bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Evolution of genome structure in the *Drosophila simulans* species complex

Mahul Chakraborty*[1], Ching-Ho Chang*[2], Danielle E. Khost[2,3], Jeffrey Vedanayagam[4], Jeffrey R. Adrion[5], Yi Liao[1], Kristi Montooth[6], Colin D. Meiklejohn[6], Amanda M. Larracuente[2], J.J. Emerson[1]

*these authors contributed equally to this work

Affiliations:

[1]Department of Ecology and Evolutionary Biology, University of California Irvine, Irvine, CA 92697

[2]Department of Biology, University of Rochester, Rochester, NY 14627

[3]FAS Informatics and Scientific Applications, Harvard University, Cambridge, MA 02138

[4]Department of Developmental Biology, Memorial Sloan-Kettering Cancer Center, New York, NY, 10065

[5]Institute of Ecology and Evolution, University of Oregon, Eugene, Oregon 97403

[6]School of Biological Sciences, University of Nebraska-Lincoln, Nebraska 68502

correspondence to alarracu@ur.rochester.edu, jje@uci.edu

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ABSTRACT

The rapid evolution of repetitive DNA sequences, including heterochromatic regions, satellite DNA, tandem duplications, and transposable elements, can underlie phenotypic evolution and contribute to hybrid incompatibilities between species. However, repetitive genomic regions are fragmented in most contemporary genome assemblies. We generated highly contiguous *de novo* assemblies for the *Drosophila simulans* species complex (*D. simulans, D. mauritiana,* and *D. sechellia*), which speciated ~250,000 years ago. These species diverged from their common ancestor with *D. melanogaster* ~3 million years ago. Our assemblies are comparable in contiguity and accuracy to the current *D. melanogaster* genome, allowing us to directly compare repetitive regions in genomes across different evolutionary times. We find a rapid turnover of satellite DNA and extensive structural variation in heterochromatic regions, while the euchromatic gene content is mostly conserved. Despite the overall preservation of synteny, euchromatin of each species has been sculpted by clade and species-specific inversions, transposable elements (TE), satellite and tRNA tandem arrays, and gene duplications. We also find Y-linked genes rapidly diverging, in terms of copy number and recent duplications from the autosomes. Our assemblies provide a valuable resource for studying genome evolution and its consequences for phenotypic evolution in these genetic model species.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

INTRODUCTION

The group of four fruit fly species composed of *D. melanogaster, D. simulans, D. sechellia* and *D. mauritiana* is collectively known as the *D. melanogaster* species complex (or mel-complex for short) (Hey and Kliman 1993) and serves as a model system for studying speciation (Tao et al. 2001; Wu 2001; Meiklejohn et al. 2018), behavior (Ding et al. 2019), population genetics (Kliman et al. 2000; Begun et al. 2007; Garrigan et al. 2012), and molecular evolution (Moriyama and Powell 1997). The members of the melcomplex are closely related and are ~96% identical in shared genomic regions (Begun et al. 2007; Garrigan et al. 2012); indeed, they are so similar as to be morphologically indistinguishable to the untrained eye. IAnd yet, despite these similarities, they exhibit profound biological differences. *D. melanogaster* and the three *D. simulans* complex species are reproductively isolated, with either sterile or lethal hybrids (Barbash 2010). They also exhibit unique ecological adaptations: *D. sechellia* larvae specialize on a fruit toxic to the other three species (R'Kha et al. 1991) whereas *D. melanogaster* can thrive in ethanol concentrations lethal to the *D. simulans* species complex (Merçot et al. 1994). Recent studies suggest that these species are more distinct at the genetic level than previously appreciated, due to variation in poorly resolved or inaccessible repetitive regions of the existing genome assemblies of the *D. simulans* species complex (Chakraborty et al. 2018; Miller et al. 2018). In both humans and fruit flies, genetic variation comprised of repetitive sequences such as transposons, satellites, and duplications affect more of the genome than all single nucleotide variants (SNVs) combined (Chakraborty et al. 2018; 1000 Genomes Project Consortium et al. 2015). Moreover, such variants often exhibit large fitness effects, underlie ecological adaptations, or are involved in genomic conflicts (e.g. (Daborn et al. 2002; Montchamp-Moreau et al. 2006; Tao et al. 2007b, 2007a; Fishman and Saunders 2008; Larracuente and Presgraves 2012; Van't Hof et al. 2016; Battlay et al. 2018; Chakraborty et al. 2019)).

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Repetitive sequences comprise a substantial fraction of the genomes of multicellular eukaryotes, occupying ~50% of human genomes and ~40% of *Drosophila melanogaster* genomes (Treangen and Salzberg 2011; Hoskins et al. 2015; Chang and Larracuente 2019). These sequences include repeated tandem arrays of non-coding sequences like satellite DNAs, self-replicating selfish elements like transposable elements (TEs), and duplications of otherwise unique sequences, including genes. Within eukaryotic genomes, there are critical structures essential for cell organization and function that consist primarily or entirely of repetitive DNA sequences. For example, chromosome segregation depends on centromeres, which are comprised of various combinations of satellites and/or transposable elements (Klein and O'Neill 2018; Chang et al. 2019; Hartley and O'Neill 2019). Similarly, telomeres, which maintain chromosome integrity, are made up of telomere-specific tandem repeats (Moyzis et al. 1988) or transposable elements (Mason et al. 2008). Additionally, protein translation requires ribosomal RNAs and tRNAs encoded in large tandem arrays of DNA (Hillis and Dixon 1991). Short tandem repeats near protein-coding genes can recruit transcription factors to regulate gene expression (Rockman and Wray 2002; Gemayel, et al. 2010). Finally, euchromatic satellite repeats contribute to X chromosome recognition during dosage compensation in *Drosophila* males through the RNA interference pathway (Menon and Meller 2012; Menon et al. 2014).

Repeats can also have major effects on genome evolution. Repetitive sequences can selfishly proliferate and create conflicts within genomes (Doolittle and Sapienza 1980; Orgel and Crick 1980), and these intragenomic conflicts can trigger wider evolutionary arms races within and between genomes (Werren et al. 1988; Aravin et al. 2007; Ellis et al. 2011; Cocquet et al. 2012; Blumenstiel 2019; Parhad and Theurkauf 2019; Rathje et al. 2019). For example, discrete genomic loci containing TE fragments produce small, Piwi-interacting RNAs (piRNA) to silence TEs (Brennecke et al. 2007), driving evolutionary arms races between selfish elements and their host genomes (but see (Cosby et al. 2019)). Tandemly repeated DNAs have been found to be involved in selfish meiotic drive systems that cheat meiosis to bias their transmission to the next

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

generation (reviewed in (Lindholm et al. 2016)), and these drivers can then trigger the rapid evolution of repeats within and between genomes (Cocquet et al. 2012; Lindholm et al. 2016; Rathje et al. 2019). Repeats at centromeres can drive in female meiosis, causing rapid evolution of centromere proteins to restore parity (Henikoff et al. 2001). Repeats are also targets of meiotic drivers in the male germline. For example, the selfish *Segregation Distorter* gene complex of *D. melanogaster* kills sperm bearing a large block of a satellite repeat (*Responder*; *Rsp*) during spermatogenesis, thus gaining a transmission advantage to the next generation (reviewed in (Larracuente and Presgraves 2012)). This conflict between driver and target may cause the rapid evolution of *Rsp* over short evolutionary time scales (Cabot et al. 1993; Larracuente 2014). The lack of recombination and male-limited transmission of Y chromosomes also create opportunities for conflicts to form— *e.g.*, antagonistic gene families on mammalian sex chromosomes (Cocquet et al. 2012; Kruger et al. 2019) and X-linked meiotic drive in Drosophila genomes (reviewed in (Jaenike 2001)). These conflicts impose strong selection pressures on Y chromosomes that may trigger the rapid turnover of Y-linked repeats (Lohe and Roberts 1990; Bachtrog 2004; Larracuente and Clark 2013; Wei et al. 2018).

While TEs are often regarded as genomic parasites, they can also fuel genome innovation and contribute to adaptive evolution. For example, acquisition of xenobiotic resistance (e.g., insecticide resistance) is often accomplished via complex structural mutations that include recurrent duplications and transposable element insertions at many loci (Aminetzach et al. 2005; Schmidt et al. 2010; Van't Hof et al. 2016; Chakraborty et al. 2018, 2019). Mounting evidence suggests TEs also participate in commensal and mutualistic interactions, leading to less antagonistic modes of coevolution of TEs with their hosts (reviewed in (Cosby et al. 2019).

The very nature of repetitive sequences makes them difficult to study. Genome assemblies cannot reliably resolve repetitive segments that substantially exceed the length of the sequence reads from which they are constructed (Alkan et al. 2011b). As a consequence, whole-genome shotgun sequencing approaches based on reads shorter

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

than common repeats yield erroneous, fragmented, and incomplete genome assemblies, which are particularly unreliable in repetitive regions (Hoskins et al. 2015, 2002). Combined with the fact that historically, reference-quality genomes were only distantly related to each other, understanding the interspecies evolutionary dynamics of sequences exhibiting such high rates of mutation and rapid turnover is challenging (Plohl et al. 2012; Lower et al. 2018; Blumenstiel 2019). Consequently, until recently, comparative genomics could offer only limited insight into the evolution of the repetitive genome. The advent of sequencing technologies that produce reads longer than common repeats solves these problems. Long-read based assemblies can be nearly complete, contiguous, and accurate (Steinberg et al. 2014; Berlin et al. 2015; Chaisson et al. 2015; Chakraborty et al. 2016, 2018; Solares et al. 2018; Chang and Larracuente 2019).

To understand how changes in repeated and copied sequences affect the evolutionary dynamics of genome structure, we sequenced and assembled reference-quality genomes of *D. simulans*, *D. sechellia*, and *D. mauritiana*. These three species, collectively known as the *Drosophila simulans* complex (or sim-complex for short; (Kliman et al. 2000)), comprise the nearest sister species to *D. melanogaster*, and are virtually equally related to each other and *D. melanogaster* (Fig. 1A), probably as a consequence of rapid speciation (Garrigan et al. 2012; Pease and Hahn 2013). The lineages that would give rise to the sim-complex and *D. melanogaster* diverged approximately 3 mya, while the sim-complex species diverged from one another approximately 250,000 years ago (Lachaise et al. 1986; Kliman et al. 2000; McDermott and Kliman 2008; Garrigan et al. 2012) and still hybridize in nature (Matute and Ayroles 2014). We used high-coverage single-molecule real-time sequencing to assemble the sim-complex species genomes *de novo*, permitting us to resolve repetitive sequences that have, until now, evaded scrutiny. These assemblies are comparable in completeness and contiguity to the latest release of the *D. melanogaster* reference genome, arguably the highest quality metazoan genome available. For each species, we annotated repetitive sequences, including TEs, satellites, tRNAs, and tandem

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

duplicates. We find that structural divergence is a dynamic process that rapidly leads to substantial variation between closely related species. For example, while TEs make up a substantial proportion of mel-complex genomes, very few are shared between species, indicating a rapid turnover of TE content. We also find extensive genomic rearrangements inside heterochromatic regions, including pericentromeric regions and the Y chromosome. We discover sim-complex specific tandem gene duplicates that likely acquired function. Altogether, we uncover a surprisingly dynamic picture of repeat evolution that leads to extensive genome variation on short timescales.

RESULTS

Contiguous, accurate, and complete assemblies resolve previous misassemblies

Highly contiguous, accurate genome assemblies permit comprehensive detection of genome-wide variation (Alkan et al. 2011a) and such assemblies are attainable with high coverage (~100X) long reads (Chakraborty et al. 2016; Koren et al. 2017; Chakraborty et al. 2018). To understand the contributions of repetitive and complex genomic regions towards genome and species diversification, we collected deep (100-150 fold autosomal coverage) long read sequence data from males (supplemental Fig. S1–2; supplemental Table S1) using Single Molecule Real Time Sequencing by Pacific Biosciences. We used these reads to create reference-quality *de novo* genome assemblies for three closely related *Drosophila* species: *D. simulans* (contig N50 = 22.7Mb; N50 = sequence of this length or longer comprise 50% of the assembly), *D. sechellia* (contig N50 = 21.2Mb), and *D. mauritiana* (contig N50 = 22.9Mb; Table 1, Fig. 1B; supplemental Fig. S3). Our assemblies are as contiguous as the *D. melanogaster* reference (contig N50 = 21.3 Mb; (Hoskins et al. 2015), which is regarded as the gold standard for metazoan genome assemblies. In all three species, the majority of each chromosome arm assembled into single contigs, including the entirety of the

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

euchromatin and large stretches of pericentric heterochromatin, revealing extensive structural genetic divergence (Fig. 1B, Fig. 2, supplemental Fig. S4). Our assemblies are less contiguous inside the pericentric heterochromatin because these genomic regions consist mainly of satellites and transposable elements and therefore require specialized assembly approaches (Khost et al. 2017; Chang and Larracuente 2019). Nonetheless, we assembled more than 20Mbp of pericentric heterochromatin, revealing structural rearrangements and divergence at these previously inaccessible complex genomic regions (Fig. 2A).

A comparison of our assemblies to the *D. melanogaster* genome shows contiguous synteny across the major chromosome arms (Fig. 2; supplemental Fig. 4), except large inversions on 3R and the 4th chromosome that were previously known (Podemski et al. 2001; Schaeffer et al. 2008). Consequently, large errors in our assemblies are unlikely, a conclusion which is further supported by the evenly distributed long-read coverage (supplemental Fig. S1–2) and mapping of BAC sequences across the assembled chromosomes (supplemental Fig. S5; supplementary text). Our assemblies are also highly accurate at the nucleotide level, as suggested by the QV estimates, which (QV = 44.0–46.3) match the *D. melanogaster* reference genome sequence accuracy (QV = 44.3; supplemental Table S2). Evaluation of assembly completeness based on a set of 2,799 single copy conserved Dipteran orthologs (BUSCO; (Simão et al. 2015) show that our assemblies contain as many conserved orthologs (98.8–99% BUSCO) as the *D. melanogaster* reference genome (98.6% BUSCO; supplemental Table S3).

In addition to assembling previously unassembled regions, we also corrected errors previously noted in the draft assemblies of these species (Schaeffer et al. 2008). For example, the ~350 Kb subtelomeric fragment from 3L that was misassembled onto the 2R scaffold in the *D. simulans* reference assembly (Schaeffer et al. 2008) is correctly placed in our *D. simulans* 3L assembly. Moreover, our *D. sechellia* assembly corrects several chimeric sequences, including the sequences involving the 3L, 3R and

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

X and misjoins between 2L, 3R, and 2R sequences in the previous assemblies of this species (Schaeffer et al. 2008).

In addition to the Drosophila genomes, we assembled 1.3 Mb contigs representing the entire *Wolbachia* genomes from *D. mauritiana* and a *D. sechellia.* Our *D. simulans* w^{XD1} strain does not carry *Wolbachia*. The Wolbachia genomes we assembled are comparable in size to the complete Wolbachia genomes previously assembled from *D. melanogaster* (GenBank ID AE017196.1) and *D. simulans* (CP003884.1 and CP003883). The *Wolbachia* in *D. sechellia* (wSech) and *D. mauritiana* (wMau) share more than 95% identity with wHA (supergroup A) and wNA (supergroup B) sequences from *D. simulans,* respectively. Notably, while we observed only one inversion between wHA and wSech (supplemental Fig. S6A), we found 15 genome rearrangement events between wNO and wMau under double-cut-and-join (DCJ) model (supplemental Fig. S6B)(Lin and Moret 2008). We found that more reads mapped to Wolbachia in *D. mauritiana* compared to *D. sechellia*–13.2x more long reads from male flies and 1,297x more Illumina reads from female flies (supplemental Table S1 and supplemental Fig. S7), which might reflect differences *in vivo* titers. We detected no other symbionts in *D. mauritiana* but found 28 bacterial contigs from *Providencia* species, a known *Drosophila* pathogen (Juneja and Lazzaro 2009), in *D. sechellia* (supplemental Table S4).

Species and clade-specific genomic rearrangements

Whole-genome alignments of the four species reveal large-scale synteny: euchromatic arms are largely collinear between the three *D. simulans* species and *D. melanogaster* (Fig. 2). Despite this, we uncover several species- and clade-specific chromosomal inversions that are located in previously un-assembled euchromatin as well as within the repeat-dense regions approaching pericentromeric heterochromatin (Fig. 2).

We first compared the genomic rearrangement rates between species and in different genomic regions. We aligned the genomes using Mauve (which builds locally collinear blocks) and then estimated the number of large genomic rearrangements

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

within the genomes (Lin and Moret 2008). Using these locally collinear blocks, we discovered 535–542 rearrangements between *D. melanogaster* and species in *D. simulans* complex (~90 mutations per million-year), and 118–182 genomic rearrangements between species in *D. simulans* complex (236–364 mutations per million-year; supplemental Table S5). Within the euchromatin, rearrangements strongly enriched on the X chromosome: more than 60% of all rearrangements between *D. melanogaster* and the sim-complex species, and all but one euchromatic rearrangement within the sim-complex species are X-linked (Fig 2A; supplemental Table S5). Among all genomic rearrangements, more than 95% are located in heterochromatic regions.

To identify different types of genomic rearrangement with higher sensitivity, we applied MUMmer and svmu, which directly analyze MUMs without building locally collinear blocks. Within the euchromatin boundaries, *D. simulans*, *D. mauritiana*, and *D. sechellia* differ by 23, 25, and 21 inversions, respectively, ranging up to 13.6 Mbp. Nine inversions are shared among all three *D. simulans* species complex , 4 of which are also present in the outgroup species *D. yakuba*. This suggests that these 4 mutations likely occurred in the *D. melanogaster* lineage, while the other five occurred on the lineage leading to the sim-complex species. One of these, a 13.6 Mbp inversion (3R:8049180–21735108) on 3R that was previously identified based on cytological evidence (Sturtevant and Plunkett 1926), is contained within a single contig in all three assemblies (Fig. 2). The sim-mau, sim-sech, mau-sech species pairs share 5, 3, and 4 euchromatic inversions, respectively, suggesting that they were polymorphic in the common ancestor of the sim-complex species. Notably, a 460-kb inversion on the X chromosome shared by *D. sechellia* (X:8744323–9203725) and *D. mauritiana* (X:8736133–9203526) spans 45 protein-coding genes, including several genes involved in chromatin assembly and germline development (HP1b, APC4, His3.3B, and mei-P26; Fig. 2, supplemental Fig. S8).

It is more difficult to confidently determine the nature of rearrangements in heterochromatic regions due to a lack of large syntenic sequence blocks, but we do observe evidence for two large (> 100-Kb) inversions within pericentromeric

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

heterochromatin. One such inversion appears to be pericentric, wherein an inverted segment from 3R pericentric heterochromatin is now attached to 3L pericentric heterochromatin (Fig. 2, Fig. S9A–D). Inspection of the outgroup *D. erecta* and *D. yakuba* genomes suggests that the inversion most likely occurred in the *D. melanogaster* lineage. We also observed a ~700-kb inversion in the X heterochromatin of sim-complex species that spans 35 genes (22.4–23.1Mb on *D. melanogaster* X; Fig. 2, supplemental Fig. S8). The FlyBase reference assemblies of *D. yakuba* and *D. erecta* do not have any evidence of this X inversion (dos Santos et al. 2015), suggesting that this inversion is sim-complex specific. We also find large species-specific heterochromatic inversions on *D. mauritiana* 2R and *D. sechellia* 3R (Fig. 2, supplemental Fig. S9–10).

Repetitive DNA

We annotated repetitive DNA across the genome using our custom repeat library (supplementary File S1). To examine the repeats underlying genome differences in the *D. simulans* complex, we used our annotations to estimate repeat abundance, and confirmed the estimates by mapping Illumina reads to the library (see Methods). For the major arms, we plotted the distribution of repetitive elements across the scaffolds (Fig. 3). We used the euchromatin-heterochromatin boundaries in *D. melanogaster* to infer these boundaries in the *D. simulans* species complex (Hoskins et al. 2015). The density of repetitive elements increases approaching the euchromatin-heterochromatin boundary, consistent with our expectations. Below we detail the patterns of the different classes of repetitive elements.

Distribution of satellites

Just beyond the highly repetitive regions at the ends of the scaffolds are large blocks of tandem repeats classified as satellite DNAs. These satellite DNAs differ greatly in genomic distribution between *D. melanogaster* and the sim-complex, and even between the sim-complex species (Jagannathan et al. 2017). In addition to previously known

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

satellites, we identified several novel complex satellite arrays in the *D. simulans* complex species using Tandem Repeat Finder (TRF), selecting satellites with a monomer size ≥ 10 and an array size of at least 30-kb, which we named for their monomer size (90U, 193XP, and 500U). One of the satellites we describe here is 500U, which are located primarily on unassigned contigs and near the centromeres (Chang et al. 2019; Talbert et al. 2018). We find the other two satellites, 90U and 193XP, near the ribosomal DNA (rDNA) loci. The 90U satellite is closely associated with rDNA arrays, where smaller blocks of this repeat correspond to one of the non-transcribed spacer (NTS) subunits (Stage and Eickbush 2007). 90U repeats are located directly adjacent to the 28S rDNA subunit and the 240-bp NTS repeat sequence, both on the scaffolded portion of the rDNA locus on the X chromosome as well as on numerous unassigned contigs. There is also a large 193XP locus in the pericentromeric heterochromatin adjacent to the rDNA locus but is not part of the locus itself. In *D. simulans* and *D. mauritiana*, the 193XP locus is approximately 79kb and 48kb, respectively, with LTR insertions towards the edges of the arrays. In both species, the centromere-proximal side of the locus is separated from the rDNA locus by a gap, so the actual size of the 193XP loci could be much larger. The scaffolded locus in *D. sechellia* is much smaller, only about 3.4kb; however, it also is adjacent to a gap and dense 193XP repeats are found on unassigned contigs, so it also could be much larger in this species. The 193XP locus is shared across the sim-complex but is absent in outgroup species, *D. melanogaster*, *D. erecta* and *D. yakuba*, suggesting that it arose in the ancestor of the *D. simulans* complex. Consistent with our assemblies, we detect fluorescence *in situ* hybridization signal for 193XP only on the X pericentromere in the sim-complex (supplemental Fig. S11).

In addition to the large blocks of heterochromatic satellite DNA, we find satDNA arrays in the euchromatin (supplemental Table S6) as previously reported (Kuhn et al. 2012; Gallach 2014; Waring and Pollack 1987; DiBartolomeis et al. 1992). Some X-linked euchromatic satDNAs may have functional roles. For example, a subset of 1.688 repeats in the X euchromatin of *D. melanogaster* play a role in dosage

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

compensation (Menon et al. 2014). Similar to *D. melanogaster*, euchromatic satDNA repeats are enriched on the X chromosome relative to the autosomes in all three sim-complex species (supplemental Table S6). Satellites on the autosomes only comprise ~0.07% of the total bases in the euchromatin, while they comprise 1% total bases on the X chromosome in *D. melanogaster* and *D. simulans*, and up to 2.4% in *D. mauritiana* and more than 3.4% in *D. sechellia*. The latter is a minimum estimate because the *D. sechellia* X chromosome contains 6 gaps in the euchromatic satellite regions. The location, abundance, and composition of euchromatic satellites differ between each species. For example, a complex satellite repeat called *Rsp-like* (Larracuente 2014) expanded recently in *D. simulans* and *D. mauritiana*, moving to new genomic locations across the X euchromatin. The new *Rsp-like* repeats inserted into existing arrays of *1.688*–another type of complex satellite repeat. Interestingly, the locations of large blocks of *Rsp-like* and *1.688* exist in different genomic locations in the heterochromatin of each of these species (Larracuente 2014), indicating that satellite repeats are dynamic in both the heterochromatin and euchromatin (Sproul et al. 2019).

Transposable elements

Transposable elements (TEs) are abundant dispersed repeats that are problematic for genome assembly with traditional shotgun sequencing approaches. In order to quantify the dynamics of TEs over short evolutionary time scales, we estimated TE content in *D. melanogaster* and the three sim-complex species (see Methods). We restricted our analyses to euchromatic TE sequence, since accurately assembling heterochromatic regions remains challenging (Khost et al. 2017). Unless otherwise noted, our results are based on comparisons of TE content (i.e. number of bases) rather than number of TE insertions (i.e. number of events).

We find that the sim-complex genomes host 67–83% as much TE sequence as *D. melanogaster* (Fig. 4A). The most striking difference in TE content between *D. melanogaster* and the sim-complex species is the enrichment of LTR retrotransposons in *D. melanogaster* (Kaminker et al. 2002; Bergman and Bensasson 2007), which

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

carries between 1.3 and 1.8 Mbp more LTR content than the three sim-complex species (Fig. 4A). DNA transposon and non-LTR content in *D. melanogaster* are similar to that of the sim-complex species (Fig. 4A). Most TE bases in the sim-complex genomes are specific to each species (Fig. 4A): between 66 and 72% of TE sequence content in these three species is not found in the other two, or in *D. melanogaster*. This is consistent with dynamic turnover of the TE content in these genomes, with most TE-derived sequence the result of recent transposable element activity, and older insertions subject to mutational decay and deletion (Drosophila 12 Genomes Consortium et al. 2007; Lerat et al. 2011; Bargues and Lerat 2017).

We also find that TE composition differs across the lineages connecting these four species (Fig. 4B, supplemental Fig. S12). Among the TE content shared by all members of the mel-complex, sequence derived from non-LTR retrotransposons is the most prevalent (52%), followed by DNA transposons (30%) and LTR retrotransposons (18%). In contrast, TE sequence shared by all three sim-complex species but not *D. melanogaster* are dramatically enriched in DNA transposons, which make up 71% of the shared TE sequence among the sim-complex species (Fig. 4B) despite being shorter than other TE classes (supplemental Fig. S13). LTR and non-LTR retrotransposons contribute similar proportions of sequence to the total amount of shared sim-complex TE sequence (13% and 16%, respectively). This enrichment of DNA transposons shared among all three sim-complex species is largely attributable to a single subclass of DNA transposon called INE-1 (also called DINE-1 or DNAREP1) (Yang and Barbash 2008). The relative proportion of INE-1-derived sequence is much higher in the branch leading to the sim-complex (46%) than that leading to the mel-complex (13.7%). These observations suggest that there was a burst of INE-1 activity in the sim-complex ancestor after diverging from their common ancestor with *D. melanogaster*. In contrast, the composition of species-specific TE branches is dominated by LTR elements (48-57%) followed by non-LTR elements (27-40%) and a smaller contribution of DNA elements (12-16%) (Fig. 4B). A low proportion of older, shared LTRs suggests that most LTR elements in these genomes are young and species-specific (Bergman and

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Bensasson 2007). In contrast, non-LTR retrotransposon sequence, which is enriched among older TE-derived sequence, appear to be retained at much higher rates in all four species.

Consistent with the deleterious nature of exonic TE insertions (Cridland et al. 2013; Chakraborty et al. 2019), most (82%) TEs inside the ~7,000 genes annotated with single molecule sequencing of full length mRNA (see Methods) in *D. simulans* are located within introns (Table 2). A similar percentage (94%) of genic TE sequences that are shared by all three sim-complex species are intronic. In contrast, among genic TEs found in all four mel-complex species, exonic TE sequence outnumbers intronic sequence (52% versus 48%). The excess of exonic TE bases acquired on the mel-complex ancestral branch compared to the sim-complex ancestral branch is mainly due to the higher abundance of non-LTR retrotransposons in the former (supplemental Fig. S14, Fig. 4B). In contrast, the DNA elements contribute similar amounts of bases to exonic TEs acquired on both of these ancestral branches (supplemental Fig. S14).

Intronic Indels

We analyzed 21,860 introns in 6,289 orthologous genes with conserved annotation positions in all four mel-complex species. Consistent with previous studies (Comeron and Kreitman 2000; Ometto et al. 2005; Presgraves 2006), we found that introns are significantly shorter in the sim-complex compared to *D. melanogaster* (supplemental Table S7). The longer introns in *D. melanogaster* appear to result from at least two distinct processes. First, introns without transposons or complex satellite sequences (‘simple introns’) are significantly longer in *D. melanogaster* than the other three species (paired t-tests, all $P < 1e-7$, supplemental Fig. S15), but the difference in mean intron length is less than 3 bp (supplemental Table S7). This small difference likely results from differential insertion/deletion bias between *D. melanogaster* (inferred from polymorphic indels (DGRP; (Huang et al. 2014)) and *D. simulans* ((Signor et al. 2018); see supplementary text)). Second, introns that contain TE-derived sequences or complex satellites (‘complex introns’) are, on average, ~700 bp longer in *D.*

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

melanogaster (paired t-tests, all P < 0.001; supplemental Fig. S16), in part due to longer intronic TEs in *D. melanogaster* (mean TE length = 4132 bp) than in the sim-complex species (mean TE lengths of *D. simulans* = 2429 bp, *D. mauritiana* = 2429 bp, *D. sechellia* = 2287 bp; supplemental Fig. S16).

Tandem duplication

To identify tandem duplications shared in the sim-complex species, we aligned each sim-complex genome to the *D. melanogaster* reference genome using MUMmer (Marçais et al. 2018) and LASTZ. Adjacent alignments that are overlapping in the *D. melanogaster* genome but non-overlapping in the sim-complex genomes were annotated as a duplication in the latter (see Methods; supplemental Table S8). We found 97 duplications within the euchromatin shared among *D. simulans*, *D. mauritiana*, and *D. sechellia*. Among these, at most 11 overlapped with duplications present in the outgroup *D. yakuba*, suggesting that at least 86 duplications are shared only among the *sim-complex* and presumably originated during the ~2.5 million years of separation between the MRCA of mel-complex and the MRCA of the sim-complex species. This indicates that duplications have originated in this clade at a rate of ~34 per million years. 72% of these duplications (62/86) overlap exons, 37% (32/86) overlap complete protein coding sequence, and 15% (13/86) overlap full-length *D. melanogaster* genes. In total, 32 complete genes were duplicated, suggesting that new genes arose by duplication in the MRCA of the sim-complex at the rate of 12.8 genes/million years. The X chromosome carries both an excess of duplicated genes (23 X-linked, 9 autosomal; $P < 2.2 \times 10^{-16}$, proportion test against the null of the proportion of X-linked genes in the genome, 0.158) as well as duplicates overlapping full genes (6/13, $P = 0.01$, proportion test against the null of the proportion of X-linked genes in the genome, 0.158), relative to the autosomes.

Several duplication events involve genes that are associated with phenotypes linked to speciation or species divergence: spermatogenesis (*nsr*; *(Ding et al. 2010)*), meiosis (*cona*), odorant binding (*obp18a*), chromosome organization (*HP1D3csd*),

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cytoskeleton organization (*RhoGAP18B*). Not all of these duplications are present in the previous assemblies of the sim-complex species. For example, we discovered a new 3,324 bp duplication that copied the genes *maternal haploid* (*mh*) and *Alg14*. *Mh* maintains genome integrity by interacting with the 359 bp satellite repeats and is essential for proper distribution of the parental genomes during zygotic cell division in *D. melanogaster* (Loppin et al. 2001; Delabaere et al. 2014; Tang et al. 2017). Analysis of *D. mauritiana* and *D. simulans* RNA-seq reads from our strains and iso-seq reads from another *D. simulans* strain (Nouhaud 2018) suggests that the transcript of *mh-d* is shorter than *mh-p* (p and d indicate proximal and distal with respect to the centromere, respectively), with *mh-p* retaining the ancestral gene structure (Fig. 5A and supplemental Fig. S17). The shortened *mh-d* transcript encodes a shorter protein sequence compared to *mh-p* or the ancestral *mh* (supplemental Fig. S18). *Mh-p*, consistent with its essential function during zygotic cell division, has female-biased expression, whereas *mh-d* has testis-biased expression (Fig. 5B and supplemental Fig. S17), suggesting that *mh-p* may have acquired a new male-specific function in *D. simulans* and *D. mauritiana*.

We also uncovered a new 4,654 bp duplication inside an inverted segment (supplementary Fig. 19) of the pericentric heterochromatin on the *D. simulans* complex X chromosome. The duplication created a partial copy of the gene *suppressor of forked* or *su(f)* (supplementary Fig. S19). This duplication is not present in the current FlyBase reference genomes of *D. simulans* (release 2.02) and *D. sechellia* (release 1.3) or the annotated published genome of *D. mauritiana* (Garrigan et al. 2014). The proximal *su(f)* copy is missing the first 12 codons from the 5' end of the ORF. Since we do not yet know whether this copy is transcribed or the structure of any putative transcripts, we cannot determine if there is now a new AUG codon upstream of the missing 12 codons. However, the proximal copy does retain the rest of the ORF of the ancestral mel-complex *su(f)* coding sequence, including the stop codon. Comparison of the *su(f)* copies gives no evidence that the proximal copy is evolving under relaxed selection

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

relative to the distal copy, though the limited divergence limits the power of such inferences (Wertheim et al. 2015) (supplementary Fig. S19).

Evolution of tRNA clusters

Nuclear tRNAs are distributed individually and in clusters that contain identical copies of tRNAs that code for the same amino acids (isoacceptor tRNAs) interspersed with those coding for different amino acids (alloacceptor tRNAs). Previous analyses of tRNAs in the 12 sequenced *Drosophila* genomes found that *D. simulans* had the smallest complement of tRNAs, though this could also be explained by assembly gaps and collapsing of tandem, nearly-identical isoacceptor tRNAs (Velandia-Huerto et al. 2016; Rogers et al. 2010). Additionally, in some genomes, including those of *Drosophila*, select tRNAs are associated with TEs (Mouse Genome Sequencing Consortium et al. 2002; Drosophila 12 Genomes Consortium et al. 2007), which are underrepresented in the previous draft genomes.

We used tRNAscan-SE v1.4 (Lowe and Eddy 1997) to annotate tRNAs in our sim-complex species genomes and the *D. melanogaster* reference genome (supplemental Table S9). Notably, we report the first annotation of genome-wide tRNAs in *D. mauritiana.* We found that genome-wide tRNA counts are very similar between the species, ranging from 295 in *D. melanogaster* to 303 copies in *D. sechellia* (supplemental Fig. S20 and supplemental Table S9). Our count of tRNAs in *D. simulans* (300 tRNAs) is substantially higher than was previously reported using an older assembly (268 and 255 tRNAs; (Rogers et al. 2010; Velandia-Huerto et al. 2016), respectively), suggesting that the high rate of tRNA loss previously reported in this species was likely due to the collapsing of tandem, nearly-identical tRNAs and gaps in the previous *D. simulans* assembly.

To characterize dynamic changes in tRNA clusters across *Drosophila* lineages, we first identified tRNAs that were likely to be orthologous. tRNAs often arise through tandem duplication, complicating the distinction between orthologs and paralogs. We

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

manually curated alignments of tRNAs, employing conservation of the gene order, strand orientation, distance between adjacent tRNAs, anticodon sequence, and intron positions to identify the putative tRNA orthologs between lineages. In doing so, we identified syntenic blocks of tRNAs that differed in copy-number, identity (isotype), anticodons, and pseudogene designations (Fig. 6). We also used a BLAST-based approach, similar to methods used by (Rogers et al. 2010)(see Methods), to identify the flanking regions of orthologous tRNA clusters to confirm gains or losses.

We identified four instances of tRNA anticodon shifts, resulting in three changes in tRNA identity (alloacceptor shifts) and one change that retained the original tRNA identity (isoacceptor shift) (Fig. 6B), confirming previous genome scans using older assemblies (Velandia-Huerto et al. 2016; Rogers and Griffiths-Jones 2014)(Rogers et al. 2010); (Velandia-Huerto et al. 2016; Rogers and Griffiths-Jones 2014). One additional alloacceptor shift (Met CAT > Thr CGT) previously identified in an older assembly of *D. simulans* (Velandia-Huerto et al. 2016; Rogers and Griffiths-Jones 2014), was not found in our analysis, though it is possible that we sequenced an alternate allele for a polymorphic anticodon. In each shift we observed, the derived tRNA sequence was otherwise similar to and retained the predicted structure of the ancestral tRNA. This suggests that these alloacceptor shifts may cause the aminoacyl tRNA synthetase (aaRS) to charge the alloacceptor-shifted tRNA with the amino acid cognate to the ancestral tRNA, integrating the wrong amino acid during translation.

Y chromosome evolution

We initially identified Y-linked contigs based on BLAST hits using *D. melanogaster* Y-linked genes as queries and found that the Y chromosome is scattered across many contigs in all 3 sim-complex assemblies. In each case, the longest Y-linked contig is < 1 Mb and the assemblies of several conserved Y-linked genes known to exist in *D. simulans* are missing exons. Some Y-linked exons (*e.g.,* exons 8–10 of *kl-3* and exons 6–8 of *kl-5*; supplemental Table S9) are present among raw reads but missing in our assemblies of the 3 species (Krsticevic et al. 2015; Chang and

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Larracuente 2019). Although X and Y-linked contigs should have the same coverage from the male reads, we instead find that Y-linked contigs only have ~60% of the coverage of X-linked contigs (supplemental Fig. S2 and supplemental Table S1). Y chromosome assemblies suffer from lower read coverage and a high density of repetitive elements and require the reconciliation of different assembly algorithms, careful curation by hand, and molecular validation—a labor-intensive process beyond the scope of the current study.

Despite these challenges, our assemblies recovered 66, 58 and 64 of 83 *D. melanogaster* Y-linked exons (70–80%; supplemental Table S10) in *D. mauritiana, D. simulans* and *D. sechellia*, respectively, allowing us to study patterns of gene duplication and acquisition. Surprisingly, all known Y-linked genes except *Ppr-Y* exist in multiple copies in at least one of the sim-complex assemblies, and one exon of *Ppr-Y* appears duplicated in the *D. mauritiana* raw long reads. Most duplication events do not involve entire genes but instead correspond to partial tandem duplications (all but *ARY, Pp1-Y1, and Pp1-Y2*), similar to a previously discovered partial duplication of *kl-2* in *D. simulans* (Kopp et al. 2006). We used PCR and Sanger sequencing to validate one duplicated exon from each of 10 Y-linked genes (excepting *Pp1-Y1,* because we could not find mutations that differentiated copies) (supplemental Table S11). Some duplicated exons (*e.g., kl-5* exons 9 and 10) are shared among species of the sim-complex, while other exons different in copy number among species. For example, *ARY* is single-copy in *D. melanogaster* and *D. simulans,* but present in > 3 copies in *D. sechellia* and *D. mauritiana*.

Aside from known Y-linked genes, we also identified 41 duplications from other chromosomes to the Y chromosome in the sim-complex (supplemental Table S12), only 11 of which were identified by a previous study that used Illumina reads (Tobler et al. 2017). Of these 41 duplicated genes, 22 of them are shared in at least 2 of the species, indicating that many duplications are not recent. 27 of these duplicates have copies on both Y-linked contigs and other contigs whose location is unknown but may be Y-linked. We designed primers over unique deletions/sites in putative Y-linked genes and used

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

PCR to confirm their Y linkage. 16 of 17 tested primer pairs amplified male-specific sequences (supplemental Table S11–12). Interestingly, we found that the Y chromosomes of *simulans* complex species share a mitochondrial insertion that are not found in *D. melanogaster* (the duplications are located on unassigned contigs in *D. mauritiana* and *D. sechellia*) . Among these duplicated gene families are the Y-linked pseudo-βCK2tes repeats (*PCKR*) that are related to *Stellate* and *Suppressor of Stellate* (Chang and Larracuente 2019; Danilevskaya et al. 1991; Usakin et al. 2005). We found 92 and 117 copies of *PCKR* on the Y chromosomes of *D. simulans* and *D. mauritiana*, respectively, but only 22 copies in *D. sechellia*.

DISCUSSION

To generate a complete map of genomic variation between closely related species, we constructed highly contiguous (contig N50 = 21.2–22.9Mb) *de novo* genome assemblies of the three sim-complex species. Our assemblies fill gaps in the existing assemblies, capturing the entirety of euchromatin and a significant amount of pericentric and telomeric heterochromatic sequences into single contigs. However, these improvements are accomplished without introducing large-scale misassemblies; indeed we correct many known misassemblies in the existing genomes (Drosophila 12 Genomes Consortium et al. 2007; Schaeffer et al. 2008). We reveal previously hidden structural rearrangements and divergence in highly repetitive pericentromeric regions (Jagannathan et al. 2017; Sproul et al. 2019). Despite the low gene density in heterochromatin, the extensive genomic rearrangements that we discovered in these regions may have broad consequences on phenotypes and species evolution. Many factors linked to genetic incompatibilities between species are located in repetitive sequences inside the pericentromeric heterochromatin (Bayes and Malik 2009; Ferree and Barbash 2009). Our assemblies now allow for comparisons of these regions, an important first step in exploring the mechanisms underlying these incompatibilities.

Our quantification of the dispersed and tandem repeat content and reveals the evolutionary dynamics of repeats between these species. The large blocks of satellite

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DNA in heterochromatin show dynamic evolution in location and abundance between the sim-complex species (Jagannathan et al. 2017; Sproul et al. 2019). Tandem satellite repeats are also found in euchromatin, especially the X chromosome (Kuhn et al. 2012; Gallach 2014). We report a striking enrichment (~15-to-50-fold) of satellite abundance on the X chromosome relative to autosomes, in contrast to previous reports of ~7.5-fold X-chromosome enrichment (Garrigan et al. 2014). The enrichment for satDNA in the X chromosome euchromatin may contribute to the extensive local structural rearrangements (Fig. 2A and Supplementary Table S5–6; (Sproul et al. 2019)). The structural rearrangements that we identify are not limited to the Drosophila genome–we also found structural evolution within the Wolbachia genomes carried by the *D. mauritiana* and *D. sechellia* strains that we sequenced. We infer 15 genome rearrangement events in wMau but only one in wSech, relative to closely related *D. simulans* Wolbachia strains. Compared to the parasitic Wolbachia strains in most *Drosophila* species, wMau is unique in that it increases *D. mauritiana* female fertility. It is possible that wMau's rearranged genome may play some part in this unusual function (Fast et al. 2011; Meany et al. 2019).

The close relatedness of the three sim-complex species, their intercrossability and genome colinearity, and their proximity to the premier model organism D. melanogaster, has made these species models for population, evolutionary, and speciation genetics. Despite their close relatedness, these species clearly have evolved significant ecological, behavioral, and genetic differences that shape their genomes and biology. One particularly salient feature of the sim-complex is the large variation in population size among its members. D. simulans exhibits the largest population in the simulans species complex, with estimates exceeding that of D. melanogaster (Akashi 1995; Akashi 1996). On the other hand, presumably as a result of being an island endemic, D. sechellia is inferred to have the smallest effective population size in the mel-complex. Consistently, D. sechellia shows signatures of a reduced selection efficacy across the genome: lower codon usage bias (Singh et al. 2007), lower levels of polymorphism (Kliman et al. 2000), and increased fixation of slightly deleterious

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mutations (McBride 2007). We observe that tRNAs are most dynamic in D. sechellia, and ¾ anticodon shifts (⅔ alloacceptor shifts) are found in this species' genome but not the other three species genomes. The enrichment of euchromatic satellites in D. sechellia may also be explained by weak selection against satellite DNA expansion as a result of its comparatively smaller population size.

Transposable elements (TEs) are the most abundant class of structural variants in *Drosophila* genomes and an important source of phenotypic variation (Feschotte 2008; Cridland et al. 2013; Chakraborty et al. 2018). For example, we identified TE insertions in the *white* gene in our assemblies that may be responsible for the white-eye phenotype of the *D. simulans* and *D. mauritiana* strains we sequenced (supplemental Fig. S21). However, TEs are poorly represented in current draft-quality genome assemblies (Salzberg and Yorke 2005; Alkan et al. 2011b) and escape detection using short-read sequencing approaches, challenges that are overcome by highly contiguous assemblies (Chakraborty et al. 2018, 2019). While contiguous assemblies have been used to study TEs in *D. melanogaster* (Chakraborty et al. 2018, 2019), the historical paucity of reference-quality genomes from closely related species has imposed limitations to inferring the evolutionary dynamics of TE-turnover on short timescales.

Our contiguous assemblies reveal that repeat content of these genomes is shaped by contemporary evolutionary forces acting on segregating TEs, as well as TE sequences acquired on ancestral lineages. We confirm that *D. melanogaster* harbors more sequences derived from TEs than the sim-complex species, due primarily to the activity of recently acquired LTR elements (Bergman and Bensasson 2007). However, the differences in TE genome content are substantially lower than previous estimates: the euchromatic portion of the *D. melanogaster* genome possesses only ~1.2 fold more TE sequence than *D. simulans*, in contrast to previously reported differences of 3-7 fold (Young and Schwartz 1981; Dowsett and Young 1982; Nuzhdin 1995; Vieira et al. 1999; Vieira and Biémont 2004; Drosophila 12 Genomes Consortium et al. 2007). Approximately 75-80% of the genomic TE content in all four species is due to

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

species-specific TE insertions (Fig. 4), and much of this content is likely polymorphic within each species (Chakraborty et al. 2018).

TE activity is generally deleterious to host organisms (Petrov et al. 2011; Cridland et al. 2013; Chakraborty et al. 2019): transposition disrupts genes and other functional elements (Cooley et al. 1988), TE sequences can act as ectopic regulatory elements (Feschotte 2008), and TEs provide templates for ectopic recombination (Langley et al. 1988). As a consequence, *Drosophila* has evolved host defenses against (e.g. the piRNA and endo-siRNA pathways (Aravin et al. 2007; Chung et al. 2008; Kelleher et al. 2018)((Aravin et al. 2007; Chung et al. 2008; Kelleher et al. 2018)) TE proliferation. These defenses likely shape the genomic TE content in *D. melanogaster* and the sim-complex species. Recent work in *D. simulans* and *D. melanogaster* demonstrates that TE insertions can alter local chromatin state, and thus the expression of adjacent sequences, with potentially deleterious consequences (Lee and Karpen 2017). Heterochromatin proteins and proteins that regulate heterochromatin formation (suppressors of variegation) are expressed at higher levels in *D. simulans*, which has been hypothesized to allow heterochromatin to spread further from TEs into nearby euchromatin in *D. simulans* than *D. melanogaster* (Lee and Karpen 2017). Euchromatic TE insertions in *D. simulans* may therefore be more deleterious and removed more efficiently than those in *D. melanogaster*. The expansion of LTR elements specifically in the *D. melanogaster* lineage points to a weakening of LTR suppression mechanisms, rather than demographic or recombination effects which would likely act on all types of TEs in a similar fashion. This balance between TE activity and host suppression appears to be dynamic across lineages: non-LTR retrotransposons comprise the majority (52%) of the TEs fixed in the mel-complex ancestor, whereas DNA elements comprise most (71%) of the TE sequences fixed in the ancestor of the sim-complex species. For example, INE-1 elements (Yang and Barbash 2008) comprise the most abundant TEs fixed in the sim-complex ancestor, and contribute the largest proportion of TEs assimilated into exons and to the formation of new genes (Yang et al. 2008). Distinct compositions of the TEs fixed in the ancestors of

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mel-complex and sim-complex underscore the lineage-specific mechanisms governing the dynamics of TEs and their functional impacts.

Differences in TE activity may also contribute to intron size differences, as euchromatic introns containing repetitive DNA are ~10% longer in *D. melanogaster* than in the sim-complex genomes. Previous studies of variation inside putatively non-functional dead on arrival TEs have demonstrated species-specific mutation patterns on smaller scales, including the relative size and frequency of insertions and deletions (Petrov et al. 1996; Petrov and Hartl 1998; Blumenstiel et al. 2002). *D. simulans* introns are shorter than *D. melanogaster*'s (Comeron and Kreitman 2000; Ometto et al. 2005) perhaps in part because insertions in introns are favored in *D. melanogaster* (Presgraves 2006). Consistent with these studies, our analysis of mutations in introns find that both *D. simulans* and *D. melanogaster* have more segregating deletions than insertions (1.35 and 1.41 deletions per insertion, respectively), but that in *D. simulans,* the average deletion is larger than the average insertion (4.45 bp vs. 3.31 bp), while segregating insertions are slightly larger than deletions (7.19 bp vs. 6.68 bp) in *D. melanogaster* (supplementary text). These biases are likely the reason that introns without repetitive DNA are ~3 bp longer in the *melanogaster* genome than in *D. simulans*. The observation that simple introns in *D. sechellia* and *D. mauritiana* are closer in size to *D. simulans* than to *D. melanogaster* suggests that all three sim-complex species share similar insertion and deletion biases. Other factors also likely contribute to interspecific differences in intron length. For example, the sim-complex has a higher recombination rate(True et al. 1996a; Brand et al. 2018) which may increase the efficacy of selection to reduce the cost of both transposon transcription and insertion.

Duplicated genes have previously been found to contribute to species and clade-specific adaptations, species differentiation and genetic incompatibilities (Lynch and Force 2000; Ting et al. 2004; Chakraborty and Fry 2015), and to be engaged in intragenomic conflicts (Tao et al. 2007b; Helleu et al. 2016; Eickbush et al. 2019). Even partial duplications of genes can have important functional consequences such as

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

creating novel protein structures (Long et al. 2003; Arguello et al. 2006; Katju and Lynch 2006; Zhou et al. 2008). While comprehensive identification of gene duplications is necessary to evaluate the impact of gene duplication on genome evolution, such duplications are often misassembled or remain unassembled in draft quality genomes (Salzberg and Yorke 2005; Alkan et al. 2011b). Our assemblies of the three sim-complex species provide a unique opportunity to evaluate the contribution of sequence duplications towards structural divergence between species. We estimate that within the mel-complex lineage, the rate of new gene acquisition via duplication is roughly one new gene every 78,000 years (~12.8 duplicates per my). Assuming ~13,000 single-copy genes in the *Drosophila* genome, this translates to a rate of duplicate acquisition of 10^{-9} duplicates / single-copy gene per year, a rate remarkably similar to previous estimates based on different data and over a different timescale (Osada and Innan 2008). When we include partial gene duplications, we find that the sim-complex species ancestor acquired a new duplicate every 40,000 years (~24.8 duplicates per my), or almost 2×10^{-9} duplicates / single-copy gene per year. These estimates suggest that the per gene rate of acquiring new genes are similar to the per nucleotide neutral mutation rate (Keightley et al. 2014). We also observed an enrichment of duplicated genes on the X chromosome compared to the autosomes, suggesting either a higher rate of duplication or fixation of new duplicates on the X chromosome.

We discovered new gene duplicates that were missing or incorrectly assembled in the existing assemblies of the sim-complex species, including mutations in genes that themselves mediate the evolution of the repetitive genome. For example, we discovered a previously hidden duplicate of *su(f)*, which facilitates mRNA 3' end processing and suppresses the expression of *Gypsy* LTR transposons (Parkhurst and Corces 1986; Mazo et al. 1989). The extra *su(f)* copy could potentially contribute to the observation of lower abundance of LTR elements in the sim-complex species compared to *D. melanogaster* (Fig. 4). Another gene duplicated in the sim-complex, *mh*, interacts with a large satellite found in the X-linked heterochromatin called 359-bp–a member of the

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1.688 family of satellites–to maintain the stability of parental genomes during embryogenesis (Tang et al. 2017). The proximal copy maintains ancestral gene structure and expression, but the distal copy of *mh* is shorter than the ancestral copy and has male-biased expression. While we do not know the novel function of the distal *mh* copy, we predict it is functional based on its presence in all three species. We also predict it is involved in maintaining genome integrity given its ancestral function of satellite-mediated chromatin remodeling and the high similarity between the ancestral and derived proteins (Fig. S16). Given that the heterochromatic location of the 359-bp satellite varies between species (Jagathannan et al 2017, Sproul et al 2019), and that the euchromatic distribution of related repeats in the 1.688 family are similarly dynamic (Sproul et al 2019), it is tempting to speculate that the role of the duplicated *mh* may be related to the dynamic evolution of 359-bp and related satellites in the sim-complex.

Sex chromosomes play a special role in the evolution of post-zygotic hybrid incompatibilities. Within the sim-complex, genomic signatures of recent gene flow and factors causing hybrid male sterility are depleted and enriched on the X chromosome, respectively (True et al. 1996b; Tao et al. 2003; Masly and Presgraves 2007; Meiklejohn et al. 2018). Although the *Drosophila* Y chromosome has a low gene density— ~20 genes in 40 Mb—it contributes to hybrid incompatibilities and affects a wide array of phenotypes including longevity, immunity (Araripe et al. 2016; Case et al. 2015; Kutch and Fedorka 2015), meiotic drive (Voelker 1972; Atlan et al. 1997; Unckless et al. 2015), male fitness (Chippindale and Rice 2001) and gene expression across the genome (Lemos et al. 2010; Branco et al. 2013). Gene duplication and gene traffic from the autosomes can shape the evolution of Y-linked gene content (Kopp et al. 2006; Koerich et al. 2008; Carvalho et al. 2015; Ellison and Bachtrog 2019). We found that gene duplication is rampant on sim-complex Y chromosomes. These duplication events involve both conserved Y-linked genes and new Y-linked genes that move from elsewhere in the genome to the Y chromosome. The *D. melanogaster* Y chromosome has a history of strong selective sweeps (Larracuente and Clark 2013), suggesting that some Y-chromosome divergence may contribute to adaptation in the mel-complex

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

species. Previous sequencing technologies restricted comparative genomic research to primarily unique sequences and neglected repetitive genomic regions. Recently reconstructed contiguous genome assemblies using single-molecule sequencing have demonstrated that these previously unassembled or mis-assembled genomic sequences harbor extensive hidden genetic variation relevant to genome evolution and organismal phenotype (Khost et al. 2017; Chakraborty et al. 2018; Stein et al. 2018; Chaisson et al. 2019; Chang and Larracuente 2019; Miga et al. 2019; Stitzer et al. 2019). However, understanding the evolution of these rapidly diverging repetitive, complex genomic regions and their effect on adaptation and species differentiation requires direct comparison between closely related species. Our assemblies revealed that the genomes of these four *Drosophila* species have diverged substantially in the regions that have been historically recalcitrant to assembly. Further studies of these previously hidden differences in the sibling species complex in *Drosophila* and other organisms will help us understand the dynamics of genome evolution underlying speciation and species diversification.

METHODS

Data collection

Genomic DNA was extracted from males following the protocol described in Chakraborty et al. (2016). To shear the DNA, it was subjected to 10 plunges of a 21 gauge needle followed by 10 plunges of a 24 gauge needle. It was then subjected to the 20-80Kb size selection protocol on the Blue Pippin instrument (Chakraborty et al. 2016). The size distribution of genomic DNA was measured with pulse-field gel electrophoresis (Chakraborty et al. 2016). The library for *D. mauritiana* (w12), *D. simulans* (wXD1), and *D. sechellia* (Rob3c / Tucson 14021-0248.25) were generated following the standard 20 kb library protocol. All the sequencing was performed at the UCI genomics core using the P6-C4 chemistry on the RS II platform.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Single-end Illumina data for *D. simulans* was generated by Meiklejohn et al. (Meiklejohn et al. 2018)SRX5065614; (Meiklejohn et al. 2018). *D. mauritiana* paired-end data was obtained from the NCBI short read archive (SRX684364 and SRX135546), and paired-end *D. sechellia* reads were obtained from D. Garrigan (Garrigan et al. 2012)PRJNA541958: SRX5807620 to SRX5807622; (Garrigan et al. 2012)).

To collect RNA sequencing (RNA-Seq) data, *D. simulans* (*w*XD1), *D. mauritiana* (*w*12) and *D. sechellia* (Rob3c / Tucson 14021-0248.25) were reared at room temperature in vials on standard cornmeal-molasses medium. Males and virgin females were collected post-eclosion and allowed to age for 3-5 days in single-sex vials. 20-30 flies were then flash-frozen in liquid nitrogen and stored at -80ºC for RNA extraction. Testes from at least 100 males were dissected in PBS buffer and stored at -80ºC. For *D. simulans* and *D. mauritiana,* Total RNA was extracted following standard protocols using Trizol (Invitrogen), chloroform, and phase-lock gel tubes (Fisher Scientific). Sequencing libraries were generated using the Illumina TruSeq Stranded mRNA kit with 125 bp inserts and paired-end reads were sequenced on an Illumina HiSeq 2500 at the University of Minnesota Genomics Center. For *D. sechellia*, total RNA was isolated from whole tissue using the RNeasy Plus Kit (Qiagen, Valencia, CA) per manufacturer's recommendations. The TruSeq RNA Sample Preparation Kit V2 (Illumina, San Diego, CA) was used for next generation sequencing library construction per manufacturer's protocols. Briefly, polyA mRNA was purified from ~100 ng total RNA with oligo-dT magnetic beads and fragmented. First-strand cDNA synthesis was performed with random hexamer priming followed by second-strand cDNA synthesis. End repair and 3' adenylation was performed on the double-stranded cDNA. Illumina indexed adaptors were ligated to both ends of the cDNA, purified by gel electrophoresis and amplified with PCR primers specific to the adaptor sequences to generate amplicons of approximately 200 to 500 bp in size. The amplified libraries were purified by AMPure (Beckman Coulter, Brea, CA) purification and hybridized to an Illumina pair-end flow cell for cluster amplification using the cBot (Illumina, San Diego, CA) at a concentration of 8 picomoles per lane (data available in PRJNA541548).

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genome assembly

Nuclear genome assembly

We obtained 21.5 Gb long reads for *D. mauritiana* (NR50 = 14.9kb), 20.9 Gb for *D. simulans* (NR50 = 15.7 kb), and 15 Gb for *D. sechellia* (NR50 = 15.2 kb). For each sequence, we generated three assemblies: a hybrid assembly, a long read only assembly with PBcR and a long read only assembly with Canu. The hybrid assembly, or the DBG2OLC assembly, was constructed with the longest 30X long reads (assuming genome size of 130Mb) from each long read dataset with 19.4, 17.68, and 20.65 Gb Illumina reads for *D. mauritiana*, *D. simulans*, and *D. sechellia*, respectively. We generated a long read only assembly for each genome with PBcR (Berlin et al. 2015) as implemented in wgs8.3rc1 with the –sensitive parameter. We also generated a long read only assembly for each genome with Canu 1.2 with genomeSize = 160m useGrid = false errorRate = 0.025 parameters for *D. mauritiana* and *D. sechellia* and with genomeSize = 160m useGrid = false errorRate = 0.035 parameters for *D. simulans* (https://github.com/marbl/canu/tree/5bd4744ad89b71243c7e52446c156956bd75672e; (Koren et al. 2017)). First, we merged the Canu and hybrid assemblies for each genome using quickmerge (Chakraborty et al. 2016; Solares et al. 2018) with the Canu assembly as the reference. We then used the merged assembly from the first step and merged it with the PBcR-sensitive assembly, with the former serving as the reference. In the first round of merging, we used the quickmerge parameters c=1.5, hc0=5.0, l = 1000000, and in the second round of merging we used c=1.5, hco=5.0, l=5000000. We increased the length cutoff for the second round of merging because at least one of the contigs aligning to the major chromosome arms was longer than 5Mb after the first round of merging. We further processed the assemblies with finisherSC (Lam et al. 2015) to use connectivity information from the raw reads that were not used by the assemblers (*e.g.*, DBG2OLC uses the longest 30X of the data and MHAP corrects the longest 40X of the data). We polished all assemblies twice with Quiver (Chin et al. 2013), followed by final

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

polishing with pilon (Walker et al. 2014). We then manually curated 10 misassemblies (supplemental Table S13), including fixing the mitochondrial and Wolbachia genomess described below.

Mitochondrial genome assembly

To assemble the mitochondrial genome for each species, we extracted uncorrected reads aligning to an existing partial mitochondrial genome using *BLASR* (Chaisson and Tesler 2012). (https://github.com/mahulchak/mito-finder). We selected the longest read exceeding a length cutoff of 18Kb (the mitochondrial genome is approximately 19Kb) and trimmed the duplicated sequences resulting from multiple polymerase reading through the sequence start point. We polished the trimmed reads twice with Quiver (Chin et al. 2013) to generate a consensus of all mitochondrial reads.

Wolbachia genome assembly

We extracted the Wolbachia genome in *D. mauritiana* from the original Canu assembly. The Wolbachia genome in *D. sechellia* was not complete in any of our draft assemblies. To assemble the Wolbachia genome in *D. sechellia*, we collected all the reads mapped to two reference Wolbachia genomes (CP003884.1 and CP003883) using BLASR v5.1 (Chaisson and Tesler 2012) with parameters (--clipping soft --bestn 1 --minPctIdentity 0.70). We assembled these reads using Canu v1.3 with the parameters (genomeSize=3m; (Koren et al. 2017)).

Assembly validation

We used the long read coverage to detect assembly errors and validate copy number variants identified by our pipeline. We used *BLASR* (version 1.3.1.142244; parameters: -bestn 1 -sam; (Chaisson and Tesler 2012)) or minimap2 (2-2.8 parameters: -ax map-pb; (Li 2016)) to map the long raw reads to our assemblies and generate sam files that we converted to sorted bam files with *samtools* 1.3 (Li et al. 2009). We calculated long read coverage across the contigs using the *samtools mpileup and depth* (*-Q 10*

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

-*aa*) command. To validate CNVs, we randomly chose 20 CNVs for each species and examined long read coverage across the assembly regions containing the CNVs.

We used Masurca v3.2.1 (Zimin et al. 2013) to detect redundant sequences in our assemblies. In summary, we used MUMmer (Marçais et al. 2018) implemented in Masura to map each contig to the assembly and detect small contigs that are also embedded in longer contigs. We designated candidate redundant contigs as smaller contigs greater than 40 kb with >90% identity, or between 10 and 40 kb with >95% identity, to the longer contigs. These putative redundant contigs may be a result of residual heterozygosity. To detect symbiont-derived sequences from symbionts in our assemblies, we used Blast+ v2.6.0 (Altschul et al. 1990) with blobtools (0.9.19.4; (Laetsch and Blaxter 2017)) to search the nt database (parameters "-task megablast -max_target_seqs 1 -max_hsps 1 -evalue 1e-25"). We calculated the Illumina coverage of each contigs using SRR483621, SRR8247551, and SRR9030358 for *D. mauritiana*, *D. simulans* and *D. sechellia*, respectively. We designated contigs with homology to bacteria and fungi as potential contaminants (supplemental Table S4).

QV estimation

We aligned the Illumina reads from each species to their Pilon polished, pre-scaffolded assemblies using bwa mem with default parameters (Li 2013). Following Koren et al. (Koren et al. 2018), we used freebayes (v0.9.21; (Garrison and Marth 2012) to estimate the number of homozygous (GT = 0/0) SNP and indel errors using the command: "freebayes -C 2 -0 -O -q 20 -z 0.10 -! 3 -E 0 -X -u -p 2 –F 0.75 -b asm.bam -v asm.bayes.vcf -f asm.pilon.fasta". We call the concordance of Illumina reads to our assemblies concordance QV, represented as $-10log_{10}E/T$, where E is the sum of total bases changed (added, deleted, substituted) and T is the total number of bases (minimum coverage of 3).

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Scaffolding

Prior to scaffolding, we masked repeats in all three assemblies with Repeatmasker using default settings. We scaffolded the assemblies of all three genomes with mscaffolder (https://github.com/mahulchak/mscaffolder; (Chakraborty et al. 2018) using the release 6 *D. melanogaster* genome (r6.09) assembly as the reference. The repeat-masked genomes were aligned to the repeat-masked *D. melanogaster* major chromosome arms (X, 2L, 2R, 3L, 3R, 4) using *MUMmer (Marçais et al. 2018)*. We filtered alignments using the delta-filter utility with the -m option and assigned contigs to specific chromosome arms based on the best alignment. We did not assign chromosomal arms to contigs with less than 40% of total alignment to any chromosome arm. To order contigs, we used the starting coordinate of the alignment that did not overlap with the preceding reference chromosome-contig alignment. Finally, we joined the contigs with 100 Ns, to represent assembly gaps of undetermined size. Finally, we prefixed all unscaffolded contigs with 'U'.

Annotation

Transcript annotation

We first mapped transcript and translated sequences from *D. melanogaster* (r6.14) to each assembly using maker2 (v2.31.9; (Holt and Yandell 2011). We further improved the annotations with transcriptome data from whole females, whole females, and testes. We mapped transcriptome (supplemental Table S14) using Hisat 2.1.0 with the maker2 annotation, and then used Stringtie 1.3.4d to generate the consensus annotations from all transcriptomes (Pertea et al. 2016). We further validated potential duplicated genes in *D. simulans* using Iso-seq data from (Nouhaud 2018). We used Blast (-evalue 1e-10; (Altschul et al. 1990) homology to assign the predicted transcripts to *D. melanogaster* transcript sequences. To identify conserved introns, we kept isoforms with the same numbers of exons and only used introns flanked by exons of similar size in each

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

species (< 10% length difference). To compare intron sizes between species, we used the longest isoform from each gene. We manually identified 61 introns from 6 genes with large introns (> 8kb).

Large structural variant detection

To show large scale synteny between the three *sim-complex* genomes and *D. melanogaster,* we created whole-genome alignments with the Mauve aligner (build 2015-2-13) using the progressiveMauve algorithm with the default parameters: default seed weight, determine LCBs (minimum weight = default), full alignment with iterative refinement. We plotted gene density based on Dm6 annotations in *D. melanogaster* was plotted using Karyoploter (Gel and Serra 2017)).

Annotation of repetitive elements

To show the distribution of repetitive elements in our assemblies, we constructed a custom repeat library by combining the latest Repbase release for *Drosophila* with the consensus sequences for complex satellites with repeat units >90 bp. In addition to previously annotated satellites, we identified unannotated complex satellites using Tandem Repeat Finder (TRF) as predicted repeats with period >= 10 and array sizes >= 30kb. We compared these candidates to the latest Repbase release as well as the NR database to remove any redundant or previously annotated repeats. We manually curated previously unannotated sequences and added them to our repeat library. To avoid any bias arising from the use of an existing TE annotation database, we annotated novel TEs in the sim-complex species using the REPET TE annotation package (Flutre et al. 2011). REPET is an umbrella package, which includes both *de novo* (Grouper, Piler, and Recon) and homology (RepeatMasker, Censor) based TE annotation programs. We first fragmented the genome sequence into multiple 200 kb sequences. Then, these genomic fragments were each aligned to themselves using BLAST (Altschul et al. 1990) to identify the repetitive High-scoring Segment Pairs (HSPs). HSPs were clustered using Recon, Grouper and/or Piler (Bao and Eddy 2002;

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Edgar and Myers 2005). After computing multiple alignments of the clustered HSPs, a consensus sequence was obtained from each multiple alignments. These consensus repetitive sequences were treated as *de novo* identified TEs for further cross-comparison with known TE sequences from the RepeatMasker and Repbase TE libraries. Briefly, this cross-comparison involves fragmenting the genome (for parallelization) to obtain a TE map using RepeatMasker and/or Censor (Kohany et al. 2006; Smit et al. 2013). We added these *de novo* TEs from the REPET pipeline into our custom repeat library that includes novel satellites and used this library (Supplementary File S1) to annotate the three *sim-complex* species and the *D. melanogaster* reference with RepeatMasker v4.0.5.

We removed redundant or overlapping annotations and grouped the repeats into larger families using Repbase categorizations, as well as manual curation. If a simple repeat was annotated as spanning >= 100bp, we categorized it as a “simple repeat”, otherwise we categorize it as “low complexity”. We calculated the proportion of each repeat family in 100-kb windows across the scaffolds containing major chromosome arms. We also calculated the total proportion of the genome comprised of each repeat family for euchromatin and heterochromatin. We determined approximate euchromatin/heterochromatin boundaries in the major scaffolds by identifying euchromatin/heterochromatin boundaries from *D. melanogaster* (Hoskins et al 2015) in each simulans clade species assembly using BLAST. We considered Chromosome 4 and all unassigned contigs to be heterochromatin.

tRNA annotation and analysis

We used tRNAscan-SE v1.4 (options: -H; (Lowe and Eddy 1997) to annotate tRNAs in the *D. melanogaster* reference (r6.09) and in our *D. mauritiana*, *D. sechellia*, and *D. simulans* assemblies. We sorted tRNAs within each lineage by position along the chromosomes and represented them as a peptide sequence based on the tRNA isotype predicted by tRNAscan-SE. We first aligned these peptide sequences using MUSCLE v3.8.31 (Edgar 2004) to generate a coarse alignment of tRNA positions for each

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

chromosome (X, 2L, 2R, 3L, 3R). Next, we manually curated each of these alignments by hand, using the conservation of gene order, strand orientation, distances between adjacent tRNAs, anticodon sequence, and intron positions to guarantee the best alignments of tRNAs between lineages.

From our manually curated alignment, we next identified syntenic blocks of neighboring tRNAs—separated by either conserved (*i.e.*, present in all species) tRNAs of a different isotype or by large physical distances along the chromosome. From these syntenic blocks, we identified changes in copy-number, isotype identity, anticodon sequence, or pseudogene designation (as predicted by tRNAscan-SE). We refer to the tRNAs within these syntenic blocks as positional orthologs, though we caution that many of these tRNAs may have arisen through duplications or more complicated local rearrangements and 1:1 orthology between any two tRNA positional orthologs is not implied. We used raw long reads to verify nucleotide changes in predicted tRNA isotypes or anticodons within syntenic blocks, as some of these changes were the result of single-base substitutions or small indels in the tRNA anticodon loop and may be highly sensitive to errors in sequencing or mapping. Visualization of the alignments of raw reads at this position using the Integrative Genome Viewer or IGV (Robinson et al. 2011) revealed that none of the observed changes in tRNA isotypes or anticodons among our assembled genomes were the result of sequencing or mapping error.

We also used a BLAST-based orthology discovery method—similar to methods described in Rogers et al (2010)—to map tRNAs from *D. mauritiana*, *D. sechellia*, or *D. simulans* that did not share positional orthologs with tRNAs in D. melanogaster. Specifically, we asked if sequences flanking these tRNAs had orthologous sequences in *D. melanogaster* and if these sequences overlapped annotated tRNA genes in *D. melanogaster*. We first masked tRNA positions in each query assembly (*D. mauritiana*, *D. sechellia*, *D. simulans*) using the maskfasta function in bedtools v2.20.1 (default options) (Quinlan and Hall 2010). We then masked repetitive sequences in the D. melanogaster reference using RepeatMasker v4.0.5 (Smit et al. 2013) (options: -species drosophila -no_is)—which served as our custom blast database. We extracted

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a 10 kb region of sequence—5 kb from each flank (including neighboring masked tRNAs)—surrounding each tRNA of interest and searched against the repeat-masked *D. melanogaster* database using BLASTN v.2.2.29 (options: -max_hsps 10000 -evalue 10^{-10}). Orthologous windows were identified when both the left- and right-flanking query sequences produced significant search hits separated by fewer than 20Kbp in *D. melanogaster*. Putatively orthologous tRNAs were then identified if these orthologous windows either overlapped or flanked a tRNA annotated in *D. melanogaster*.

Genomewide SV annotation

To discover the genetic variation that changes genome structure (ie structural variants/SVs, eg insertions, duplications, deletions, or inversions SVs), we aligned *D. simulans*, *D. mauritiana*, and *D. sechellia* genomes individually to the *D. melanogaster* reference genome (Hoskins et al. 2015) using MUMmer 4.0 (nucmer -maxmatch) (Marçais et al. 2018) and LASTZ (Harris 2007). We used both of these aligners because they exhibit complementary alignment sensitivity for the detection of copy number variation. The MUMmer alignments were processed used a custom program called *SVMU* v0.3 (Structural Variants from MUMMER; https://github.com/mahulchak/svmu commit 9a20a2d; (Chakraborty et al. 2019) to annotate the SVs as duplicates originating in either the sim-complex or in *D. melanogaster*. We also examined the LASTZ alignments that were not reported by MUMmer and added the duplicates detected by these alignments. The LASTZ alignments were first processed with the CHAIN/NET (Schwartz et al. 2003). The resulting syntenic net output file was further processed to identify the potential structural variations including copy number variants. The detailed workflow is available at https://github.com/yiliao1022/LASTZ_SV_pipeline. We tested the enrichment of the duplicates overlapping full genes on the X chromosome using the exact binomial test (binom.test() function in R). We calculated the proportion of total genes on the euchromatic X chromosome (2244/13861) based on the *D. melanogaster* release 6 GFF file (r6.09). We assumed that the sim-complex duplications overlapping *D. yakuba* duplications by at least a mutually 50% overlap

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(bedtools intersect -f 0.5 -F 0.5) were orthologous. We annotated inversions and insertions (>100 bp) using SVMU, based on whether the sim-complex sequences were inverted or possessed a larger gap between two syntenic segments compared to the *D. melanogaster* reference genome (Hoskins et al. 2015), respectively. To identify species or strain-specific TE insertions, we identified TE insertions in one species that were detected in genome comparisons between that species and the three other species genomes, requiring that the annotated insertions overlap 80% of their length (bedtools intersect -u -f 0.8 -a ins.A.bed -b ins.B.bed). We manually inspected inversion breakpoints from the SVMU output in the nucmer- generated dotplots (nucmer --maxmatch) to validate the breakpoints. We annotated TE overlapping insertions as insertions overlapping 90% of the TE sequence from the repeat library, based on bedtools (bedtools intersect -u -wa -F 0.9 -a asm.svmu.ins.txt -b asm.te.bed).

Shared TE analysis

Due to the high divergence, the unambiguous identification of the syntenic region was difficult in the heterochromatic region. We therefore limited the shared TE analysis to euchromatic regions. To identify TEs shared between species, we aligned the 3 sim-complex species to each other (*D. sechellia-D. mauritiana*, *D. simulans-D. mauritiana*, and *D. simulans-D. sechellia*) using nucmer -maxmatch -g 1000 in MUMmer v4. Similarly, we aligned assemblies of these species individually to the *D. melanogaster* release 6 assembly. We extracted syntenic regions between species pairs from the “cm.txt” output of svmu 0.3 and validated these regions by inspecting the dotplots of the syntenic alignment coordinates (supplemental Fig. S22–23). To identify TE sequences completely contained with syntenic regions between species pairs, we used bedtools (bedtools -u -f 1.0 -a te.bed -b cm.eu.txt). We identified TEs shared among the members of the *D. simulans* species complex and *D. melanogaster* using *D. mauritiana* genome as the reference. TEs shared between *D. mauritiana-D. sechellia* (A) and *D. mauritiana-D. simulans* species pairs (B) were inferred to be derived from either the sim-complex or mel-complex ancestral lineages (Fig. 4), whereas TEs shared

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

between A, B, and *D. mauritiana-D. melanogaster* pair were inferred to be derived from the TEs fixed only in the mel-complex ancestral lineage (bedtools intersect -u -a te.simclade.bed -b te.dmau-dmel.bed). We also repeated this analysis using *D. simulans* genome instead of *D. mauritiana* as the reference, reaching the same conclusion as obtained with *D. mauritiana* as reference.

Y chromosome analyses

We used BLAST to identify the orthologs of all known *D. melanogaster* Y-linked genes in the sim-complex assemblies (Altschul et al. 1990). The sequences of new Y-linked genes were extracted based on Blast results. All alignments of duplicates are manually examined by the eyes.

Cytological validation

We conducted FISH following the protocol from (Larracuente and Ferree 2015). Briefly, brains from third instar larva were dissected and collected in 1X PBS, followed by a 8-min treat of hypotonic solution (0.5% sodium citrate), and fixed in 1.8% paraformaldehyde, 45% acetic acid, and dehydrated in ethanol. The 193XP probe was made by IDT with 5'-/56-FAM/ACATTGGTCAAATGTCAATATGTGGTTATGAATCC-3'. Slides are mounted in Diamond Antifade Mountant with DAPI (Invitrogen) and visualized on a Leica DM5500 upright fluorescence microscope, imaged with a Hamamatsu Orca R2 CCD camera and analyzed using Leica's LAX software.

DATA ACCESS

All raw genomic data and RNAseq have been deposited to NCBI. The accession numbers of the assemblies, Illumina and Pacific Biosciences raw reads are provided in [supplemental Table S15]

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DISCLOSURE DECLARATION

The authors do not declare any conflict of interest.

ACKNOWLEDGMENTS

This work was funded by the National Institutes of Health (K99GM129411 to M.C., R35GM119515 to A.M.L., R01GM123303 to J.J.E.) and National Science Foundation (NSF MCB 1844693 to A.M.L., NSF DDIG 1209536 to J.V., IOS-1656260 to J.J.E., NSF GRFP 1342962 to J.R.A.) and funding from the University of Nebraska-Lincoln to C.D.M and K.L.M. and the University of California, Irvine to J.J.E. A.M.L. is supported by a Stephen Biggar and Elisabeth Asaro fellowship in Data Science. C.-H.C. is supported by the Messersmith Fellowship from the U of Rochester and the Government Scholarship to Study Abroad from Taiwan. We would also like to thank Nishant Nirale and Luna Thanh Ngo for help with data collection and management.

References

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74.

Alkan C, Coe BP, Eichler EE. 2011a. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**: 363.

Alkan C, Sajjadian S, Eichler EE. 2011b. Limitations of next-generation genome sequence assembly. *Nat Methods* **8**: 61–65.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Aminetzach YT, Macpherson JM, Petrov DA. 2005. Pesticide resistance via transposition-mediated adaptive gene truncation in Drosophila. *Science* **309**:

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

764–767.

Araripe LO, Tao Y, Lemos B. 2016. Interspecific Y chromosome variation is sufficient to rescue hybrid male sterility and is influenced by the grandparental origin of the chromosomes. *Heredity* **116**: 516–522.

Aravin AA, Hannon GJ, Brennecke J. 2007. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318**: 761–764.

Arguello JR, Chen Y, Yang S, Wang W, Long M. 2006. Origination of an X-linked testes chimeric gene by illegitimate recombination in Drosophila. *PLoS Genet* **2**: e77.

Atlan A, Merçot H, Landre C, Montchamp-Moreau C. 1997. THE SEX-RATIO TRAIT IN DROSOPHILA SIMULANS: GEOGRAPHICAL DISTRIBUTION OF DISTORTION AND RESISTANCE. *Evolution* **51**: 1886–1895.

Bachtrog D. 2004. Evidence that positive selection drives Y-chromosome degeneration in Drosophila miranda. *Nat Genet* **36**: 518–522.

Bao Z, Eddy SR. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res* **12**: 1269–1276.

Barbash DA. 2010. Ninety years of Drosophila melanogaster hybrids. *Genetics* **186**: 1–8.

Bargues N, Lerat E. 2017. Evolutionary history of LTR-retrotransposons among 20 Drosophila species. *Mob DNA* **8**: 7.

Battlay P, Leblanc PB, Green L, Garud NR, Schmidt JM, Fournier-Level A, Robin C. 2018. Structural Variants and Selective Sweep Foci Contribute to Insecticide Resistance in the Drosophila Genetic Reference Panel. *G3* **8**: 3489–3497.

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh Y-P, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN, et al. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS Biol* **5**: e310.

Bergman CM, Bensasson D. 2007. Recent LTR retrotransposon insertion contrasts with waves of non-LTR insertion since speciation in Drosophila melanogaster. *Proc Natl Acad Sci U S A* **104**: 11340–11345.

Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM. 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol* **33**: 623–630.

Blumenstiel JP. 2019. Birth, School, Work, Death, and Resurrection: The Life Stages and Dynamics of Transposable Element Proliferation. *Genes* **10**.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

http://dx.doi.org/10.3390/genes10050336.

Blumenstiel JP, Hartl DL, Lozovsky ER. 2002. Patterns of insertion and deletion in contrasting chromatin domains. *Mol Biol Evol* **19**: 2211–2225.

Branco AT, Tao Y, Hartl DL, Lemos B. 2013. Natural variation of the Y chromosome suppresses sex ratio distortion and modulates testis-specific gene expression in Drosophila simulans. *Heredity* **111**: 8–15.

Brand CL, Cattani MV, Kingan SB, Landeen EL, Presgraves DC. 2018. Molecular Evolution at a Meiosis Gene Mediates Species Differences in the Rate and Patterning of Recombination. *Curr Biol* **28**: 1289–1295.e4.

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* **128**: 1089–1103.

Cabanettes F, Klopp C. 2018. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**: e4958.

Cabot EL, Doshi P, Wu ML, Wu CI. 1993. Population genetics of tandem repeats in centromeric heterochromatin: unequal crossing over and chromosomal divergence at the Responder locus of Drosophila melanogaster. *Genetics* **135**: 477–487.

Carvalho AB, Vicoso B, Russo CAM, Swenor B, Clark AG. 2015. Birth of a new gene on the Y chromosome of Drosophila melanogaster. *Proc Natl Acad Sci U S A* **112**: 12450–12455.

Case LK, Wall EH, Osmanski EE, Dragon JA, Saligrama N, Zachary JF, Lemos B, Blankenhorn EP, Teuscher C. 2015. Copy number variation in Y chromosome multicopy genes is linked to a paternal parent-of-origin effect on CNS autoimmune disease in female offspring. *Genome Biol* **16**: 28.

Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, et al. 2015. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**: 608–611.

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10**: 1784.

Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**: 238.

Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. 2016. Contiguous and

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res* **44**: e147.

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

Chakraborty M, Fry JD. 2015. Parallel functional changes in independent testis-specific duplicates of Aldehyde dehydrogenase in Drosophila. *Mol Biol Evol* **32**: 1029–1038.

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

Chang C-H, Chavan A, Palladino J, Wei X, Martins NMC, Santinello B, Chen C-C, Erceg J, Beliveau BJ, Wu C-T, et al. 2019. Islands of retroelements are major components of Drosophila centromeres. *PLoS Biol* **17**: e3000241.

Chang C-H, Larracuente AM. 2019. Heterochromatin-Enriched Assemblies Reveal the Sequence and Organization of the Drosophila melanogaster Y Chromosome. *Genetics* **211**: 333–348.

Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**: 563–569.

Chippindale AK, Rice WR. 2001. Y chromosome polymorphism is a strong determinant of male fitness in Drosophila melanogaster. *Proc Natl Acad Sci U S A* **98**: 5677–5682.

Chung W-J, Okamura K, Martin R, Lai EC. 2008. Endogenous RNA interference provides a somatic defense against Drosophila transposons. *Curr Biol* **18**: 795–802.

Cocquet J, Ellis PJI, Mahadevaiah SK, Affara NA, Vaiman D, Burgoyne PS. 2012. A genetic basis for a postmeiotic X versus Y chromosome intragenomic conflict in the mouse. *PLoS Genet* **8**: e1002900.

Comeron JM, Kreitman M. 2000. The correlation between intron length and recombination in drosophila. Dynamic equilibrium between mutational and selective forces. *Genetics* **156**: 1175–1190.

Cooley L, Kelley R, Spradling A. 1988. Insertional mutagenesis of the Drosophila genome with single P elements. *Science* **239**: 1121–1128.

Cosby RL, Chang N-C, Feschotte C. 2019. Host-transposon interactions: conflict,

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cooperation, and cooption. *Genes Dev* **33**: 1098–1116.

Cridland JM, Macdonald SJ, Long AD, Thornton KR. 2013. Abundance and Distribution of Transposable Elements in Two Drosophila QTL Mapping Resources. *Mol Biol Evol* **30**: 2311–2327.

Daborn PJ, Yen JL, Bogwitz MR, Le Goff G, Feil E, Jeffers S, Tijet N, Perry T, Heckel D, Batterham P, et al. 2002. A single p450 allele associated with insecticide resistance in Drosophila. *Science* **297**: 2253–2256.

Danilevskaya ON, Kurenova EV, Pavlova MN, Bebehov DV, Link AJ, Koga A, Vellek A, Hartl DL. 1991. He-T family DNA sequences in the Y chromosome of Drosophila melanogaster share homology with the X-linked stellate genes. *Chromosoma* **100**: 118–124.

Delabaere L, Orsi GA, Sapey-Triomphe L, Horard B, Couble P, Loppin B. 2014. The Spartan ortholog maternal haploid is required for paternal chromosome integrity in the Drosophila zygote. *Curr Biol* **24**: 2281–2287.

DiBartolomeis SM, Tartof KD, Jackson FR. 1992. A superfamily of Drosophila satellite related (SR) DNA repeats restricted to the X chromosome euchromatin. *Nucleic Acids Res* **20**: 1113–1116.

Ding Y, Lillvis JL, Cande J, Berman GJ, Arthur BJ, Long X, Xu M, Dickson BJ, Stern DL. 2019. Neural Evolution of Context-Dependent Fly Song. *Curr Biol* **29**: 1089–1099.e7.

Ding Y, Zhao L, Yang S, Jiang Y, Chen Y, Zhao R, Zhang Y, Zhang G, Dong Y, Yu H, et al. 2010. A young Drosophila duplicate gene plays essential roles in spermatogenesis by regulating several Y-linked male fertility genes. *PLoS Genet* **6**: e1001255.

Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**: 601–603.

dos Santos G, Schroeder AJ, Goodman JL, Strelets VB, Crosby MA, Thurmond J, Emmert DB, Gelbart WM, FlyBase Consortium. 2015. FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res* **43**: D690–7.

Dowsett AP, Young MW. 1982. Differing levels of dispersed repetitive DNA among closely related species of Drosophila. *Proc Natl Acad Sci U S A* **79**: 4570–4574.

Drosophila 12 Genomes Consortium, Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**: 203–218.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797.

Edgar RC, Myers EW. 2005. PILER: identification and classification of genomic repeats. *Bioinformatics* **21 Suppl 1**: i152–8.

Eickbush MT, Young JM, Zanders SE. 2019. Killer Meiotic Drive and Dynamic Evolution of the wtf Gene Family. *Mol Biol Evol* **36**: 1201–1214.

Ellison C, Bachtrog D. 2019. Recurrent gene co-amplification on Drosophila X and Y chromosomes. *PLoS Genet* **15**: e1008251.

Ellis PJI, Bacon J, Affara NA. 2011. Association of Sly with sex-linked gene amplification during mouse evolution: a side effect of genomic conflict in spermatids? *Hum Mol Genet* **20**: 3010–3021.

Fast EM, Toomey ME, Panaram K, Desjardins D, Kolaczyk ED, Frydman HM. 2011. Wolbachia enhance Drosophila stem cell proliferation and target the germline stem cell niche. *Science* **334**: 990–992.

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**: 397–405.

Fishman L, Saunders A. 2008. Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers. *Science* **322**: 1559–1562.

Flutre T, Duprat E, Feuillet C, Quesneville H. 2011. Considering transposable element diversification in de novo annotation approaches. *PLoS One* **6**: e16526.

Gallach M. 2014. Recurrent turnover of chromosome-specific satellites in Drosophila. *Genome Biol Evol* **6**: 1279–1286.

Garrigan D, Kingan SB, Geneva AJ, Andolfatto P, Clark AG, Thornton KR, Presgraves DC. 2012. Genome sequencing reveals complex speciation in the Drosophila simulans clade. *Genome Res* **22**: 1499–1511.

Garrigan D, Kingan SB, Geneva AJ, Vedanayagam JP, Presgraves DC. 2014. Genome diversity and divergence in Drosophila mauritiana: multiple signatures of faster X evolution. *Genome Biol Evol* **6**: 2444–2458.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv [q-bioGN]*. http://arxiv.org/abs/1207.3907.

Gel B, Serra E. 2017. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**: 3088–3090.

Hartley G, O'Neill RJ. 2019. Centromere Repeats: Hidden Gems of the Genome. *Genes*

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

10. http://dx.doi.org/10.3390/genes10030223.

Helleu Q, Gérard PR, Dubruille R, Ogereau D, Prud'homme B, Loppin B, Montchamp-Moreau C. 2016. Rapid evolution of a Y-chromosome heterochromatin protein underlies sex chromosome meiotic drive. *Proc Natl Acad Sci U S A* **113**: 4110–4115.

Henikoff S, Ahmad K, Malik HS. 2001. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**: 1098–1102.

Hey J, Kliman RM. 1993. Population genetics and phylogenetics of DNA sequence variation at multiple loci within the Drosophila melanogaster species complex. *Mol Biol Evol* **10**: 804–822.

Hillis DM, Dixon MT. 1991. Ribosomal DNA: molecular evolution and phylogenetic inference. *Q Rev Biol* **66**: 411–453.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**: 491.

Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, Booth BW, Pfeiffer BD, George RA, Svirskas R, et al. 2015. The Release 6 reference sequence of the Drosophila melanogaster genome. *Genome Res* **25**: 445–458.

Hoskins RA, Smith CD, Carlson JW, Carvalho AB, Halpern A, Kaminker JS, Kennedy C, Mungall CJ, Sullivan BA, Sutton GG, et al. 2002. Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. *Genome Biol* **3**: RESEARCH0085.

Huang W, Massouras A, Inoue Y, Peiffer J, Ramia M, Tarone AM, Turlapati L, Zichner T, Zhu D, Lyman RF, et al. 2014. Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. *Genome Res* **24**: 1193–1208.

Jaenike J. 2001. Sex Chromosome Meiotic Drive. *Annu Rev Ecol Syst* **32**: 25–49.

Jagannathan M, Warsinger-Pepe N, Watase GJ, Yamashita YM. 2017. Comparative Analysis of Satellite DNA in the Drosophila melanogaster Species Complex. *G3* **7**: 693–704.

Juneja P, Lazzaro BP. 2009. Providencia sneebia sp. nov. and Providencia burhodogranariea sp. nov., isolated from wild Drosophila melanogaster. *Int J Syst Evol Microbiol* **59**: 1108–1111.

Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, Frise E, Wheeler DA, Lewis SE, Rubin GM, et al. 2002. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome Biol* **3**:

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RESEARCH0084.

Katju V, Lynch M. 2006. On the formation of novel genes by duplication in the Caenorhabditis elegans genome. *Mol Biol Evol* **23**: 1056–1067.

Keightley PD, Ness RW, Halligan DL, Haddrill PR. 2014. Estimation of the spontaneous mutation rate per nucleotide site in a Drosophila melanogaster full-sib family. *Genetics* **196**: 313–320.

Kelleher ES, Jaweria J, Akoma U, Ortega L, Tang W. 2018. QTL mapping of natural variation reveals that the developmental regulator bruno reduces tolerance to P-element transposition in the Drosophila female germline. *PLoS Biol* **16**: e2006040.

Khost DE, Eickbush DG, Larracuente AM. 2017. Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in Drosophila melanogaster. *Genome Res*. http://genome.cshlp.org/content/early/2017/04/03/gr.213512.116.abstract.

Klein SJ, O'Neill RJ. 2018. Transposable elements: genome innovation, chromosome diversity, and centromere conflict. *Chromosome Res* **26**: 5–23.

Kliman RM, Andolfatto P, Coyne JA, Depaulis F, Kreitman M, Berry AJ, McCarter J, Wakeley J, Hey J. 2000. The population genetics of the origin and divergence of the Drosophila simulans complex species. *Genetics* **156**: 1913–1931.

Koerich LB, Wang X, Clark AG, Carvalho AB. 2008. Low conservation of gene content in the Drosophila Y chromosome. *Nature* **456**: 949–951.

Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* **7**: 474.

Kopp A, Frank AK, Barmina O. 2006. Interspecific divergence, intrachromosomal recombination, and phylogenetic utility of Y-chromosomal genes in Drosophila. *Mol Phylogenet Evol* **38**: 731–741.

Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy AM. 2018. De novo assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol*. http://dx.doi.org/10.1038/nbt.4277.

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Krsticevic FJ, Schrago CG, Carvalho AB. 2015. Long-Read Single Molecule Sequencing to Resolve Tandem Gene Copies: The Mst77Y Region on the Drosophila melanogaster Y Chromosome. *G3* **5**: 1145–1150.

Kruger AN, Brogley MA, Huizinga JL, Kidd JM, de Rooij DG, Hu Y-C, Mueller JL. 2019. A Neofunctionalized X-Linked Ampliconic Gene Family Is Essential for Male Fertility and Equal Sex Ratio in Mice. *Curr Biol* **29**: 3699–3706.e5.

Kuhn GC, Kuttler H, Moreira-Filho O, Heslop-Harrison JS. 2012. The 1.688 repetitive DNA of Drosophila: concerted evolution at different genomic scales and association with genes. *Mol Biol Evol* **29**: 7–11.

Kutch IC, Fedorka KM. 2015. Y-linked variation for autosomal immune gene regulation has the potential to shape sexually dimorphic immunity. *Proc Biol Sci* **282**: 20151301.

Lachaise D, David JR, Lemeunier F, Tsacas L, Ashburner M. 1986. THE REPRODUCTIVE RELATIONSHIPS OF DROSOPHILA SECHELLIA WITH D. MAURITIANA, D. SIMULANS, AND D. MELANOGASTER FROM THE AFROTROPICAL REGION. *Evolution* **40**: 262–271.

Laetsch DR, Blaxter ML. 2017. BlobTools: Interrogation of genome assemblies. *F1000Res* **6**: 1287.

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

Larracuente AM. 2014. The organization and evolution of the Responder satellite in species of the Drosophila melanogaster group: dynamic evolution of a target of meiotic drive. *BMC Evol Biol* **14**: 233.

Larracuente AM, Clark AG. 2013. Surprising differences in the variability of Y chromosomes in African and cosmopolitan populations of Drosophila melanogaster. *Genetics* **193**: 201–214.

Larracuente AM, Ferree PM. 2015. Simple method for fluorescence DNA in situ hybridization to squashed chromosomes. *J Vis Exp* **95**: 52288.

Larracuente AM, Presgraves DC. 2012. The selfish Segregation Distorter gene complex of Drosophila melanogaster. *Genetics* **192**: 33–53.

Lee YCG, Karpen GH. 2017. Pervasive epigenetic effects of Drosophila euchromatic transposable elements impact their evolution. *Elife* **6**. http://dx.doi.org/10.7554/eLife.25762.

Lemos B, Branco AT, Hartl DL. 2010. Epigenetic effects of polymorphic Y chromosomes modulate chromatin components, immune response, and sexual conflict. *Proc Natl*

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acad Sci U S A **107**: 15826–15831.

Lerat E, Burlet N, Biémont C, Vieira C. 2011. Comparative analysis of transposable elements in the melanogaster subgroup sequenced genomes. *Gene* **473**: 100–109.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bioGN]*. http://arxiv.org/abs/1303.3997.

Li H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**: 2103–2110.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Lindholm AK, Dyer KA, Firman RC, Fishman L, Forstmeier W, Holman L, Johannesson H, Knief U, Kokko H, Larracuente AM, et al. 2016. The Ecology and Evolutionary Dynamics of Meiotic Drive. *Trends Ecol Evol* **31**: 315–326.

Lin Y, Moret BME. 2008. Estimating true evolutionary distances under the DCJ model. *Bioinformatics* **24**: i114–22.

Lohe AR, Roberts PA. 1990. An unusual Y chromosome of Drosophila simulans carrying amplified rDNA spacer without rRNA genes. *Genetics* **125**: 399–406.

Long M, Betrán E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. *Nat Rev Genet* **4**: 865–875.

Loppin B, Berger F, Couble P. 2001. Paternal chromosome incorporation into the zygote nucleus is controlled by maternal haploid in Drosophila. *Dev Biol* **231**: 383–396.

Lower SS, McGurk MP, Clark AG, Barbash DA. 2018. Satellite DNA evolution: old ideas, new approaches. *Curr Opin Genet Dev* **49**: 70–78.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**: 955–964.

Lynch M, Force AG. 2000. The Origin of Interspecific Genomic Incompatibility via Gene Duplication. *Am Nat* **156**: 590–605.

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol* **14**: e1005944.

Masly JP, Presgraves DC. 2007. High-resolution genome-wide dissection of the two

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

rules of speciation in Drosophila. *PLoS Biol* **5**: e243.

Mason JM, Frydrychova RC, Biessmann H. 2008. Drosophila telomeres: an exception providing new insights. *Bioessays* **30**: 25–37.

Matute DR, Ayroles JF. 2014. Hybridization occurs between Drosophila simulans and D. sechellia in the Seychelles archipelago. *J Evol Biol* **27**: 1057–1068.

Mazo AM, Mizrokhi LJ, Karavanov AA, Sedkov YA, Krichevskaja AA, Ilyin YV. 1989. Suppression in Drosophila: su(Hw) and su(f) gene products interact with a region of gypsy (mdg4) regulating its transcriptional activity. *EMBO J* **8**: 903–911.

McDermott SR, Kliman RM. 2008. Estimation of isolation times of the island species in the Drosophila simulans complex from multilocus DNA sequence data. *PLoS One* **3**: e2442.

Meany MK, Conner WR, Richter SV, Bailey JA, Turelli M, Cooper BS. 2019. Loss of cytoplasmic incompatibility and minimal fecundity effects explain relatively low Wolbachia frequencies in Drosophila mauritiana. *Evolution* **73**: 1278–1295.

Meiklejohn CD, Landeen EL, Gordon KE, Rzatkiewicz T, Kingan SB, Geneva AJ, Vedanayagam JP, Muirhead CA, Garrigan D, Stern DL, et al. 2018. Gene flow mediates the role of sex chromosome meiotic drive during complex speciation. *Elife* **7**. http://dx.doi.org/10.7554/eLife.35468.

Menon DU, Coarfa C, Xiao W, Gunaratne PH, Meller VH. 2014. siRNAs from an X-linked satellite repeat promote X-chromosome recognition in Drosophila melanogaster. *Proc Natl Acad Sci U S A* **111**: 16460–16465.

Menon DU, Meller VH. 2012. A role for siRNA in X-chromosome dosage compensation in Drosophila melanogaster. *Genetics* **191**: 1023–1028.

Merçot H, Defaye D, Capy P, Pla E, David JR. 1994. ALCOHOL TOLERANCE, ADH ACTIVITY, AND ECOLOGICAL NICHE OF DROSOPHILA SPECIES. *Evolution* **48**: 746–757.

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, et al. 2019. Telomere-to-telomere assembly of a complete human X chromosome. *bioRxiv* 735928. https://www.biorxiv.org/content/10.1101/735928v3.abstract (Accessed December 20, 2019).

Miller DE, Staber C, Zeitlinger J, Hawley RS. 2018. GENOME REPORT: Highly Contiguous Genome Assemblies of 15 Drosophila Species Generated Using Nanopore Sequencing. *G3* . http://dx.doi.org/10.1534/g3.118.200160.

Montchamp-Moreau C, Ogereau D, Chaminade N, Colard A, Aulard S. 2006.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Organization of the sex-ratio Meiotic Drive Region in Drosophila simulans. *Genetics* **174**: 1365–1371. http://dx.doi.org/10.1534/genetics.105.051755.

Moriyama EN, Powell JR. 1997. Codon usage bias and tRNA abundance in Drosophila. *J Mol Evol* **45**: 514–523.

Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520–562.

Moyzis RK, Buckingham JM, Cram LS, Dani M, Deaven LL, Jones MD, Meyne J, Ratliff RL, Wu J-R. 1988. A highly conserved repetitive DNA sequence,(TTAGGG) n, present at the telomeres of human chromosomes. *Proceedings of the National Academy of Sciences* **85**: 6622–6626.

Nouhaud P. 2018. Long-read based assembly and annotation of a Drosophila simulans genome. *bioRxiv* 425710. https://www.biorxiv.org/content/early/2018/09/24/425710 (Accessed April 1, 2019).

Nuzhdin SV. 1995. The distribution of transposable elements on X chromosomes from a natural population of Drosophila simulans. *Genet Res* **66**: 159–166.

Ometto L, Stephan W, De Lorenzo D. 2005. Insertion/deletion and nucleotide polymorphism data reveal constraints in Drosophila melanogaster introns and intergenic regions. *Genetics* **169**: 1521–1527.

Orgel LE, Crick FH. 1980. Selfish DNA: the ultimate parasite. *Nature* **284**: 604–607.

Osada N, Innan H. 2008. Duplication and gene conversion in the Drosophila melanogaster genome. *PLoS Genet* **4**: e1000305.

Parhad SS, Theurkauf WE. 2019. Rapid evolution and conserved function of the piRNA pathway. *Open Biol* **9**: 180181.

Parkhurst SM, Corces VG. 1986. Mutations at the suppressor of forked locus increase the accumulation of gypsy-encoded transcripts in Drosophila melanogaster. *Mol Cell Biol* **6**: 2271–2274.

Pease JB, Hahn MW. 2013. More accurate phylogenies inferred from low-recombination regions in the presence of incomplete lineage sorting. *Evolution* **67**: 2376–2384.

Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* **11**: 1650–1667.

Petrov DA, Fiston-Lavier A-S, Lipatov M, Lenkov K, González J. 2011. Population

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genomics of transposable elements in Drosophila melanogaster. *Mol Biol Evol* **28**: 1633–1644.

Petrov DA, Hartl DL. 1998. High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups. *Mol Biol Evol* **15**: 293–302.

Petrov DA, Lozovskaya ER, Hartl DL. 1996. High intrinsic rate of DNA loss in Drosophila. *Nature* **384**: 346–349.

Plohl M, Meštrović N, Mravinac B. 2012. Satellite DNA evolution. *Genome Dyn* **7**: 126–152.

Podemski L, Ferrer C, Locke J. 2001. Whole arm inversions of chromosome 4 in Drosophila species. *Chromosoma* **110**: 305–312.

Presgraves DC. 2006. Intron length evolution in Drosophila. *Mol Biol Evol* **23**: 2203–2213.

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

Rathje CC, Johnson EEP, Drage D, Patinioti C, Silvestri G, Affara NA, Ialy-Radio C, Cocquet J, Skinner BM, Ellis PJI. 2019. Differential Sperm Motility Mediates the Sex Ratio Drive Shaping Mouse Sex Chromosome Evolution. *Curr Biol* **29**: 3692–3698.e4.

R'Kha S, Capy P, David JR. 1991. Host-plant specialization in the Drosophila melanogaster species complex: a physiological, behavioral, and genetical analysis. *Proc Natl Acad Sci U S A* **88**: 1835–1839.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24–26.

Rogers HH, Bergman CM, Griffiths-Jones S. 2010. The evolution of tRNA genes in Drosophila. *Genome Biol Evol* **2**: 467–477.

Rogers HH, Griffiths-Jones S. 2014. tRNA anticodon shifts in eukaryotic genomes. *RNA* **20**: 269–281.

Salzberg SL, Yorke JA. 2005. Beware of mis-assembled genomes. *Bioinformatics* **21**: 4320–4321.

Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O'Grady PM, Rohde C, Valente VLS, Aguadé M, Anderson WW, et al. 2008. Polytene chromosomal maps of 11 Drosophila species: the order of genomic scaffolds inferred from genetic and physical maps. *Genetics* **179**: 1601–1655.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Schmidt JM, Good RT, Appleton B, Sherrard J, Raymant GC, Bogwitz MR, Martin J, Daborn PJ, Goddard ME, Batterham P, et al. 2010. Copy number variation and transposable elements feature in recent, ongoing adaptation at the Cyp6g1 locus. *PLoS Genet* **6**: e1000998.

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

Signor SA, New, F. N., Nuzhdin S. 2018. A Large Panel of Drosophila simulans Reveals an Abundance of Common Variants. *Genome Biol Evol* **10**: 189–206.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210–3212.

Smit AFA, Hubley R, Green P. 2013. RepeatMasker. *Open-40*. http://www.repeatmasker.org.

Solares EA, Chakraborty M, Miller DE, Kalsow S, Hall K, Perera AG, Emerson JJ, Hawley RS. 2018. Rapid Low-Cost Assembly of the Drosophila melanogaster Reference Genome Using Low-Coverage, Long-Read Sequencing. *G3* . http://dx.doi.org/10.1534/g3.118.200162.

Sproul JS, Khost DE, Eickbush DG, Negm S, Wei X, Wong I, Larracuente AM. 2019. Dynamic evolution of euchromatic satellites on the X chromosome in Drosophila melanogaster and the simulans clade. *bioRxiv* 846238. https://www.biorxiv.org/content/10.1101/846238v1 (Accessed November 25, 2019).

Stage DE, Eickbush TH. 2007. Sequence variation within the rRNA gene loci of 12 Drosophila species. *Genome Res* **17**: 1888–1897.

Steinberg KM, Schneider VA, Graves-Lindsay TA, Fulton RS, Agarwala R, Huddleston J, Shiryev SA, Morgulis A, Surti U, Warren WC, et al. 2014. Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Res* **24**: 2066–2076.

Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang C, Chougule K, Gao D, Iwata A, Goicoechea JL, et al. 2018. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. *Nat Genet* **50**: 285–296.

Stitzer MC, Anderson SN, Springer NM, Ross-Ibarra J. 2019. The Genomic Ecosystem of Transposable Elements in Maize. *bioRxiv* 559922. https://www.biorxiv.org/content/10.1101/559922v1 (Accessed December 8, 2019).

Sturtevant AH, Plunkett CR. 1926. SEQUENCE OF CORRESPONDING THIRD-CHROMOSOME GENES IN DROSOPHILA MELANOGASTER AND D.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SIMULANS. *Biol Bull* **50**: 56–60.

Talbert P, Kasinathan S, Henikoff S. 2018. Simple and Complex Centromeric Satellites in Drosophila Sibling Species. *Genetics* **2018/01/07**. https://www.ncbi.nlm.nih.gov/pubmed/29305387.

Tang X, Cao J, Zhang L, Huang Y, Zhang Q, Rong YS. 2017. Maternal Haploid, a Metalloprotease Enriched at the Largest Satellite Repeat and Essential for Genome Integrity in Drosophila Embryos. *Genetics* **206**: 1829–1839.

Tao Y, Araripe L, Kingan SB, Ke Y, Xiao H, Hartl DL. 2007a. A sex-ratio meiotic drive system in Drosophila simulans. II: an X-linked distorter. *PLoS Biol* **5**: e293.

Tao Y, Chen S, Hartl DL, Laurie CC. 2003. Genetic dissection of hybrid incompatibilities between Drosophila simulans and D. mauritiana. I. Differential accumulation of hybrid male sterility effects on the X and autosomes. *Genetics* **164**: 1383–1397.

Tao Y, Hartl DL, Laurie CC. 2001. Sex-ratio segregation distortion associated with reproductive isolation in Drosophila. *Proc Natl Acad Sci U S A* **98**: 13183–13188.

Tao Y, Masly JP, Araripe L, Ke Y, Hartl DL. 2007b. A sex-ratio meiotic drive system in Drosophila simulans. I: an autosomal suppressor. *PLoS Biol* **5**: e292.

Ting C-T, Tsaur S-C, Sun S, Browne WE, Chen Y-C, Patel NH, Wu C-I. 2004. Gene duplication and speciation in Drosophila: evidence from the Odysseus locus. *Proc Natl Acad Sci U S A* **101**: 12232–12235.

Tobler R, Nolte V, Schlotterer C. 2017. High rate of translocation-based gene birth on the Drosophila Y chromosome. *Proc Natl Acad Sci U S A* **114**: 11721–11726.

Treangen TJ, Salzberg SL. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13**: 36–46.

True JR, Mercer JM, Laurie CC. 1996a. Differences in crossover frequency and distribution among three sibling species of Drosophila. *Genetics* **142**: 507–523.

True JR, Weir BS, Laurie CC. 1996b. A genome-wide survey of hybrid incompatibility factors by the introgression of marked segments of Drosophila mauritiana chromosomes into Drosophila simulans. *Genetics* **142**: 819–837.

Unckless RL, Larracuente AM, Clark AG. 2015. Sex-ratio meiotic drive and Y-linked resistance in Drosophila affinis. *Genetics* **199**: 831–840.

Usakin LA, Kogan GL, Kalmykova AI, Gvozdev VA. 2005. An alien promoter capture as a primary step of the evolution of testes-expressed repeats in the Drosophila melanogaster genome. *Mol Biol Evol* **22**: 1555–1560.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Van't Hof AE, Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, Hall N, Darby AC, Saccheri IJ. 2016. The industrial melanism mutation in British peppered moths is a transposable element. *Nature* **534**: 102–105.

Velandia-Huerto CA, Berkemer SJ, Hoffmann A, Retzlaff N, Romero Marroquín LC, Hernández-Rosales M, Stadler PF, Bermúdez-Santana CI. 2016. Orthologs, turn-over, and remolding of tRNAs in primates and fruit flies. *BMC Genomics* **17**: 617.

Vieira C, Biémont C. 2004. Transposable element dynamics in two sibling species: Drosophila melanogaster and Drosophila simulans. *Genetica* **120**: 115–123.

Vieira C, Lepetit D, Dumont S, Biémont C. 1999. Wake up of transposable elements following Drosophila simulans worldwide colonization. *Mol Biol Evol* **16**: 1251–1255.

Voelker RA. 1972. Preliminary characterization of "sex ratio" and rediscovery and reinterpretation of "male sex ratio" in Drosophila affinis. *Genetics* **71**: 597–606.

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

Waring GL, Pollack JC. 1987. Cloning and characterization of a dispersed, multicopy, X chromosome sequence in Drosophila melanogaster. *Proc Natl Acad Sci U S A* **84**: 2843–2847.

Wei KHC, Lower SE, Caldas IV, Sless TJ, Barbash DA, Clark AG. 2018. Variable rates of simple satellite gains across the Drosophila phylogeny. *Mol Biol Evol* msy005–msy005.

Werren JH, Nur U, Wu CI. 1988. Selfish genetic elements. *Trends Ecol Evol* **3**: 297–302.

Wertheim JO, Murrell B, Smith MD, Kosakovsky Pond SL, Scheffler K. 2015. RELAX: detecting relaxed selection in a phylogenetic framework. *Mol Biol Evol* **32**: 820–832.

Wu C-I. 2001. The genic view of the process of speciation. *J Evol Biol* **14**: 851–865.

Yang H-P, Barbash DA. 2008. Abundant and species-specific DINE-1 transposable elements in 12 Drosophila genomes. *Genome Biol* **9**: R39.

Yang S, Arguello JR, Li X, Ding Y, Zhou Q, Chen Y, Zhang Y, Zhao R, Brunet F, Peng L, et al. 2008. Repetitive element-mediated recombination as a mechanism for new

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

gene origination in Drosophila. *PLoS Genet* **4**: e3.

Young MW, Schwartz HE. 1981. Nomadic gene families in Drosophila. *Cold Spring Harb Symp Quant Biol* **45 Pt 2**: 629–640.

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in Drosophila. *Genome Res* **18**: 1446–1455.

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. *Bioinformatics* **29**: 2669–2677.

Figures

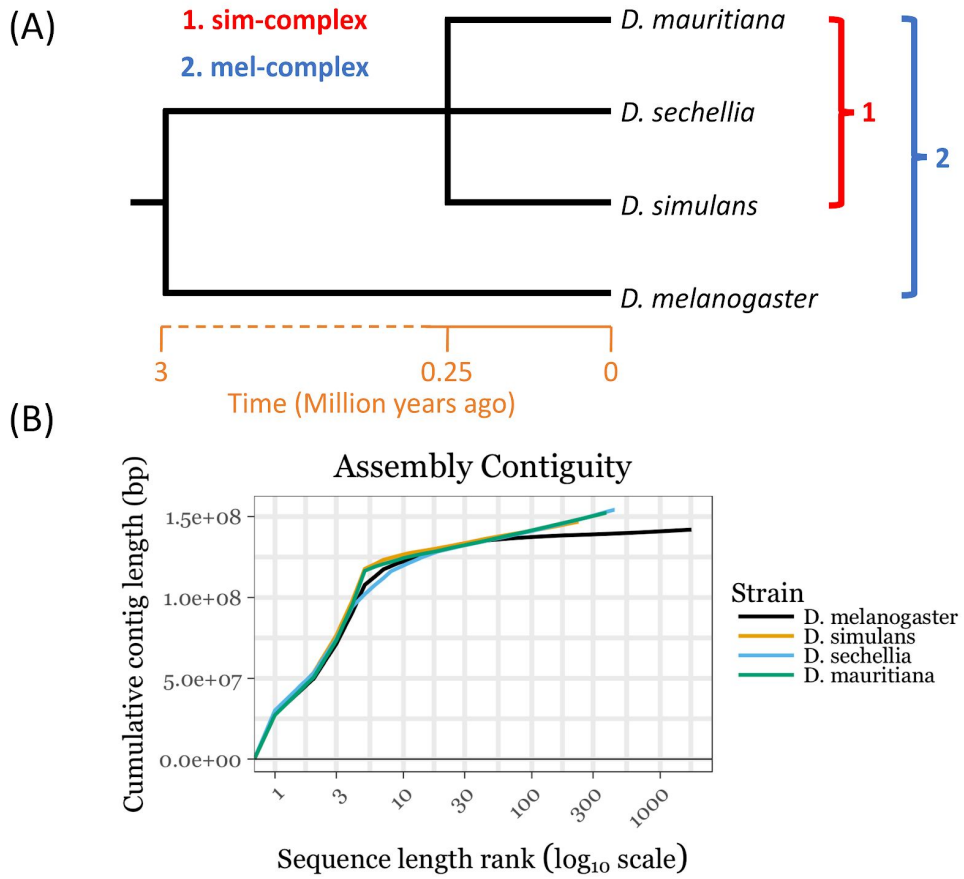


Figure 1. Reference quality *de novo* genome assemblies of the *Drosophila melanogaster* species complex. (A) Phylogeny showing the evolutionary relationship among the members of the *Drosophila simulans* complex and *D. melanogaster.* (B) Contiguities of the new assemblies from *Drosophila simulans* clade and the reference assembly of *D. melanogaster* (R6). The contigs were ranked by their lengths and their cumulative lengths were plotted on the Y-axis. The colors represent different species.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

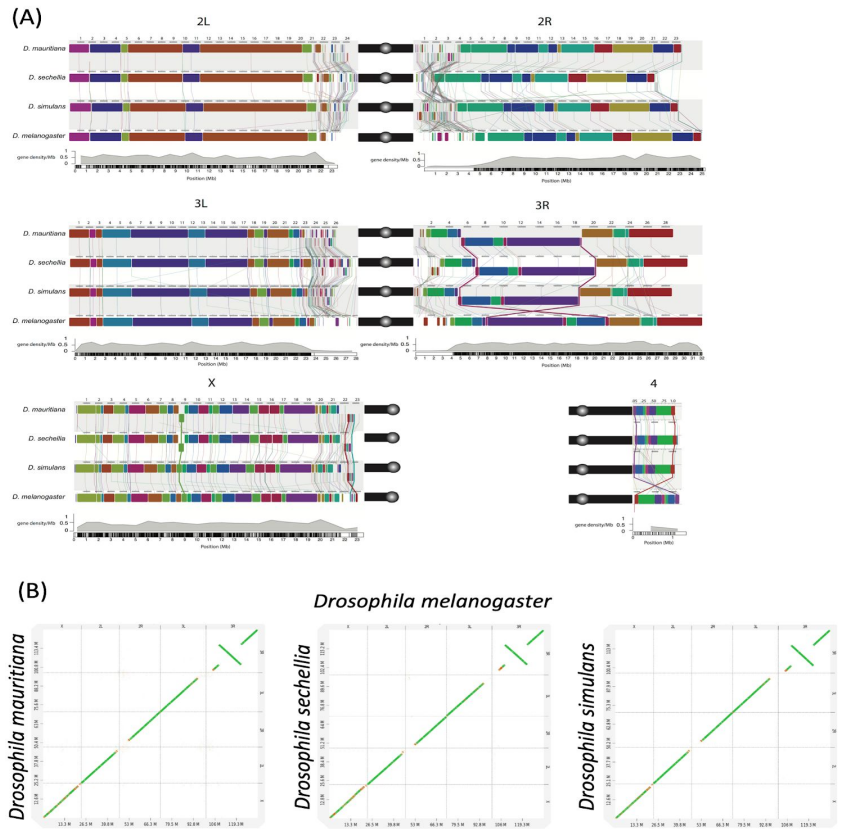


Figure 2. Chromosomal rearrangements in the sim-complex species. We used mauve (A) and D-Genies (Cabanettes and Klopp 2018) (B) to show the syntenies between the members of the simulans species complex and *D. melanogaster*. A) Colored rectangles show positions of syntenic collinear blocks compared to the *D. melanogaster* reference (v6; see details in Materials and Methods). Each chromosome arm is plotted with its own scale with position in megabase indicated above each chromosome. Lines connecting boundaries of colored blocks indicate structural rearrangements. Along the euchromatic chromosome arms, there are three major inversion events (X, 3R, and 4). The heterochromatic regions have significantly more rearrangements than the euchromatin (see text). The heterochromatic regions are marked with a shaded grey bar and the position of the last gene on the scaffolds is indicated for each chromosome arm. B) The dot plots for the whole genome and each chromosome arm between the simulans complex species and *D. melanogaster*. The green segments represent high identity of alignment while the orange segments represent lower identity.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

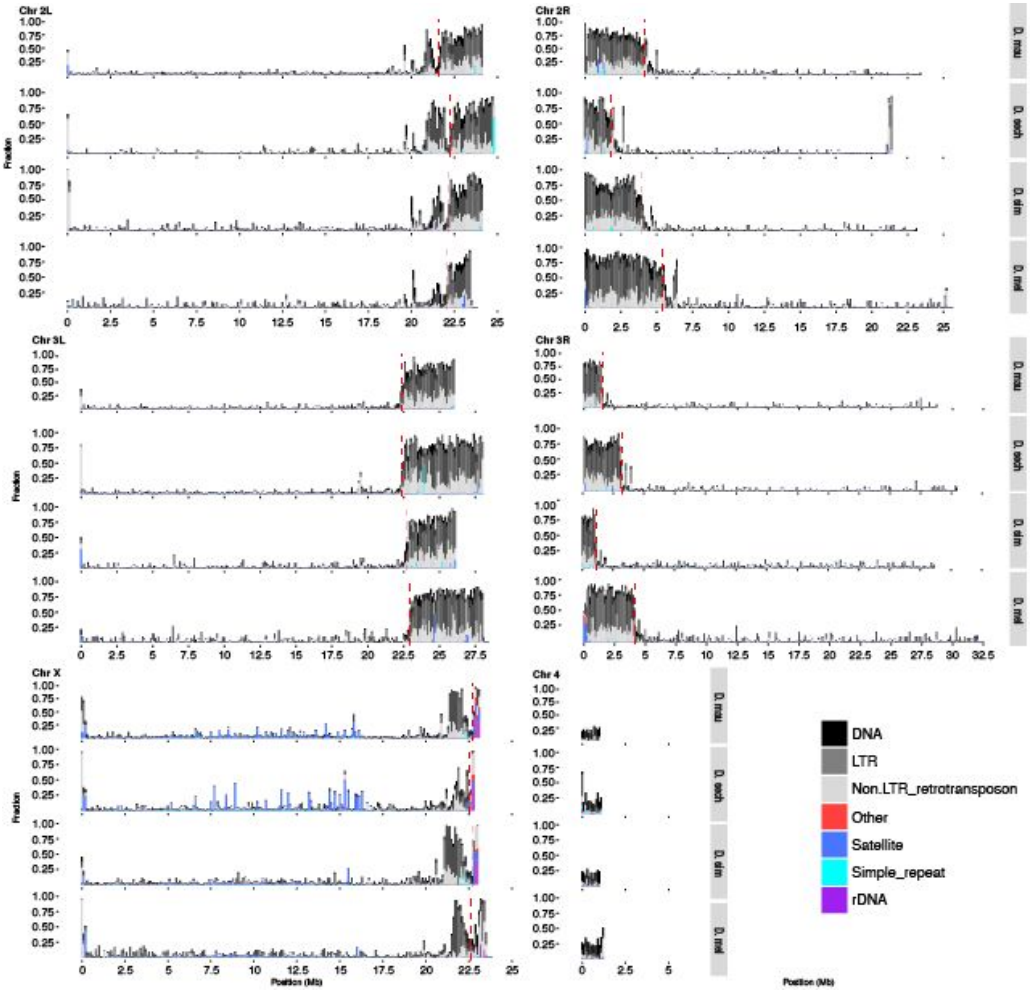


Figure 3. The repeat content across the chromosome arms in *D. simulans* species complex. We estimated the repeat content in the genome using RepeatMasker. Each bar represents the proportion of different repeat types in 100-kb windows. The red dotted lines indicate the euchromatin-heterochromatin boundaries.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

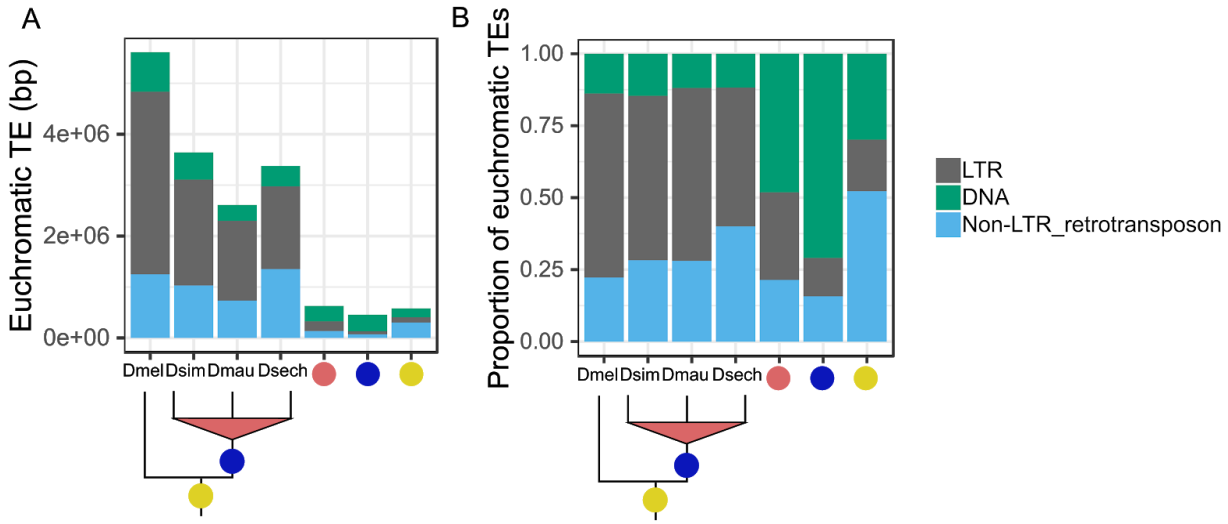


Figure 4. Euchromatic transposon sequence content in each species and their ancestral lineages in the *D. melanogaster* species complex. In each lineage, the bars represent the total content (A) or relative proportion (B) of TE bases due to LTR, DNA, non-LTR retrotransposon TEs. The total bases due to TEs fixed in the ancestral lineages is much smaller than species-specific TE content. The proportion of different TE types is dynamic across the lineages shown here.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

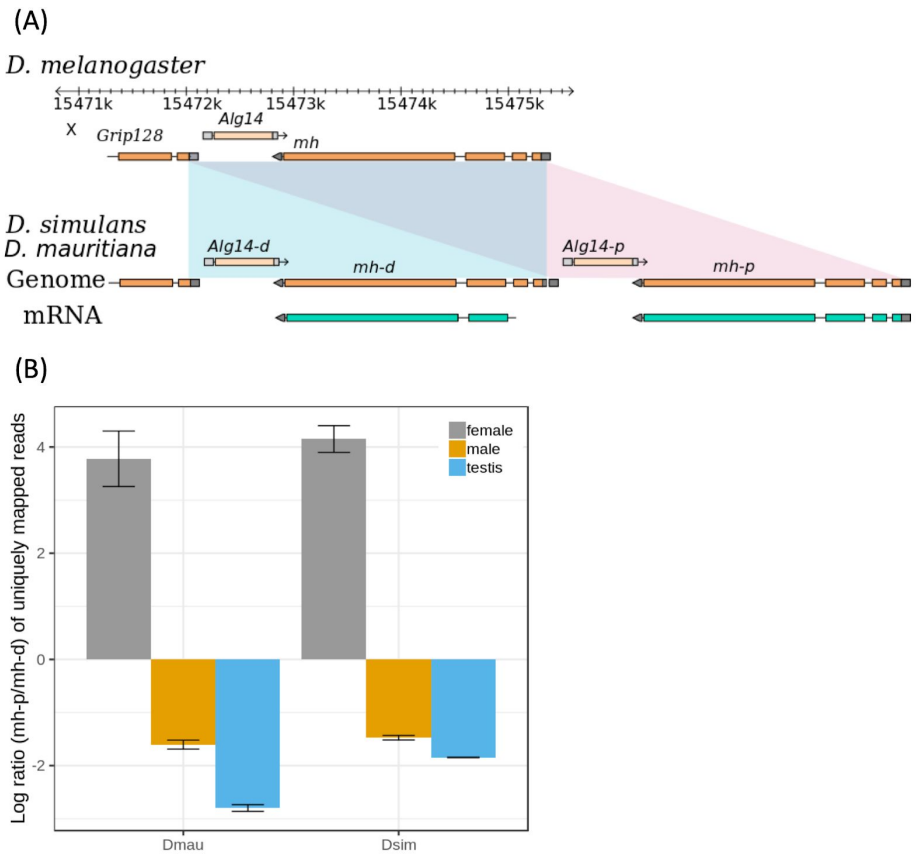


Figure 5. The expression divergence of *maternal haploid* (*mh*) duplicates in the *D. simulans* species complex. A) The sim-complex shares a tandem duplication of *mh* and *alg14* genes. The expression of both *mh* copies is supported by Isoseq and Illumina transcriptome data. B) The proximal copy of *mh* (*mh-p*) is primarily expressed in females and the distal copy (*mh-d*) shows testis-biased expression in both *D. mauritiana* and *D. simulans*.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.27.968743; this version posted February 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

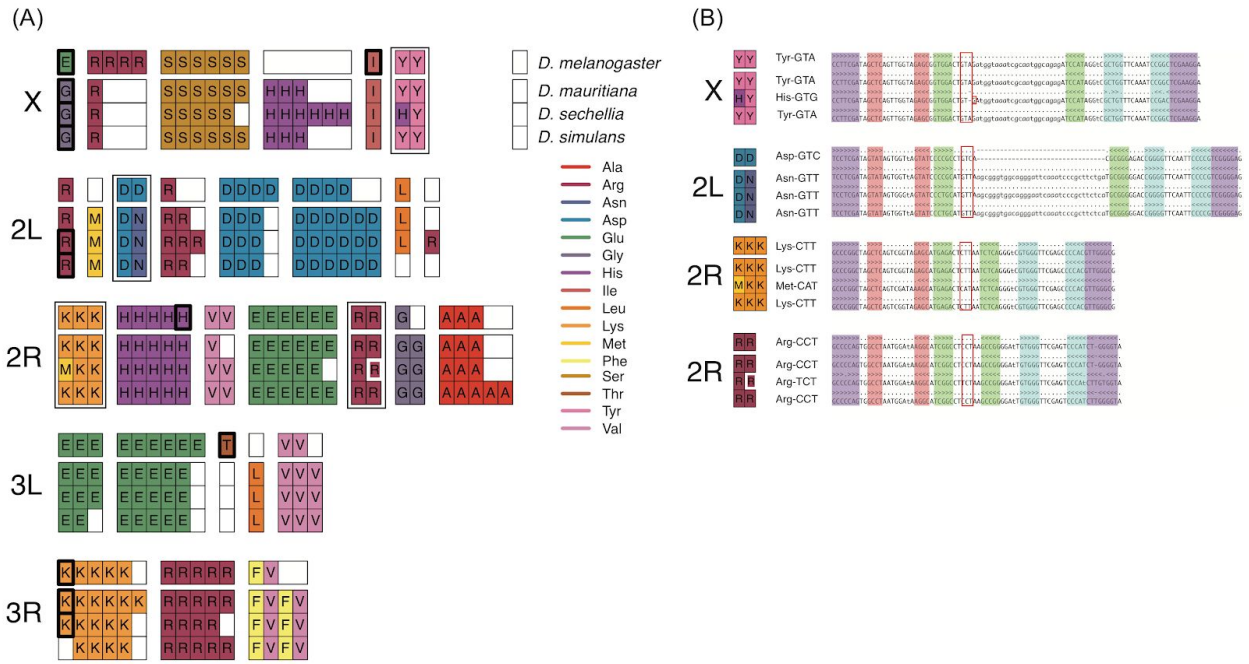
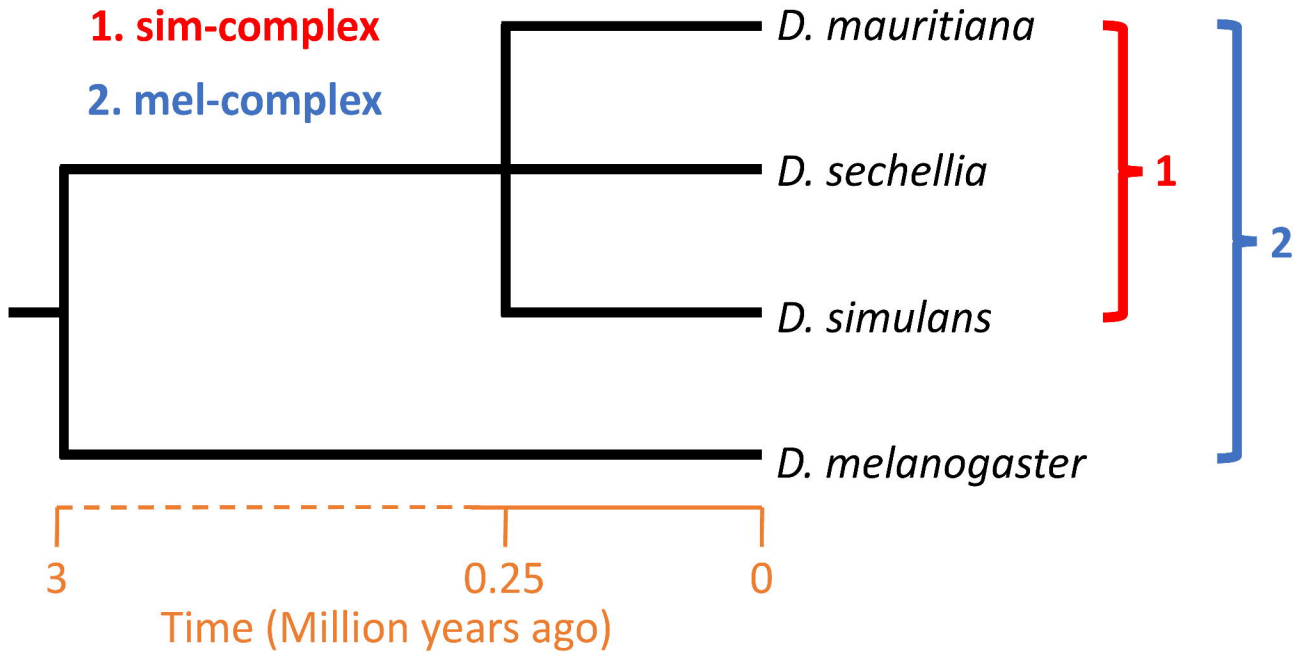
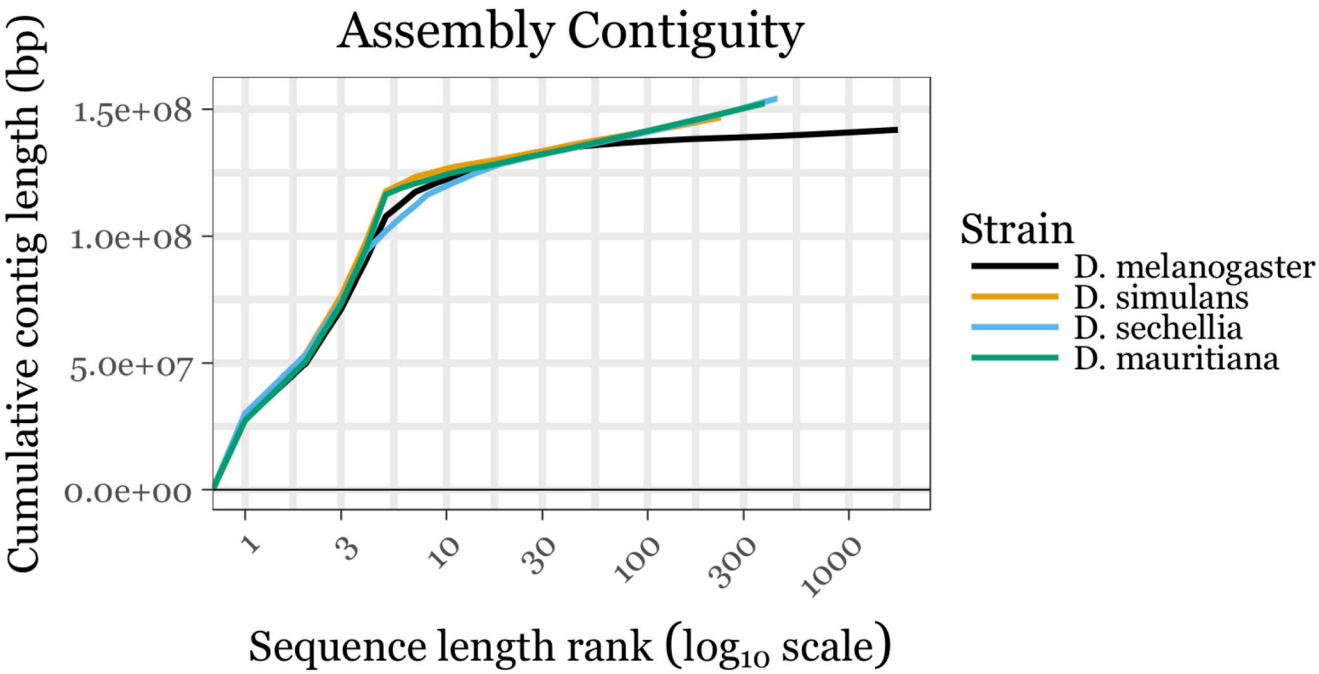


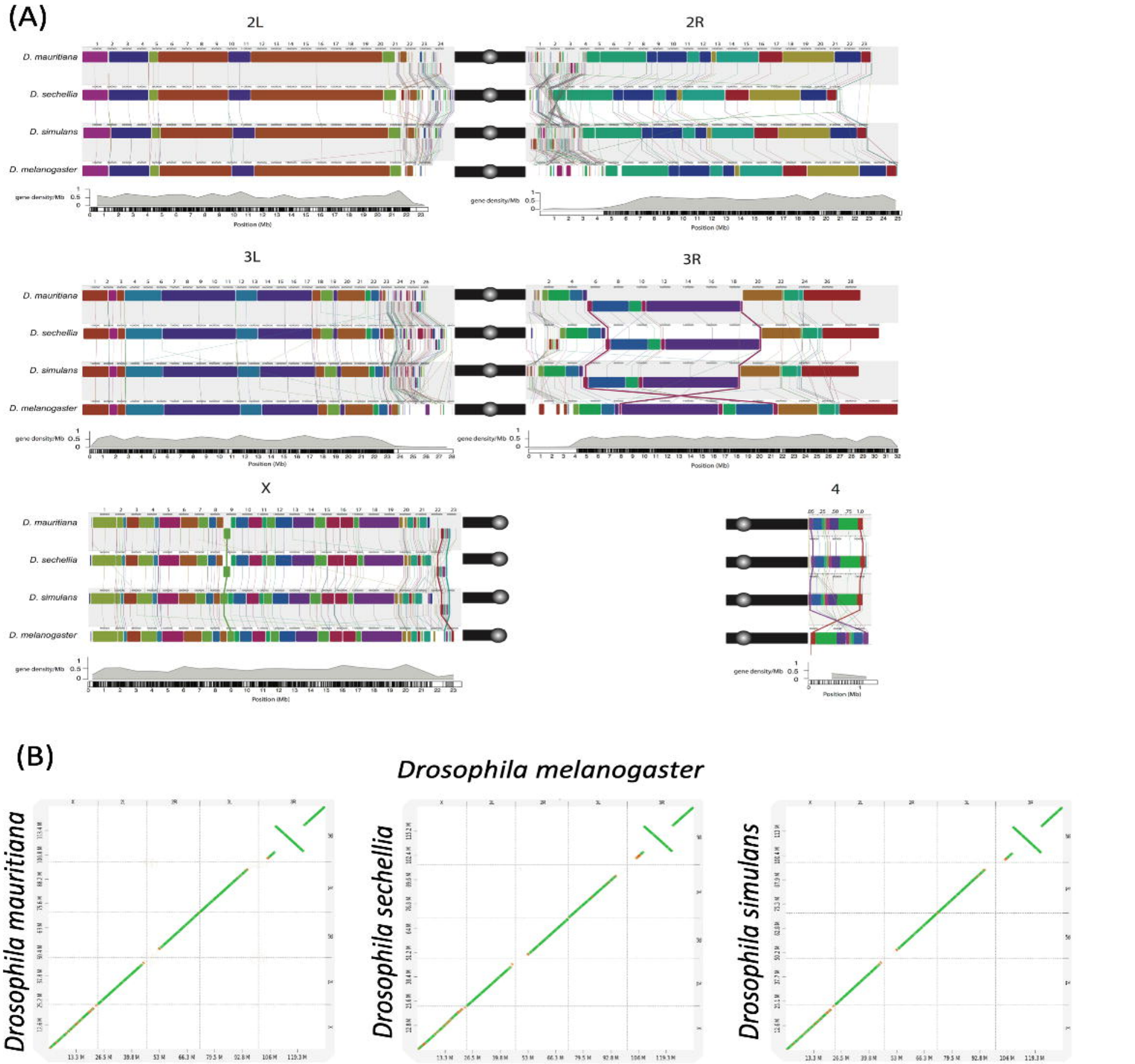
Figure 6. A) Dynamic changes in nuclear tRNA copy-number, isotype identity, and anticodon sequence between *D. melanogaster* and members of the *D. simulans* species complex. Each rectangle represents a single tRNA copy located within a larger syntenic block of tRNAs. Thick black outlines show tRNAs predicted to be pseudogenes. The thick white outline shows a tRNA predicted to utilize a different anticodon. B) Secondary structure alignment of orthologous nuclear tRNAs that experienced an anticodon shift (right). The tRNA anticodon (red box), acceptor stem (highlighted purple), D arm (highlighted red), anticodon arm (highlighted green), and T arm (highlighted blue) are shown in the alignment. The tRNA isotype and anticodon are both shown in the context of their syntenic blocks (left). Secondary structure predictions were generated by tRNAscan-SE (Lowe and Eddy 1997) and secondary structure alignments were performed by hand.

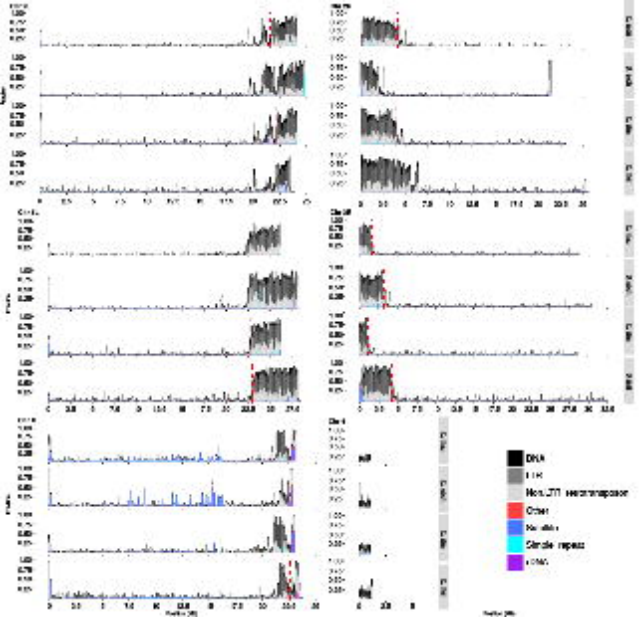
(A)

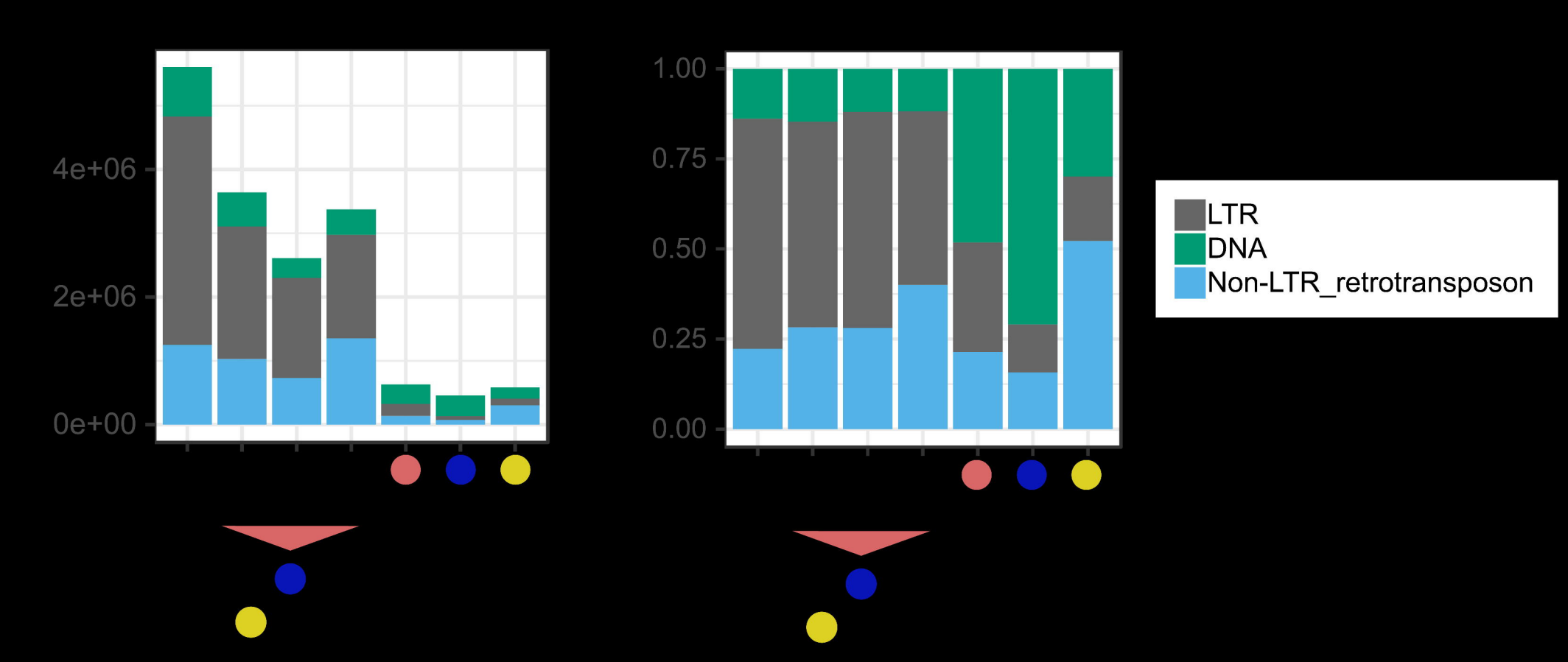

1. sim-complex
2. mel-complex
D. mauritiana
D. sechellia
D. simulans
D. melanogaster
1
2
3
0.25
0
Time (Million years ago)

(B)

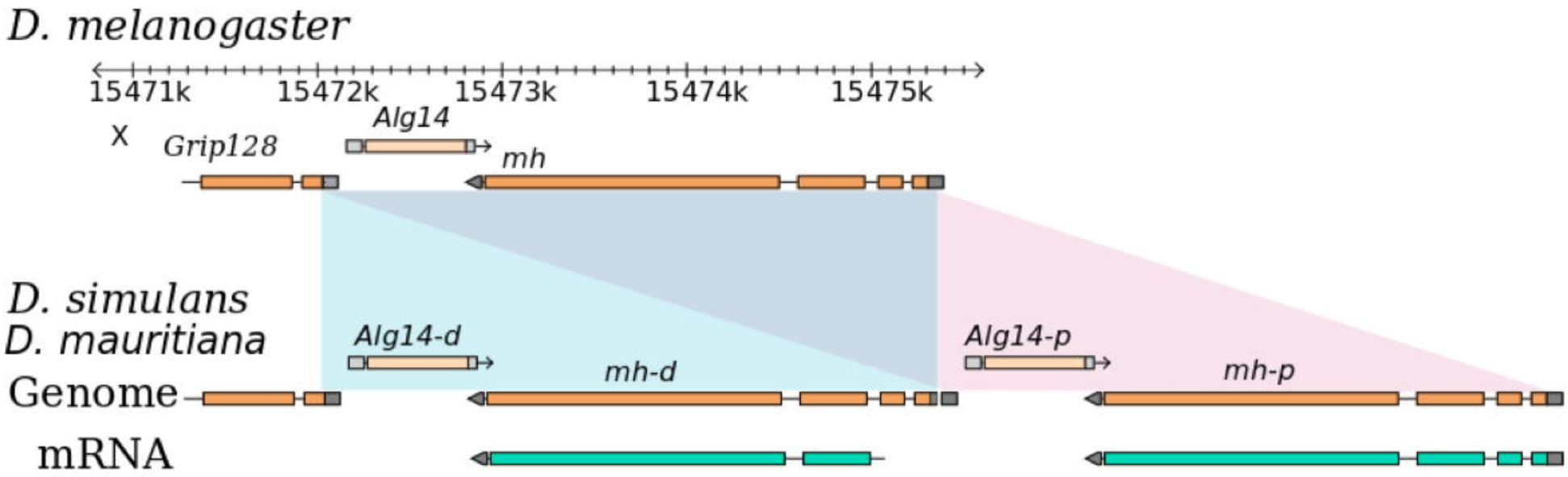

Assembly Contiguity
Cumulative contig length (bp)
1.5e+08
1.0e+08
5.0e+07
0.0e+00
1
3
10
30
100
300
1000
Sequence length rank (log10 scale)
Strain
D. melanogaster
D. simulans
D. sechellia
D. mauritiana


(A)
2L
2R
3L
3R
X
4
D. mauritiana
D. sechellia
D. simulans
D. melanogaster
gene density/Mb
Position (Mb)
(B)
Drosophila melanogaster
Drosophila mauritiana
Drosophila sechellia
Drosophila simulans

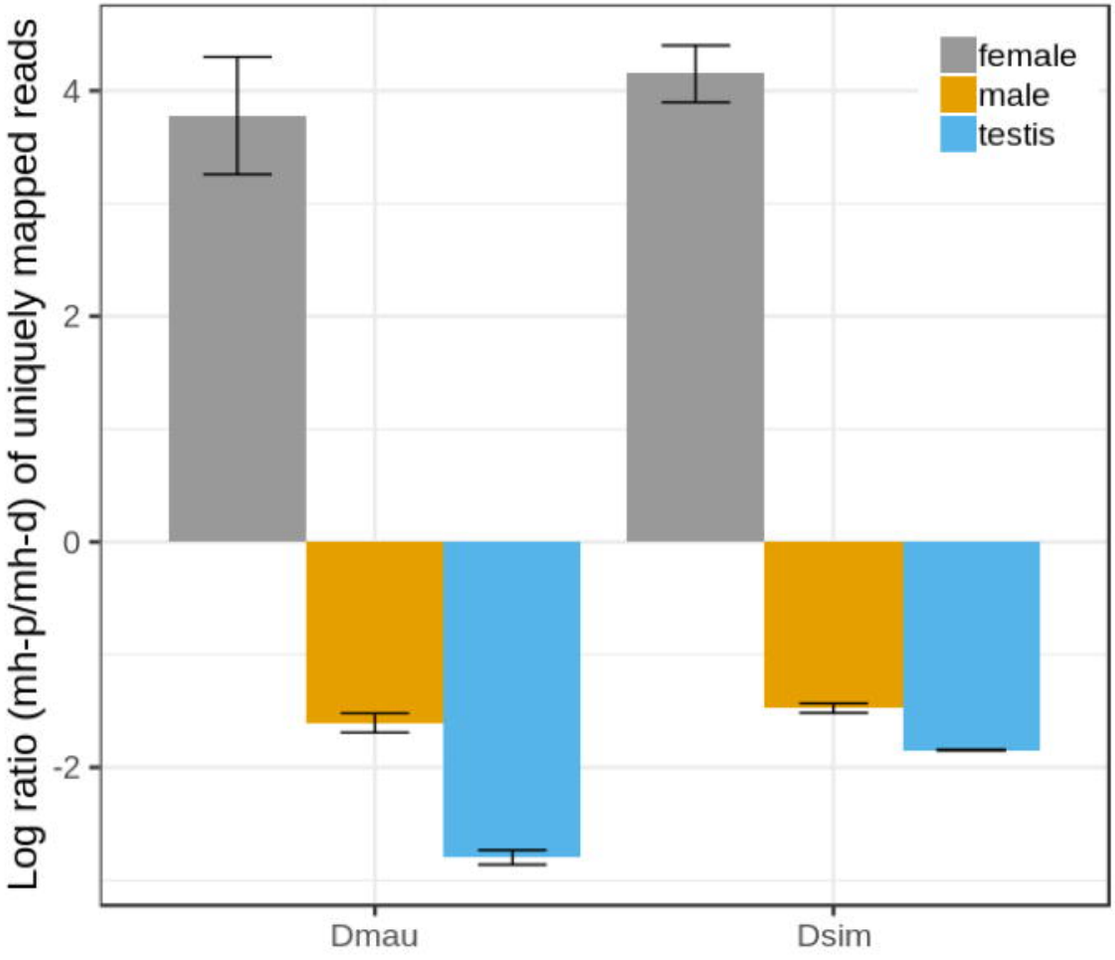

DNA
LINE
Non-LTR retrotransposon
Other
Satellite
Simple repeat
LTR

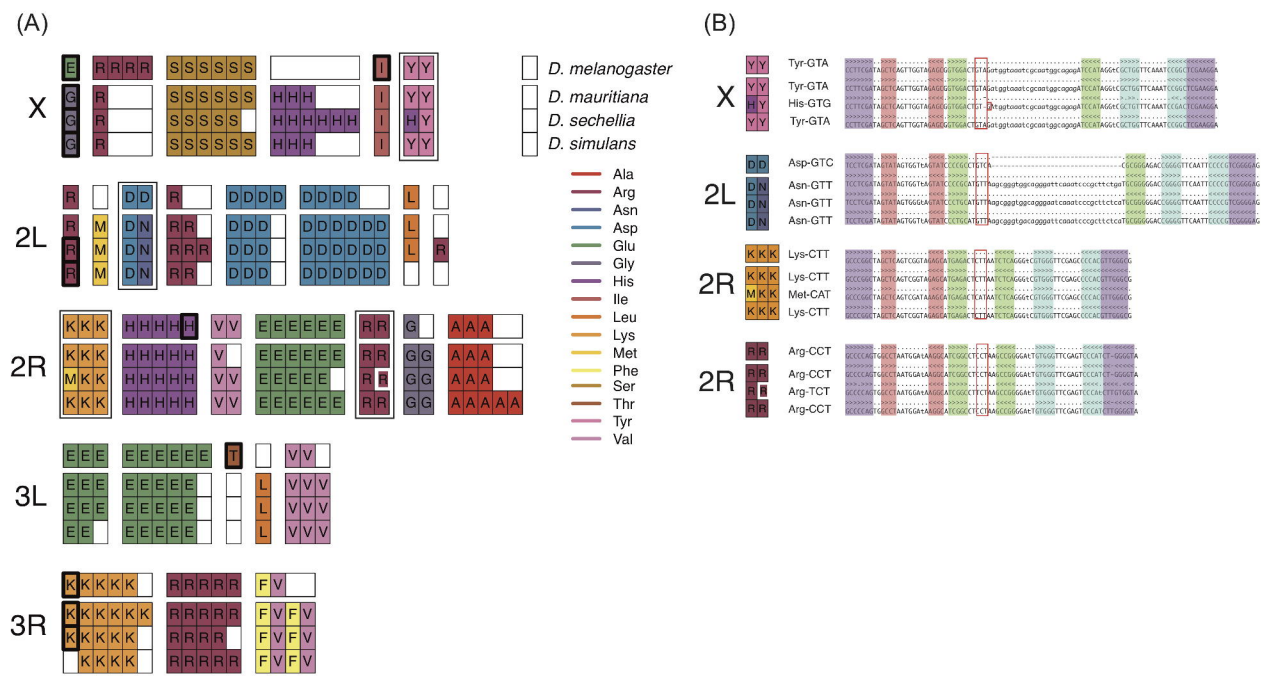

4e+06
2e+06
0e+00
1.00
0.75
0.50
0.25
0.00
LTR
DNA
Non-LTR_retrotransposon

(A)


D. melanogaster
15471k
15472k
15473k
15474k
15475k
X
Grip128
Alg14
mh
D. simulans
D. mauritiana
Genome
Alg14-d
mh-d
Alg14-p
mh-p
mRNA

(B)


Log ratio (mh-p/mh-d) of uniquely mapped reads
4
2
0
-2
female
male
testis
Dmau
Dsim


(A)
X
2L
2R
3L
3R
D. melanogaster
D. mauritiana
D. sechellia
D. simulans
Ala
Arg
Asn
Asp
Glu
Gly
His
Ile
Leu
Lys
Met
Phe
Ser
Thr
Tyr
Val
(B)
X
Tyr-GTA
Tyr-GTA
His-GTG
Tyr-GTA
2L
Asp-GTC
Asn-GTT
Asn-GTT
Asn-GTT
2R
Lys-CTT
Lys-CTT
Met-CAT
Lys-CTT
2R
Arg-CCT
Arg-CCT
Arg-TCT
Arg-CCT

Assembly	Assembly Length	NG10	NG25	NG50	NG75	NG90	LG10	LG25	LG50	LG75	LG90
D. melanogaster contig	142,575,135	27,905,053	21,907,215	19,478,218	5,107,766	83,949	1	2	4	6	43
D. simulans contig	146,772,009	27,689,984	25,882,106	22,909,593	19,705,053	198,872	1	2	3	5	37
D. sechellia contig	154,185,598	30,225,992	23,008,689	19,907,079	4,183,003	168,616	1	2	4	8	45
D. mauritiana contig	152,314,381	27,404,031	23,215,807	22,120,385	21,002,293	174,794	1	2	4	5	44
D. melanogaster scaffold	143,726,002	32,079,331	28,110,227	25,286,936	23,513,712	3,667,352	1	2	3	5	6
D. simulans scaffold	146,773,509	28,742,805	26,221,060	24,213,528	23,123,156	222,038	1	2	3	5	28
D. sechellia scaffold	154,189,398	30,464,902	28,131,630	24,956,976	21,536,224	201,855	1	2	3	5	19
D. mauritiana scaffold	152,317,181	28,872,862	26,143,691	24,228,064	23,222,241	228,686	1	2	3	5	26