin Assembly." *Science (New York, N.Y.)* 292 (5514): 110–13. https://doi.org/10.1126/science.1060118.

Nurk, Sergey, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, et al. 2021. "The Complete Sequence of a Human Genome." *BioRxiv*, May, 2021.05.26.445798. https://doi.org/10.1101/2021.05.26.445798.

O'Grady, Patrick M., and Rob DeSalle. 2018. "Phylogeny of the Genus Drosophila." *Genetics* 209 (1): 1–25. https://doi.org/10.1534/genetics.117.300583.

O'Neill, Kathryn, David Brocks, and Molly Gale Hammell. 2020. "Mobile Genomics: Tools and Techniques for Tackling Transposons." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 375 (1795): 20190345. https://doi.org/10.1098/rstb.2019.0345.

O'Neill, R. J., M. J. O'Neill, and J. A. Graves. 1998. "Undermethylation Associated with Retroelement Activation and Chromosome Remodelling in an Interspecific Mammalian Hybrid." *Nature* 393 (6680): 68–72. https://doi.org/10.1038/29985.

Orr, H. A. 1993. "Haldane's Rule Has Multiple Genetic Causes." *Nature* 361 (6412): 532–33. https://doi.org/10.1038/361532a0.

Ozata, Deniz M., Ildar Gainetdinov, Ansgar Zoch, Dónal O'Carroll, and Phillip D. Zamore. 2019. "PIWI-Interacting RNAs: Small RNAs with Big Functions." *Nature Reviews. Genetics* 20 (2): 89–108. https://doi.org/10.1038/s41576-018-0073-3.

Panta, Manisha, Avdesh Mishra, Md Tamjidul Hoque, and Joel Atallah. 2021. “ClassifyTE: A Stacking Based Prediction of Hierarchical Classification of Transposable Elements.” *Bioinformatics (Oxford, England)*, March, btab146. https://doi.org/10.1093/bioinformatics/btab146.

Papaceit, Montserrat, Montserrat Aguadé, and Carmen Segarra. 2006. “Chromosomal Evolution of Elements B and C in the Sophophora Subgenus of Drosophila: Evolutionary Rate and Polymorphism.” *Evolution; International Journal of Organic Evolution* 60 (4): 768–81.

Parhad, Swapnil S., and William E. Theurkauf. 2019. “Rapid Evolution and Conserved Function of the PiRNA Pathway.” *Open Biology* 9 (1): 180181. https://doi.org/10.1098/rsob.180181.

Pertea, Mihaela, Geo M. Pertea, Corina M. Antonescu, Tsung-Cheng Chang, Joshua T. Mendell, and Steven L. Salzberg. 2015. “StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads.” *Nature Biotechnology* 33 (3): 290–95. https://doi.org/10.1038/nbt.3122.

Pevzner, Pavel, and Glenn Tesler. 2003. “Human and Mouse Genomic Sequences Reveal Extensive Breakpoint Reuse in Mammalian Evolution.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (13): 7672–77. https://doi.org/10.1073/pnas.1330369100.

Phadnis, Nitin, and H. Allen Orr. 2009. “A Single Gene Causes Both Male Sterility and Segregation Distortion in Drosophila Hybrids.” *Science (New York, N.Y.)* 323 (5912): 376–79. https://doi.org/10.1126/science.1163934.

Phifer-Rixey, Megan, Matthew Bomhoff, and Michael W. Nachman. 2014. “Genome-Wide Patterns of Differentiation among House Mouse Subspecies.” *Genetics* 198 (1): 283–97. https://doi.org/10.1534/genetics.114.166827.

Pope, A. 1987. “Inversion Polymorphism in Species of the Drosophila Nasuta Subgroup from Thailand.” *Genetica* 72 (1): 55–64. https://doi.org/10.1007/BF00126978.

Porubsky, David, Ashley D. Sanders, Wolfram Höps, PingHsun Hsieh, Arvis Sulovari, Ruiyang Li, Ludovica Mercuri, et al. 2020. “Recurrent Inversion Toggling and Great Ape Genome Evolution.” *Nature Genetics* 52 (8): 849–58. https://doi.org/10.1038/s41588-020-0646-x.

Powell, Jeffrey R. 1997. *Progress and Prospects in Evolutionary Biology: The Drosophila Model*. New York: Oxford University Press.

Presgraves, Daven C. 2003. “A Fine-Scale Genetic Analysis of Hybrid Incompatibilities in Drosophila.” *Genetics* 163 (3): 955–72.

———. 2018. “Evaluating Genomic Signatures of ‘the Large X-Effect’ during Complex Speciation.” *Molecular Ecology* 27 (19): 3822–30. https://doi.org/10.1111/mec.14777.

Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics (Oxford, England)* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

Rahman, Reazur, Gung-wei Chirn, Abhay Kanodia, Yuliya A. Sytnikova, Björn Brembs, Casey M. Bergman, and Nelson C. Lau. 2015. “Unique Transposon Landscapes Are Pervasive across Drosophila Melanogaster Genomes.” *Nucleic Acids Research* 43 (22): 10655–72. https://doi.org/10.1093/nar/gkv1193.

Ranjini, Mysore S., and Nallur B. Ramachandra. 2013. “Rapid Evolution of a Few Members of Nasuta-Albomicans Complex of Drosophila: Study on Two Candidate Genes, Sod1 and Rpd3.” *Journal of Molecular Evolution* 76 (5): 311–23. https://doi.org/10.1007/s00239-013-9560-5.

Ranz, José M., Damien Maurin, Yuk S. Chan, Marcin von Grotthuss, LaDeana W. Hillier, John Roote, Michael Ashburner, and Casey M. Bergman. 2007. “Principles of Genome Evolution in the Drosophila Melanogaster Species Group.” *PLoS Biology* 5 (6): e152. https://doi.org/10.1371/journal.pbio.0050152.

Ratner, V. A., S. A. Zabanov, O. V. Kolesnikova, and L. A. Vasilyeva. 1992. “Induction of the

Mobile Genetic Element Dm-412 Transpositions in the Drosophila Genome by Heat Shock Treatment." *Proceedings of the National Academy of Sciences of the United States of America* 89 (12): 5650–54. https://doi.org/10.1073/pnas.89.12.5650.

Reis, Micael, Cristina P. Vieira, Rodrigo Lata, Nico Posnien, and Jorge Vieira. 2018. "Origin and Consequences of Chromosomal Inversions in the Virilis Group of Drosophila." *Genome Biology and Evolution* 10 (12): 3152–66. https://doi.org/10.1093/gbe/evy239.

RepeatMasker Open-4.0. 2013. http://www.repeatmasker.org.

Revell, Liam J. 2012. "Phytools: An R Package for Phylogenetic Comparative Biology (and Other Things)." *Methods in Ecology and Evolution* 3 (2): 217–23. https://doi.org/10.1111/j.2041-210X.2011.00169.x.

Richards, Eric J., and Sarah C. R. Elgin. 2002. "Epigenetic Codes for Heterochromatin Formation and Silencing: Rounding up the Usual Suspects." *Cell* 108 (4): 489–500. https://doi.org/10.1016/s0092-8674(02)00644-x.

Richards, Stephen, Yue Liu, Brian R. Bettencourt, Pavel Hradecky, Stan Letovsky, Rasmus Nielsen, Kevin Thornton, et al. 2005. "Comparative Genome Sequencing of Drosophila Pseudoobscura: Chromosomal, Gene, and Cis-Element Evolution." *Genome Research* 15 (1): 1–18. https://doi.org/10.1101/gr.3059305.

Rosin, Leah F., and Barbara G. Mellone. 2017. "Centromeres Drive a Hard Bargain." *Trends in Genetics: TIG* 33 (2): 101–17. https://doi.org/10.1016/j.tig.2016.12.001.

Ruan, Jue, and Heng Li. 2020. "Fast and Accurate Long-Read Assembly with Wtdbg2." *Nature Methods* 17 (2): 155–58. https://doi.org/10.1038/s41592-019-0669-3.

Schaack, Sarah, Clément Gilbert, and Cédric Feschotte. 2010. "Promiscuous DNA: Horizontal Transfer of Transposable Elements and Why It Matters for Eukaryotic Evolution." *Trends in Ecology & Evolution* 25 (9): 537–46. https://doi.org/10.1016/j.tree.2010.06.001.

Schaefer, R. E., M. G. Kidwell, and A. Fausto-Sterling. 1979. "Hybrid Dysgenesis in DROSOPHILA MELANOGASTER: Morphological and Cytological Studies of Ovarian Dysgenesis." *Genetics* 92 (4): 1141–52.

Serio, Carmela, Silvia Castiglione, Gianmarco Tesone, Martina Piccolo, Marina Melchionna, Alessandro Mondanaro, Mirko Di Febbraro, and Pasquale Raia. 2019. "Macroevolution of Toothed Whales Exceptional Relative Brain Size." *Evolutionary Biology* 46 (4): 332–42. https://doi.org/10.1007/s11692-019-09485-7.

Shao, H., D. Li, X. Zhang, H. Yu, X. Li, D. Zhu, Y. Zhou, and Z. Geng. 1997. "Study on the recognition and evolutionary genetics of the courtship song of species in Drosphila nasuta species subgroup." *Yi Chuan Xue Bao = Acta Genetica Sinica* 24 (4): 311–21.

Sievers, Fabian, Andreas Wilm, David Dineen, Toby J. Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, et al. 2011. "Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega." *Molecular Systems Biology* 7 (October): 539. https://doi.org/10.1038/msb.2011.75.

Silva, Danubia Guimarães, Hermes José Schmitz, Hermes Fonseca de Medeiros, Cláudia Rohde, Martín Alejandro Montes, and Ana Cristina Lauer Garcia. 2020. "Geographic Expansion and Dominance of the Invading Species Drosophila Nasuta (Diptera, Drosophilidae) in Brazil." *Journal of Insect Conservation* 24 (3): 525–34. https://doi.org/10.1007/s10841-020-00219-1.

Simkin, Alfred, Alex Wong, Yu-Ping Poh, William E. Theurkauf, and Jeffrey D. Jensen. 2013. "Recurrent and Recent Selective Sweeps in the PiRNA Pathway." *Evolution; International Journal of Organic Evolution* 67 (4): 1081–90. https://doi.org/10.1111/evo.12011.

Simpson, Jared T., and Mihai Pop. 2015. "The Theory and Practice of Genome Sequence Assembly." *Annual Review of Genomics and Human Genetics* 16: 153–72. https://doi.org/10.1146/annurev-genom-090314-050032.

Smith, A, and R Hubley. 2008. "RepeatModeler Open-1.0." httpwww.repeatmasker.org.

Smith, A, R Hubley, and P Green. 2013. “RepeatMasker Open-4.0.” http://www.repeatmasker.org.

Sonoda, Eiichiro, Helfrid Hochegger, Alihossein Saberi, Yoshihito Taniguchi, and Shunichi Takeda. 2006. “Differential Usage of Non-Homologous End-Joining and Homologous Recombination in Double Strand Break Repair.” *DNA Repair* 5 (9–10): 1021–29. https://doi.org/10.1016/j.dnarep.2006.05.022.

Sperlich, D., and Pfreim. 1986. “Chromosomal Polymorphism in Natural and Experimental Populations.” *The Genetics and Biology of Drosophila* 3e: 257–309.

Spieth, HT. 1969. “Courtship and Mating Behavior of the Drosophila Nasuta Subgroup of Species.” *Univ. Texas Publ.*, no. 6918: 255–70.

Stadler, Michael R., Jenna E. Haines, and Michael B. Eisen. 2017. “Convergence of Topological Domain Boundaries, Insulators, and Polytene Interbands Revealed by High-Resolution Mapping of Chromatin Contacts in the Early Drosophila Melanogaster Embryo.” *ELife* 6 (November). https://doi.org/10.7554/eLife.29550.

Stamatakis, Alexandros. 2014. “RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies.” *Bioinformatics (Oxford, England)* 30 (9): 1312–13. https://doi.org/10.1093/bioinformatics/btu033.

Sturtevant, A. H. 1917. “Genetic Factors Affecting the Strength of Linkage in Drosophila.” *Proceedings of the National Academy of Sciences of the United States of America* 3 (9): 555–58. https://doi.org/10.1073/pnas.3.9.555.

Sturtevant, A. H., and G. W. Beadle. 1936. “The Relations of Inversions in the X Chromosome of Drosophila Melanogaster to Crossing over and Disjunction.” *Genetics* 21 (5): 554–604.

Sundaram, Vasavi, and Joanna Wysocka. 2020. “Transposable Elements as a Potent Source of Diverse Cis-Regulatory Sequences in Mammalian Genomes.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 375 (1795): 20190347. https://doi.org/10.1098/rstb.2019.0347.

Tamura, Koichiro, Sankar Subramanian, and Sudhir Kumar. 2004. “Temporal Patterns of Fruit Fly (Drosophila) Evolution Revealed by Mutation Clocks.” *Molecular Biology and Evolution* 21 (1): 36–44. https://doi.org/10.1093/molbev/msg236.

Tesler, Glenn. 2002. “GRIMM: Genome Rearrangements Web Server.” *Bioinformatics (Oxford, England)* 18 (3): 492–93. https://doi.org/10.1093/bioinformatics/18.3.492.

Traverse, K. L., and M. L. Pardue. 1988. “A Spontaneously Opened Ring Chromosome of Drosophila Melanogaster Has Acquired He-T DNA Sequences at Both New Telomeres.” *Proceedings of the National Academy of Sciences of the United States of America* 85 (21): 8116–20. https://doi.org/10.1073/pnas.85.21.8116.

Turissini, David A., and Daniel R. Matute. 2017. “Fine Scale Mapping of Genomic Introgressions within the Drosophila Yakuba Clade.” *PLoS Genetics* 13 (9): e1006971. https://doi.org/10.1371/journal.pgen.1006971.

Vaser, Robert, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. 2017. “Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads.” *Genome Research* 27 (5): 737–46. https://doi.org/10.1101/gr.214270.116.

Vendrell-Mir, Pol, Fabio Barteri, Miriam Merenciano, Josefa González, Josep M. Casacuberta, and Raúl Castanera. 2019. “A Benchmark of Transposon Insertion Detection Tools Using Real Data.” *Mobile DNA* 10: 53. https://doi.org/10.1186/s13100-019-0197-9.

Vieira, J., C. P. Vieira, D. L. Hartl, and E. R. Lozovskaya. 1997. “Discordant Rates of Chromosome Evolution in the Drosophila Virilis Species Group.” *Genetics* 147 (1): 223–30.

Vilela, Carlos Ribeiro, and Beatriz Goñi. 2015. “Is *Drosophila Nasuta* Lamb (Diptera, Drosophilidae) Currently Reaching the Status of a Cosmopolitan Species?” *Revista Brasileira de Entomologia* 59 (December): 346–50.

https://doi.org/10.1016/j.rbe.2015.09.007.

Villanueva-Cañas, José Luis, Vivien Horvath, Laura Aguilera, and Josefa González. 2019. "Diverse Families of Transposable Elements Affect the Transcriptional Regulation of Stress-Response Genes in Drosophila Melanogaster." *Nucleic Acids Research* 47 (13): 6842–57. https://doi.org/10.1093/nar/gkz490.

Walker, Bruce J., Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, et al. 2014. "Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement." *PloS One* 9 (11): e112963. https://doi.org/10.1371/journal.pone.0112963.

Wang, Chen, and Haifan Lin. 2021. "Roles of PiRNAs in Transposon and Pseudogene Regulation of Germline MRNAs and LncRNAs." *Genome Biology* 22 (1): 27. https://doi.org/10.1186/s13059-020-02221-x.

Wei, Kevin H.-C., Carolus Chan, and Doris Bachtrog. 2021. "Establishment of H3K9me3-Dependent Heterochromatin during Embryogenesis in Drosophila Miranda." *ELife* 10 (June): e55612. https://doi.org/10.7554/eLife.55612.

Wei, Kevin H.-C., Lauren Gibilisco, and Doris Bachtrog. 2020. "Epigenetic Conflict on a Degenerating Y Chromosome Increases Mutational Burden in Drosophila Males." *Nature Communications* 11 (1): 5537. https://doi.org/10.1038/s41467-020-19134-9.

Wei, Kevin H.-C., Hemakumar M. Reddy, Chandramouli Rathnam, Jimin Lee, Deanna Lin, Shuqing Ji, James M. Mason, Andrew G. Clark, and Daniel A. Barbash. 2017. "A Pooled Sequencing Approach Identifies a Candidate Meiotic Driver in Drosophila." *Genetics* 206 (1): 451–65. https://doi.org/10.1534/genetics.116.197335.

Wessler, S. R. 1996. "Turned on by Stress. Plant Retrotransposons." *Current Biology: CB* 6 (8): 959–61. https://doi.org/10.1016/s0960-9822(02)00638-3.

Wilson, FD, MR Wheeler, M Harget, and M Kambysellis. 1969. "Cytogenetic Relations in the Drosophila Nasuta Subgroup of the Immigrans Group of Species." *Univ. Texas Publ.*, no. 6918: 207–53.

Wong Miller, Karen M., Ryan R. Bracewell, Michael B. Eisen, and Doris Bachtrog. 2017. "Patterns of Genome-Wide Diversity and Population Structure in the Drosophila Athabasca Species Complex." *Molecular Biology and Evolution* 34 (8): 1912–23. https://doi.org/10.1093/molbev/msx134.

Wu, Ying, Yue Sun, Kun Shen, Shuai Sun, Jie Wang, Tingting Jiang, Shuai Cao, et al. 2015. "Immediate Genetic and Epigenetic Changes in F1 Hybrids Parented by Species with Divergent Genomes in the Rice Genus (Oryza)." *PloS One* 10 (7): e0132911. https://doi.org/10.1371/journal.pone.0132911.

Xiao, Y. L., X. Li, and T. Peterson. 2000. "Ac Insertion Site Affects the Frequency of Transposon-Induced Homologous Recombination at the Maize P1 Locus." *Genetics* 156 (4): 2007–17.

Xing, Jinchuan, Hui Wang, Victoria P. Belancio, Richard Cordaux, Prescott L. Deininger, and Mark A. Batzer. 2006. "Emergence of Primate Genes by Retrotransposon-Mediated Sequence Transduction." *Proceedings of the National Academy of Sciences of the United States of America* 103 (47): 17608–13. https://doi.org/10.1073/pnas.0603224103.

Yang, Hsiao-Pei, and Daniel A. Barbash. 2008. "Abundant and Species-Specific DINE-1 Transposable Elements in 12 Drosophila Genomes." *Genome Biology* 9 (2): R39. https://doi.org/10.1186/gb-2008-9-2-r39.

Yu, H., W. Wang, S. Fang, Y. P. Zhang, F. J. Lin, and Z. C. Geng. 1999. "Phylogeny and Evolution of the Drosophila Nasuta Subgroup Based on Mitochondrial ND4 and ND4L Gene Sequences." *Molecular Phylogenetics and Evolution* 13 (3): 556–65. https://doi.org/10.1006/mpev.1999.0667.

Zhang, Chao, Maryam Rabiee, Erfan Sayyari, and Siavash Mirarab. 2018. "ASTRAL-III: Polynomial Time Species Tree Reconstruction from Partially Resolved Gene Trees."

BMC Bioinformatics 19 (Suppl 6): 153. https://doi.org/10.1186/s12859-018-2129-y.

Zhang, Hua-Hao, Jean Peccoud, Min-Rui-Xuan Xu, Xiao-Gu Zhang, and Clément Gilbert. 2020. “Horizontal Transfer and Evolution of Transposable Elements in Vertebrates.” *Nature Communications* 11 (1): 1362. https://doi.org/10.1038/s41467-020-15149-4.

Zhang, Jianbo, Chuanhe Yu, Lakshminarasimhan Krishnaswamy, and Thomas Peterson. 2011. “Transposable Elements as Catalysts for Chromosome Rearrangements.” *Methods in Molecular Biology (Clifton, N.J.)* 701: 315–26. https://doi.org/10.1007/978-1-61737-957-4_18.

Zhang, Zhang, Jun Li, Xiao-Qian Zhao, Jun Wang, Gane Ka-Shu Wong, and Jun Yu. 2006. “KaKs_Calculator: Calculating Ka and Ks through Model Selection and Model Averaging.” *Genomics, Proteomics & Bioinformatics* 4 (4): 259–63. https://doi.org/10.1016/S1672-0229(07)60007-2.

Zhao, Shaying, Jyoti Shetty, Lihua Hou, Arthur Delcher, Baoli Zhu, Kazutoyo Osoegawa, Pieter de Jong, William C. Nierman, Robert L. Strausberg, and Claire M. Fraser. 2004. “Human, Mouse, and Rat Genome Large-Scale Rearrangements: Stability Versus Speciation.” *Genome Research* 14 (10a): 1851–60. https://doi.org/10.1101/gr.2663304.

Zhou, Qi, and Doris Bachtrog. 2012. “Chromosome-Wide Gene Silencing Initiates Y Degeneration in Drosophila.” *Current Biology: CB* 22 (6): 522–25. https://doi.org/10.1016/j.cub.2012.01.057.

———. 2015. “Ancestral Chromatin Configuration Constrains Chromatin Evolution on Differentiating Sex Chromosomes in Drosophila.” *PLoS Genetics* 11 (6): e1005331. https://doi.org/10.1371/journal.pgen.1005331.

Zhou, Qi, Christopher E. Ellison, Vera B. Kaiser, Artyom A. Alekseyenko, Andrey A. Gorchakov, and Doris Bachtrog. 2013. “The Epigenome of Evolving Drosophila Neo-Sex Chromosomes: Dosage Compensation and Heterochromatin Formation.” *PLoS Biology* 11 (11): e1001711. https://doi.org/10.1371/journal.pbio.1001711.

Zhou, Qi, Hong-mei Zhu, Quan-fei Huang, Li Zhao, Guo-jie Zhang, Scott W. Roy, Beatriz Vicoso, et al. 2012. “Deciphering Neo-Sex and B Chromosome Evolution by the Draft Genome of Drosophila Albomicans.” *BMC Genomics* 13 (March): 109. https://doi.org/10.1186/1471-2164-13-109.