bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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 4.0 International license.

Unique structure and positive selection promote the rapid divergence of Drosophila Y chromosomes

Ching-Ho Chang[1#*], Lauren E. Gregory[1], Kathleen E. Gordon[2^], Colin D. Meiklejohn[2] and Amanda M. Larracuente[1*]

Affiliations:

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

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

[#]Current address: Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109

^Current address: Department of Molecular Biology and Genetics, Field of Genetics, Genomics and Development, Cornell University, Ithaca, NY, 14853

*correspondence to cchang2@fredhutch.org, alarracu@ur.rochester.edu

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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 4.0 International license.

Abstract

Y chromosomes across diverse species convergently evolve a gene-poor, heterochromatic organization enriched for duplicated genes, LTR retrotransposable elements, and satellite DNA. Sexual antagonism and a loss of recombination play major roles in the degeneration of young Y chromosomes. However, the processes shaping the evolution of mature, already degenerated Y chromosomes are less well-understood. Because Y chromosomes evolve rapidly, comparisons between closely related species are particularly useful. We generated de novo long read assemblies complemented with cytological validation to reveal Y chromosome organization in three closely related species of the *Drosophila simulans* complex, which diverged only 250,000 years ago and share >98% sequence identity. We find these Y chromosomes are divergent in their organization and repetitive DNA composition and discover new Y-linked gene families whose evolution is driven by both positive selection and gene conversion. These Y chromosomes are also enriched for large deletions, suggesting that the repair of double-strand breaks on Y chromosomes may be biased toward microhomology-mediated end joining over canonical non-homologous end-joining. We propose that this repair mechanism generally contributes to the convergent evolution of Y chromosome organization.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

Introduction

Most sex chromosomes evolved from a pair of homologous gene-rich autosomes that acquired sex-determining factors and subsequently differentiated. Y chromosomes gradually lose most of their genes, while their X chromosome counterparts tend to retain the original autosomal complement of genes. This Y chromosome degeneration follows a suppression of recombination [1], which limits the efficacy of natural selection, and causes the accumulation of deleterious mutations through Muller's ratchet, background selection, and hitchhiking effects [2-6]. As a consequence, many Y chromosomes present a seemingly hostile environment for genes, with their mutational burden, high repeat content and abundant silent chromatin.

Genomic studies of Y chromosome evolution focus primarily on young sex chromosomes, addressing how the suppression of recombination promotes Y chromosome degeneration at both the epigenetic and genetic levels [2, 7]. Although sexually antagonistic selection is traditionally cited as the cause of recombination suppression on the Y chromosome, direct evidence for its role is still lacking [8] and new models propose that regulatory evolution is the initial trigger for recombination suppression [9]. Sexually antagonistic selection may accelerate Y-linked gene evolution to optimize male-specific functions. Indeed, Y-linked genes tend to have slightly higher rates of protein evolution than their orthologs on other chromosomes [10, 11]. Higher rates of Y-linked gene evolution are driven by positive selection, relaxed selective constraints and male-biased mutation patterns, with most Y-linked genes evolving under

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

at least some functional constraint [11]. Although there is evidence suggesting that some Y chromosomes have experienced recent selective sweeps [12, 13], the relative importance of positive selection for Y chromosome evolution remains unclear.

Y chromosomes harbor extensive structural divergence between species, in part through the acquisition of genes from other genomic regions [14-21]. However, the functions of most Y-linked genes are unknown [18, 21-23]. Some Y-linked genes are duplicated and, in extreme cases, amplified into so-called ampliconic genes—gene families with tens to hundreds of highly similar sequences. Y chromosomes of both *Drosophila* and mammals have independently acquired and amplified gene families, which turnover rapidly between closely related species [14, 17, 20, 24-26]. Following Y-linked gene amplification, gene conversion between gene copies may enhance the efficacy of selection on Y-linked genes in the absence of crossing over [15, 27].

Detailed analyses of old Y chromosomes have been restricted to a few species with reference-quality assemblies, *e.g.*, mouse and human. The challenges of cloning and assembling repeat-rich regions of the genome have stymied progress towards a complete understanding of Y chromosome evolution [28-30]. Recent advances in long-read sequencing make it feasible to assemble large parts of Y chromosomes [19, 21, 22, 31] enabling comparative studies of a majority of Y-linked sequences in closely related species.

Drosophila melanogaster and three related species in the *D. simulans* clade are ideally suited to study Y chromosome evolution. These Y chromosomes are functionally

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

divergent, contribute to hybrid sterility [32-35], and at least four X-linked meiotic drive systems likely shape Y chromosome evolution in these species [36-43]. Previous genetic and transcriptomic studies suggest that Y chromosome variation can impact male fitness and gene regulation [44-51]. Since there is minimal nucleotide variation and divergence in Y-linked protein-coding sequences within and between these *Drosophila* species [11, 12, 40], structural variation may be responsible for the majority of these effects. For example, 20-40% of *D. melanogaster* Y-linked regulatory variation (YRV) comes from differences in ribosomal DNA (rDNA) copy numbers [52, 53]. The chromatin on *Drosophila* Y chromosomes has genome-wide effects on expression level and chromatin states [54], but aside from the rDNA, the molecular basis of Y chromosome divergence and variation in these species remains elusive.

To better understand Y chromosome structure and evolution, we assembled the Y chromosomes of the three species in the *D. simulans* clade and compared them to *D. melanogaster*. We observe that the Y chromosomes of the *D. simulans* clade species have high duplication and gene conversion rates that, along with strong positive selection, shaped the evolution of two new ampliconic protein-coding gene families. We propose that, in addition to positive selection, sexual antagonism, and genetic conflict, differences in the usage of DNA repair pathways may give rise to the unique patterns of Y-linked mutations. Together these effects may drive the convergent evolution of Y chromosome structure across taxa.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

Results

Improving the Y chromosome assemblies using long-read assembly and fluorescence in situ hybridization (FISH)

Long reads have enabled the assembly of many repetitive genome regions, but have had limited success in assembling Y chromosomes [17, 19, 21, 22]. To improve Y chromosome assemblies for comparative genomic analyses, we applied our heterochromatin-sensitive assembly pipeline [22] with long reads that we previously generated [55] to reassemble the Y chromosome from the three species in the *Drosophila simulans* clade. We also resequenced male genomes using PCR-free Illumina libraries to polish these assemblies. Our heterochromatin-enriched methods improve contiguity compared to previous *D. simulans* clade assemblies. We recovered all known exons of the 11 canonical Y-linked genes conserved across the *melanogaster* group, including 58 exons missed in previous assemblies (Table S1; [55]). Based on the median male-to-female coverage [22], we assigned 13.7 to 18.9 Mb of Y-linked sequences per species with N50 ranging from 0.6 to 1.2 Mb. The quality of these new *D. simulans* clade Y assemblies are comparable to *D. melanogaster* (Table 1; [22]). We evaluated our methods by comparing our assignments for every 10-kb window of assembled sequences to its known chromosomal location. Our assignments have 96, 98, and 99% sensitivity and 5, 0, and 3% false-positive rates in *D. mauritiana*, *D. simulans*, and *D. sechellia*, respectively (Table S2). We have lower confidence in our *D. mauritiana* assignments, because the male and female Illumina reads are from different library construction methods. Therefore, we applied an additional criterion only in *D. mauritiana* based on the female-to-male total mapped reads ratio (<0.1), which reduces

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

the false-positive rate from 13 to 5% in regions with known chromosomal location (Table S2; Fig S1). Based on these chromosome assignments, we find 40–44% lower PacBio coverage on Y than X chromosomes in all three species (Fig S2).

Table 1. Contiguity statistics for heterochromatin-enriched assemblies

Y chromosome assembly	# of contigs	Total length	Contigs N50
D. melanogaster[a]	80	14,578,684	416,887
D. mauritiana[b]	55	17,880,069	1,628,994
D. simulans[b]	38	13,717,056	1,031,383
D. sechellia[b]	63	14,899,148	555,130

[a] Chang and Larracuente 2019
[b] This paper

The cytological organization of the *D. simulans* clade Y chromosomes is not well-described [56-58]. Therefore, we generated new physical maps of the Y chromosomes by combining our assemblies with cytological data. We performed FISH on mitotic chromosomes using probes for 12 Y-linked sequences (Fig 1 and S3–4; Table S3) to determine Y chromosome organization at the cytological level. We also determined the location of the centromeres using immunostaining with a Cenp-C antibody (Fig S4; [59]). These cytological data permit us to 1) validate our assemblies, and 2) infer the overall organization of the Y chromosome by orienting our scaffolds on cytological maps. Of the 11 Y-linked genes, we successfully ordered 10, 11, and 7 genes on the cytological bands of *D. simulans, D. mauritiana* and *D. sechellia*, respectively (Fig 1 and S3). We find evidence for extensive Y chromosomal structural rearrangements, including changes in satellite distribution, gene order, and centromere position. These

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

rearrangements are dramatic even among the *D. simulans* clade species, which diverged less than 250 KYA (Fig 1 and S3). The Y chromosome centromere position appears to be the same as determined by Berloco et al. for different strains of *D. simulans* and *D. mauritiana*, but not for *D. sechellia* [58]. One explanation for this discrepancy could be between-strain variation in *D. sechellia* Y chromosome centromere location. Together, our new physical maps and assemblies provide both large and fine-scale resolution on Y chromosome organization in the *D. simulans* clade.

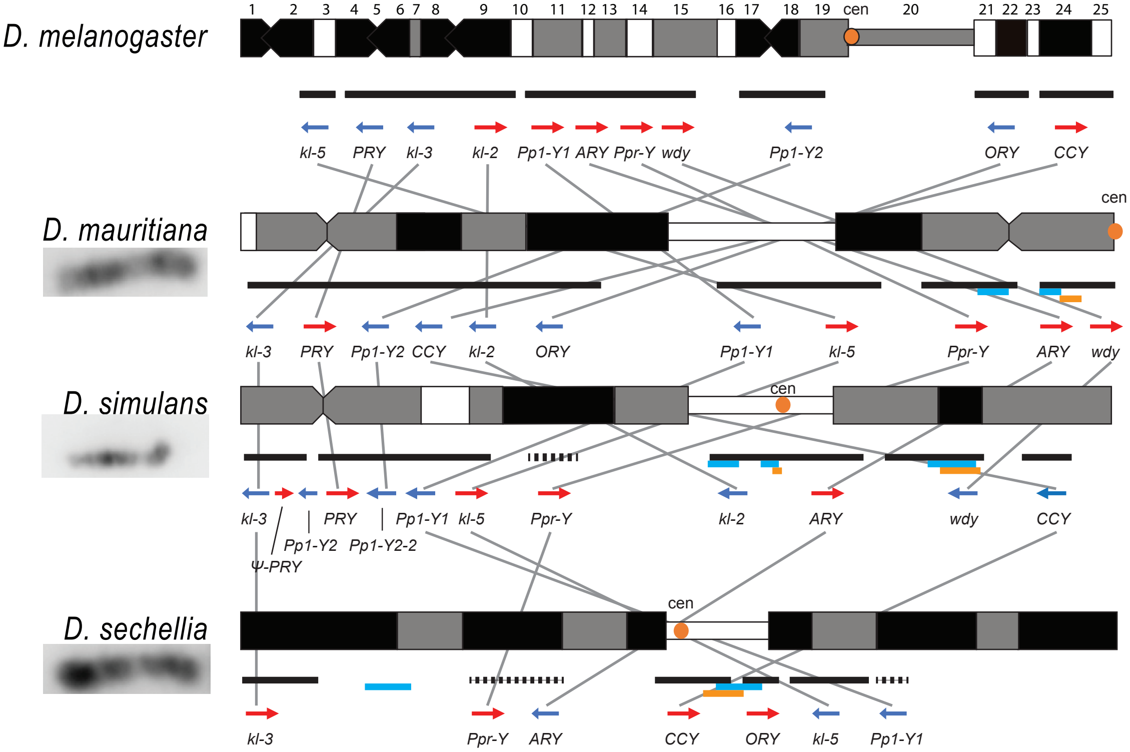


Fig 1. Y chromosome organization in *D. melanogaster* and the three *D. simulans* clade species. Schematics of the cytogenetic maps note the locations of Y-linked genes in *D. melanogaster* and *D. simulans* clade species. The bars show the relative placement of the scaffolds on the cytological bands based on FISH results. The solid black and dotted bars represent the scaffolds with known and unknown orientation information, respectively. The light blue and orange bars represent two new Y-linked gene families, *Lhk* and *CK2ßtes-Y* in the *D. simulans* clade, respectively. The arrows indicate the orientation of the genes (blue- minus strand; red- plus strand).

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

Y-linked sequence and copy number divergence across three species

Although the *D. simulans* clade species diverged only recently, Y chromosome introgression between pairs of species disrupts male fertility and influences patterns of genome-wide gene expression [32, 34]. One candidate locus that may contribute to functional divergence and possibly hybrid lethality is the Y-linked rDNA [52, 60]. Y-linked rDNA, specifically 28S rDNA, have been lost in *D. simulans* and *D. sechellia*, but not in *D. mauritiana* [57, 61, 62]. However, the intergenic spacer (IGS) repeats between rDNA genes, which are responsible for X-Y pairing in *D. melanogaster* males [63], are retained on both sex chromosomes in all three species [57, 61, 62]. Consistent with previous cytological studies [57, 61, 62], we find that *D. simulans* and *D. sechellia* lost most Y-linked 18S and 28S rDNA sequences (Fig S5). Our assemblies indicate that, despite this loss of the rRNA coding sequences, all three species still retain IGS repeats. However, we and others do not detect Y-linked IGS repeats at the cytological level in *D. sechellia* (Fig S3–4; [57, 61, 62]), suggesting that their abundance is below the level of detection by FISH in this species.

Structural variation at Y-linked genes may also contribute to functional variation and divergence in the *D. simulans* clade. Previous studies reported many duplications of canonical Y-linked genes in *D. simulans* [40, 55, 64]. We find that all three species have at least one intact copy of the 11 canonical Y-linked genes, but there is also extensive copy number variation in Y-linked exons across these species (Figure 2 and S6–7, Table S1; [65]). Using Illumina reads, we confirm the copy number variation in our assemblies, and further reveal some Y-linked duplicated exons, particularly in *kl-3*, *wdy* and *Ppr-Y*, that are not assembled in *D. sechellia* (Fig S6). Some duplicates may be

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

functional because they are expressed and have complete open reading frames, (*e.g.*, *ARY*, *Ppr-Y1* and *Ppr-Y2*). The *D. simulans* Y chromosome has four complete copies of *ARY*, all of which show similar expression levels from RNA-seq data (Figure 2B and Table S4), but two copies have inverted exons 1 and 2. *D. sechellia* also contains at least five duplicated copies of *ARY,* some of which also have the inverted exons 1 and 2, but the absence of RNA-seq data from testes of this species prevents inferences regarding whether all copies of *ARY* are expressed. However, most duplications include only a subset of exons, and in many cases, the duplicated exons are located on the periphery of the presumed functional gene copy (Figure 2B and S7, Table S4). For example, both *D. simulans* and *D. mauritiana* have multiple copies of exons 8-12 located at the 3' end of *kl-2* (Figure 2B). In *D. simulans,* most of these extra exons have low to no expression, while in *D. mauritiana*, there appears to be a substantial expression from many of the duplicated terminal exons, as well as an internal duplication of exon 5. It is unclear what effects these duplicated exons have on the protein sequences of these fertility-essential genes.

All exon-intron junctions are conserved within full-length copies of the canonical Y-linked genes, yet intron lengths vary between these species (Fig 3). The length of longer introns (>100 bp in any species) is more dynamic than that of short introns (Fig 3; Table S5). The dramatic size differences in most introns cannot be attributed to a single deletion or duplication (see an example of *ORY* in Fig S8). Some Y-linked genes contain mega-base sized introns (*i.e.*, mega-introns) whose transcription manifests as cytologically visible lampbrush-like loops (Y-loops) in primary spermatocytes [66, 67]. While Y-loops are found across the *Drosophila* genus [68, 69], their potential functions

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

are unknown [70-74] and the genes/introns that produce Y-loops differs among species [75] (Supplemental text). *D. melanogaster* has three Y-loops transcribed from introns of *ORY* (*ks-1* in previous literature), *kl-3*, and *kl-5* [66]. Based on cytological evidence, *D. simulans* has three Y-loops, whereas *D. mauritiana* and *D. sechellia* only have two [69]. Of all potential loop-producing introns, we find that only the *kl-3* mega-intron is conserved in all four species and has the same intron structure and sequences (*i.e.,* $(AATAT)_n$ repeats). While both *kl-5* and *ORY* produce Y-loops with $(AAGAC)_n$ repeats in *D. melanogaster*, $(AAGAC)_n$ is missing from the genomes of the *D. simulans* clade species. This observation is supported by our assemblies, the Illumina raw reads (Table S3), and published FISH results [76]. In the *D. simulans* clade, the *ORY* introns do not carry any long tandem repeats. However, *kl-5* has introns with $(AATAT)_n$ repeats that may form a Y-loop in the *D. simulans* clade species. These data suggest that, while mega-introns and Y-loops may be conserved features of spermatogenesis in *Drosophila*, they turn over at both the sequence and gene levels over short periods of evolutionary time (*i.e.*, ~2 My between *D. melanogaster* and the *D. simulans* clade).

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

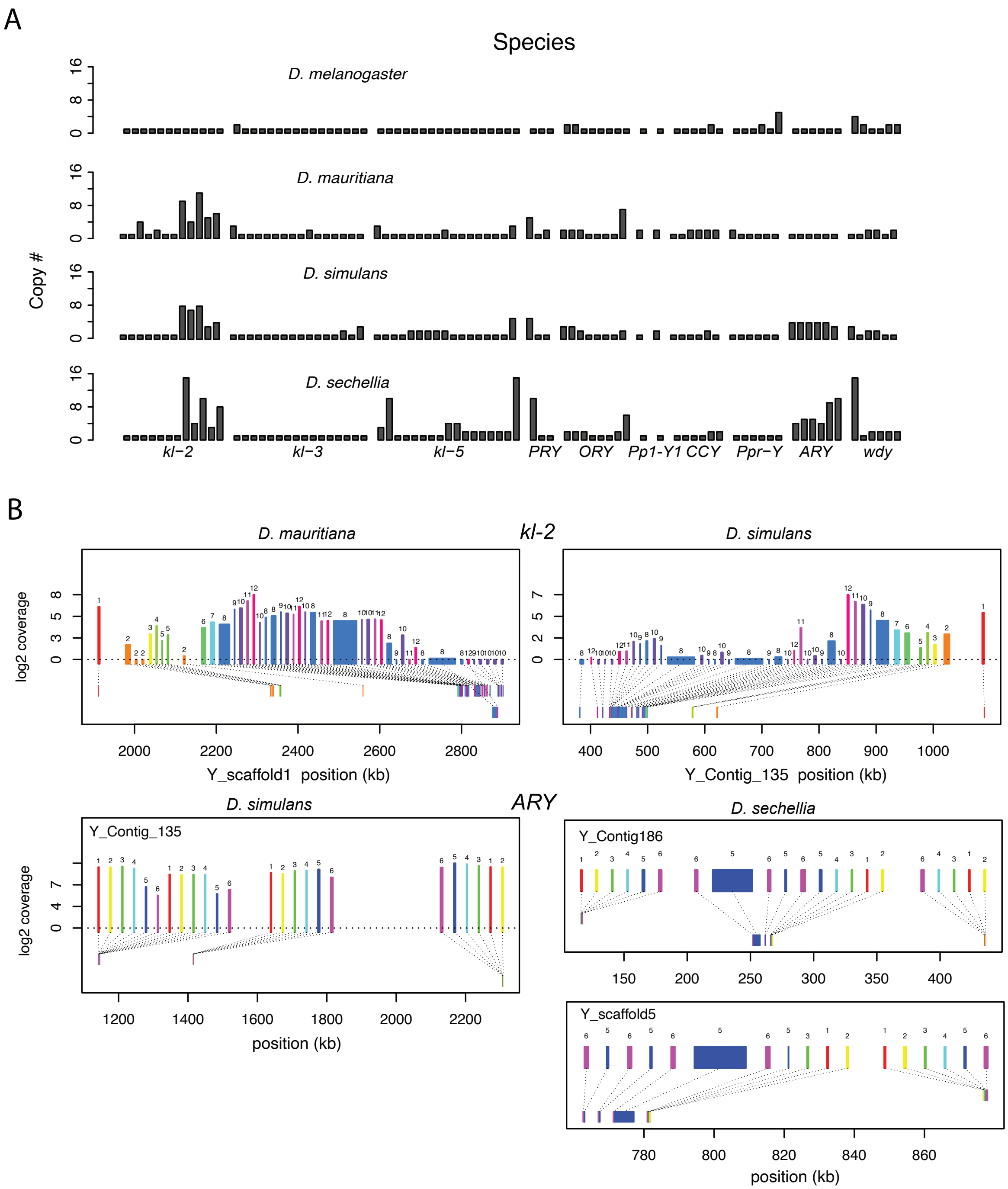


Fig 2. Duplication of canonical Y-linked exons. A) Exon copy number is highly variable across the three *D. simulans* clade species and generally greater than in *D. melanogaster.* B) Gene structure of *kl-2* and *ARY* inferred from assemblies and RNA-seq data. Upper bars indicate exons that are colored and numbered, with their height showing average read depth from sequenced testes RNA (*D. simulans* and *D. mauritiana* only). Lower bars indicate exon positions on the assembly and position on the Y-axis indicates coding strand.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

Consistent with previous studies [18, 55], we identify high rates of gene duplication to the *D. simulans* clade Y chromosome from other chromosomes. We find 49 independent duplications to the Y chromosome in our heterochromatin-enriched assemblies (Fig 4; Table S6), including eight newly discovered duplications [18, 55]. Twenty-eight duplications are DNA-based, 13 are RNA-based, and the rest are unknown due to the limited sequence information (Table S6). The rate of transposition to the Y chromosome is about 3–4 times higher in the *D. simulans* clade compared to *D. melanogaster* [22]. We also infer that 17 duplicated genes were independently deleted from *D. simulans* clade Y chromosomes. Based on transcriptomes from *D. simulans* and *D. mauritiana* testes, we suspect that more than half of the duplicated genes are likely pseudogenes that either show no expression in testes (< 3 TPM) or lack open reading frames (< 100 amino acids; Table S6). We also detect intrachromosomal duplications of these Y-linked pseudogenes (Table S6), suggesting a high duplication rate within these Y chromosomes.

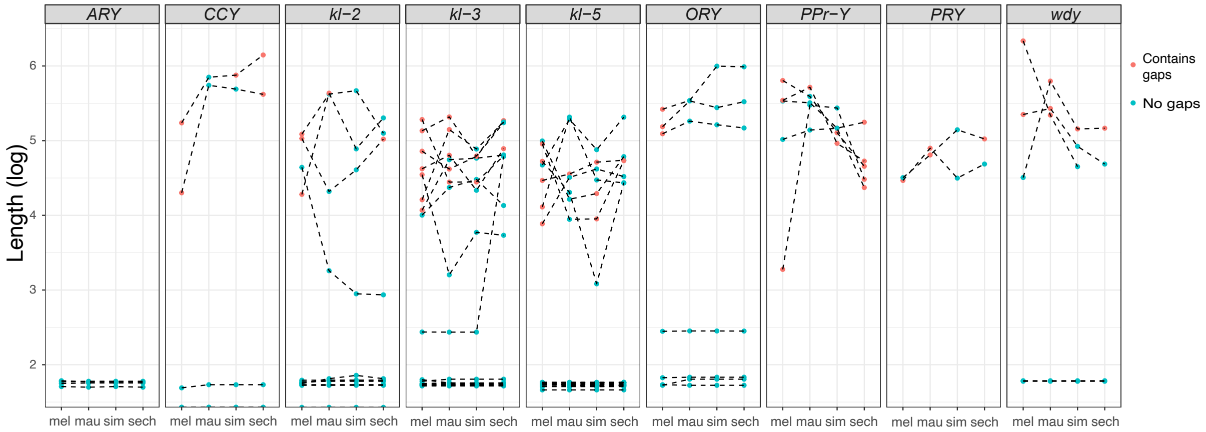


Fig 3. Evolution of intron lengths in canonical Y-linked genes. The intron length in canonical Y-linked genes is different between *D. melanogaster* and the three *D.*

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

simulans clade species. Orthologous introns are connected by dotted lines. Completely assembled introns are in blue and introns with gaps in the assembly are in red, and are therefore minimum intron lengths.

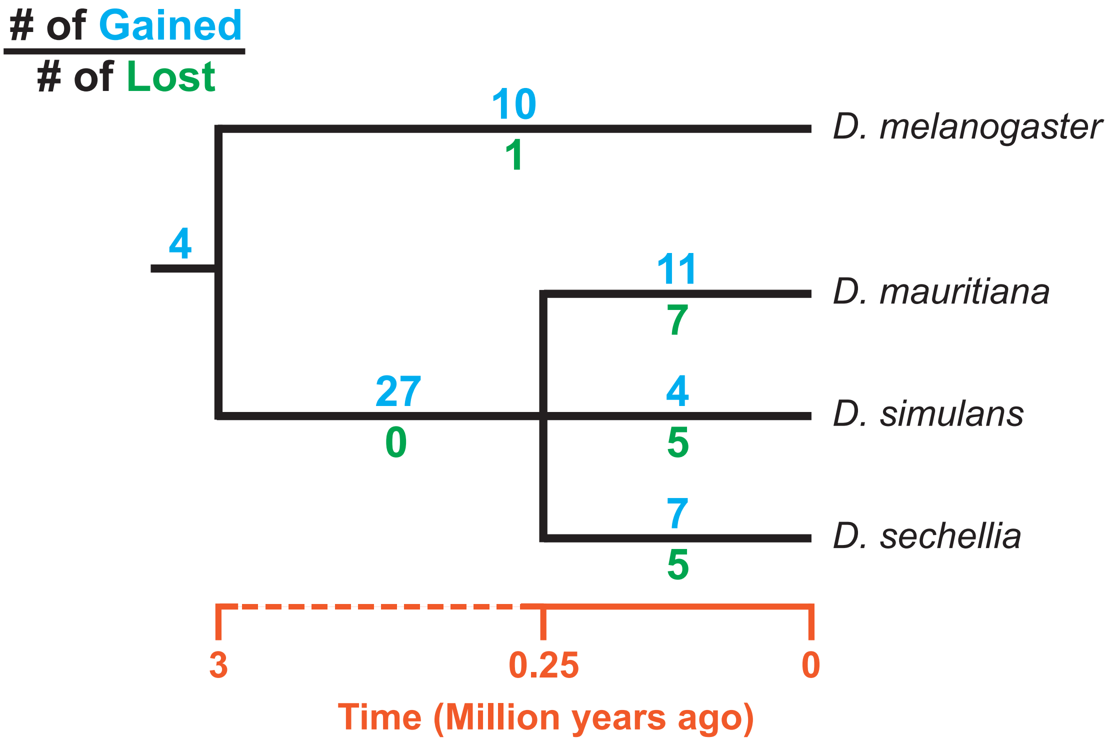


Fig 4. The turnover of new duplications to Y chromosomes in *D. melanogaster* and three species in the *D. simulans* clade. Using phylogenetic analyses, we inferred the evolutionary histories of new Y-linked duplications. The blue and green numbers represent the number of independent duplications and deletions observed in each branch, respectively. The deletion events that happened in the ancestor of these four species cannot be inferred without a Y chromosome assembly in the outgroup.

Most new Y-linked genes in *D. melanogaster* and the *D. simulans* clade have presumed functions in chromatin modification, cell division, and sexual reproduction (Table S7), consistent with other *Drosophila* species [17, 77]. Y-linked duplicates of genes with these functions may be selectively beneficial, but a duplication bias could also contribute to this enrichment, as genes expressed in the testes may be more likely to duplicate to the Y chromosome due to its open chromatin structure and transcriptional activity during spermatogenesis [78-80].

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

The evolution of new Y-linked gene families

Ampliconic gene families are found on Y chromosomes in multiple *Drosophila* species [24]. We discovered two new gene families that have undergone extensive amplification on *D. simulans* clade Y chromosomes. Both families appear to encode functional protein-coding genes with complete open reading frames and high expression in mRNA-seq data (Table S8), and have 36–146 copies in each species' Y chromosome. We also confirm that >90% of the variants in our assembled Y-linked gene families are represented in Illumina DNA-seq data (Supplemental text).

The first amplified Y-linked gene family, *SR Protein Kinase* (*SRPK*), is derived from an autosome-to-Y duplication of the sequence encoding the testis-specific isoform of the gene *SR Protein Kinase (SRPK)*. After the duplication of *SRPK* to the Y chromosome, the ancestral autosomal copy subsequently lost its testis-specific exon via a deletion (Figure 5A). The movement of the male-specific isoform inspired us to name the Y-linked *SRPK* gene family *Lo-han-kha (Lhk)*, which is the Taiwanese term for the male vagabonds that moved from mainland China to Taiwan during the Qing dynasty. In *D. melanogaster, SRPK* is essential for both male and female reproduction [81], suggesting the hypothesis that the relocation of the testis-specific isoform to the *D. simulans* clade Y chromosomes may have relieved intralocus sexual antagonism over these two functions. Our phylogenetic analysis identified two subfamilies of *Lhk* that we designate *Lhk-1* and *Lhk-2* (Figure 5B). Both subfamilies are shared by all *D. simulans* clade species and show a 5.5% protein divergence between species. The two subfamilies are found in different locations in our Y chromosome assemblies; consistent

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

with this observation, we detect two to three *Lhk* foci on Y chromosomes in the *D. simulans* clade using FISH (Figure 5B and 5D and Fig S3 and S4).

The second amplified gene family comprises both X-linked and Y-linked duplicates of the *Ssl* gene located on chromosome 2R; it is unclear whether the X- or Y-linked copies originated first. The X-linked copies are known as *CK2ßtes-like* in *D. simulans* [82]. The Y-linked copies are also found in *D. melanogaster*, but are degenerated and have little or no expression [22, 83], leading to their designation as pseudogenes. In the *D. simulans* clade species, however, the Y-linked paralogs have high levels of expression (> 50 TPM in testes, Table S8) and complete open reading frames, so we refer to this gene family as *CK2ßtes-Y*. Both *CK2ßtes-like* (4–9 copies) and *CK2ßtes-Y* (36–123 copies based on the assemblies) are amplified on the X and Y chromosome in the *D. simulans* clade relative to *D. melanogaster* (Table S8) [82]. The Y-linked copies in *D. melanogaster, Su(Ste),* are known to be a source of piRNAs [84]. We did not detect any testis piRNAs from either gene family in two small RNA-seq datasets (SRR7410589 and SRR7410590), however, we do find some short (< 23-nt) reads (0.003–0.005% of total mapped reads) mapped to these gene families (Table S9).

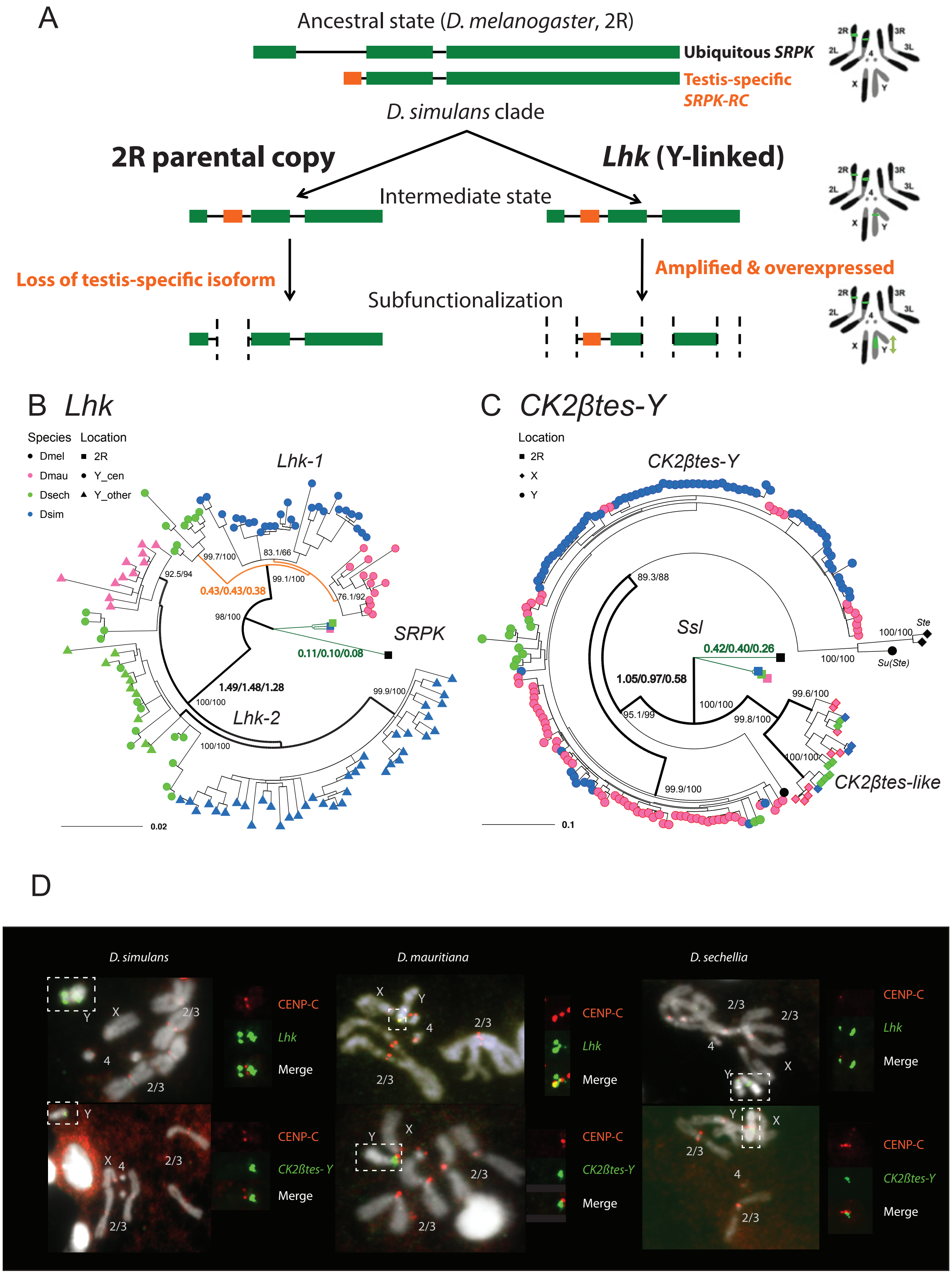
We inferred gene conversion rates and the strength of selection on these Y-linked gene families using phylogenetic analyses on coding sequences. We estimated the gene conversion rate in *D. simulans* clade Y-linked gene families based on four-gamete tests and gene similarity [15, 22, 85, 86]. In general, *D. simulans* clade species show similar gene conversion rates (on the order of 10^{-4} to 10^{-6}) in both of these families compared to our previous estimates in *D. melanogaster* (Table S10; [22]). These higher gene

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

conversion rates compared to the other chromosomes might be a shared feature of Y chromosomes across taxa [15].

To estimate rates of molecular evolution, we conducted branch-model and branch-site-model tests on the reconstructed ancestral sequences of *Lhk-1*, *Lhk-2*, *CK2ßtes-Y,* and two *CK2ßtes-like* using PAML (Fig 5B and 5C; [87]). We used reconstructed ancestral sequences for our analyses to avoid sequencing errors in the assemblies, which appear as singletons. We infer that after the divergence of *D. simulans* clade species, *Lhk-1* evolved under purifying selection, whereas *Lhk-2* evolved under positive selection (Fig 5B; Fig S9; Table S11). Using transcriptome data, we observe that highly expressed *Lhk-1* copies have fewer nonsynonymous mutations than lowly expressed copies in *D. simulans*, consistent with purifying selection (Chi-square test's P=0.01; Fig S10 and Table S12). Both *Lhk* gene families are expressed 2 to 7-fold higher than the ancestral copy on 2R in the same species, and 1.9 to 64-fold higher than their ortholog, *SRPK-RC,* in *D. melanogaster*, suggesting that gene amplification may confer increased expression. In both *D. simulans* and *D. mauritiana*, *Lhk-1* is shorter due to deletions following its origin and has a higher expression level than *Lhk-2*. Both *Lhk* gene families have higher copy numbers in *D. simulans* than *D. mauritiana,* which likely contributes to their higher expression level in *D. simulans* (Table S8). For both *Lhk-1* and *Lhk-2,* copies from the same species are more similar than copies from other species—a signal of concerted evolution [88].

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

Fig 5. The rapid evolution and gene conversion of Y-linked ampliconic genes. A) Schematic showing the inferred evolutionary history of *SRPK-Y*. *SRPK* duplicated to the ancestral Y chromosome in the *D. simulans* clade. The Y-linked copy (*Lhk*) retained an exon with testis-specific expression, which was lost in the parental copy on 2R. The Y-linked copy (*Lhk*) further duplicated and increased their expression in testes. B) The inferred maximum likelihood phylogeny for *Lhk.* Node labels indicate SH-aLRT and ultrafast bootstrap (*e.g.* 100/100) or rates of protein evolution from PAML with CodonFreq = 0,1, or 2 (*e.g.* 1.01/1.02/1.03) (Fig S9 and S11). *Lhk* shows evidence for positive selection (branch tests and branch-site tests with ω>1) after the duplication from 2R (*SRPK*) to the Y chromosome in the *D. simulans* clade. One *Lhk* subfamily (*Lhk-1*) is under recent purifying selection and is located close to the centromere, but the other (*Lhk-2*) is rapidly evolving across the species of the *D. simulans* clade. C) Same as B but for *CK2ßtes-Y.* Both Y-linked *CK2ßtes-Y* and X-linked *CK2ßtes-like* also show positive selection. All ω values shown are statistically significant (LRT tests, P≤0.05; Table S11 and S12). D) On the Y chromosomes, *Lhk* FISH signals are located in 2–3 cytological locations. *CK2ßtes-Y* signals are only located nearby centromeres in the immunolabelling with fluorescent in situ hybridization (immunoFISH) experiments. Based on our analysis of sequence information, we suggest that most *Lhk-1* copies are located close to *CK2ßtes-Y* and centromere.

Table 2. PAML analyses reveal positive selection on Y-linked ampliconic gene families

	Branch test with CodonFreq=0						Branch-site test site class						
Lhk	ω1	ω2	ω3	L	2ΔlnL	LRT's P	ω0	ω1	ω2a	ω2b	2ΔlnL	LRT's P	Positively selected sites (BEB > 0.95)[d]
one ω	0.17			-3250.74									
two ω[a]	0.11	1.05		-3218.26	64.94	7.71E-16	0.01	1	4.87	4.87	13.04	3.05E-04	I4, H11, V32, V75, N99, Y100, D193, D199
three ω[c]	0.11	1.49	0.43	-3216.30	3.92	0.05							
CK2ßtes													
one ω	0.35			-3295.01									
two ω[b]	0.25	1.05		-3272.00	46.01	1.18E-11	0.05	1	2.21	2.21	6.54	1.06E-02	D33, T38, K44, K100, F101, K104, M152, M155
three ω[c]	0.20	0.42	1.05	-3266.33	11.35	7.56E-04							

[a] Autosomal and Y lineage have protein evolution of ω1 and ω2, respectively.

[b] Autosomal and sex chromosomal (X and Y) have protein evolution of ω1 and ω2, respectively.

[c] See Figure 3C and 3D for the assignment of lineage.

[d] See Table S11 and S13 for all the sites.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

The ancestral *Ssl* gene experienced a slightly increased rate of protein evolution after it duplicated to the X and Y chromosomes (ω= 0.41 vs. 0.23; P = 0.03; Fig 5C; Fig S11; Table S13). We find that both *CK2ßtes-like* and *CK2ßtes-Y* share strong signals of positive selection, based on branch-model and branch-site-model tests (P = 8.8E-9; Fig 5C; Fig S11; Table S13). In *D. melanogaster,* the overexpression of the *CK2ßtes-like* X-linked homolog, *Stellate,* can drive in the male germline by killing Y-bearing sperm and generating female-biased offspring [89-91]. We suspect that *CK2ßtes-like* and *CK2ßtes-Y* might have similar functions and may also have a history of conflict. Therefore, the co-amplification of sex-linked genes and positive selection on their coding sequences may be a consequence of an arms race between sex chromosome drivers.

Y chromosome evolution driven by specific mutation patterns

The specific DNA-repair mechanisms used on Y chromosomes might contribute to their high rates of intrachromosomal duplication and structural rearrangements. Because Y chromosomes lack a homolog, they must repair double-strand breaks (DSBs) by non-homologous end joining (NHEJ) or microhomology-mediated end joining (MMEJ), which relies on short homology (usually > 2 bp) to repair DSBs [92]. Compared to NHEJ, MMEJ is more error-prone and can result in translocations and duplications [93]. Preferential use of MMEJ instead of NHEJ could contribute to the high duplication rate and extensive genome rearrangements that we observed on Y chromosomes. To infer the mechanisms of DSB repair on Y chromosomes, we counted indels between Y-linked duplicates and their parent genes for a set of 17 putative pseudogenes—both NHEJ and MMEJ can generate indels, but NHEJ usually produces smaller indels (1–3 bp) compared to MMEJ (> 3 bp) [93, 94]. We also cataloged short stretches of homology

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

between each duplicate and its parent. To compare Y-linked patterns of DSB repair to other regions of the genome, we measured the size of polymorphic indels in intergenic regions and pseudogenes on the autosomes and X chromosomes from population data in *D. melanogaster* (DGRP [95]) and *D. simulans* [96]. To the extent that these indels do not experience selection, their sizes should reflect the mutation patterns on each chromosome. We observe proportionally more large deletions on Y chromosomes (25% of Y-linked indels are ≥10-bp deletions; Table S14) compared to other chromosomes in both *D. melanogaster* (12.8% and 15.2% of indels are ≥10-bp deletions in intergenic regions and pseudogenes) and *D. simulans* (7.3% of indels are ≥10-bp deletions in intergenic regions; all pairwise chi-square's P < 1e-6; Fig 4A; Table S15). The pattern of excess large deletions is shared in the three *D. simulans* clade species Y chromosomes, but is not obvious in *D. melanogaster* (Fig 6B). However, because all *D. melanogaster* Y-linked indels in our analyses are from copies of a single pseudogene (*CR43975*), it is difficult to compare to the larger samples in the simulans clade species (duplicates from 16 genes). The differences in deletion sizes between the Y and other chromosomes are unlikely to be driven by heterochromatin or the lack of recombination —the non-recombining and heterochromatic dot chromosome has a deletion size profile more similar to the other autosomes in *D. simulans* (10.9% of indels are ≥10-bp deletions). These results suggest that Y chromosomes may use MMEJ over NHEJ compared to other chromosomes, particularly in the simulans clade species. We also find that across the genome larger deletions (>7bp) share a similar length of microhomologies for repairing DSBs (39.5–57% deletions have ≥ 2 bp microhomology; Chi-square test for microhomology length between Y and other chromosomes, P > 0.24;

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

Table S14 and S15), consistent with most being a consequence of MMEJ-mediated repair.

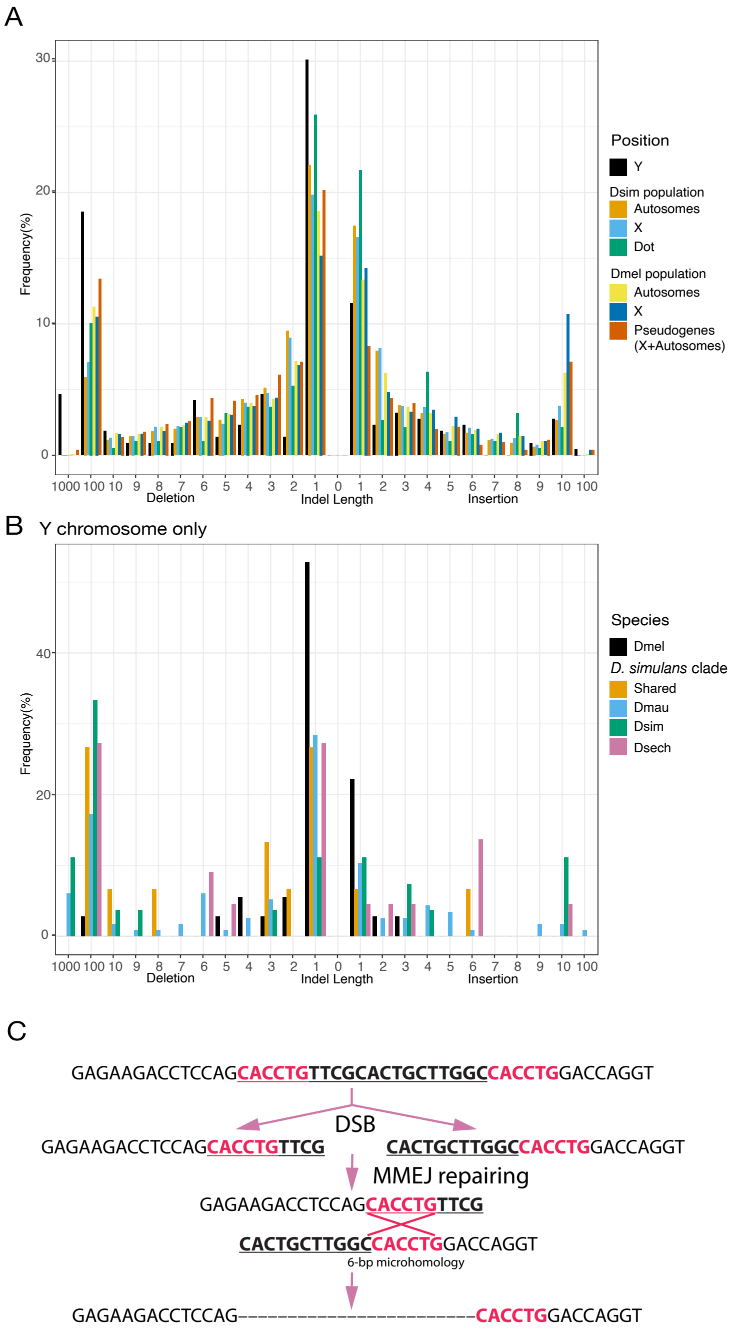


Fig 6. An excess of large deletions on Y chromosomes, compared to population data suggests a preference for MMEJ. A) We compared the size of 216 indels on 17 recently duplicated Y-linked genes in *D. melanogaster* and the *D. simulans* clade species to the indels polymorphic in the *D. melanogaster* and *D. simulans* populations.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

For the indels in *D. melanogaster* and *D. simulans* populations, we separated them based on their location, including autosomes (excluding dot chromosomes), X chromosomes, and dot chromosomes. We excluded the *D. melanogaster* dot-linked indels due to the small sample size (12). B) We classify Y-linked indels by whether they are shared between species or specific in one species C) The excess of large deletions (underlined) on the Y chromosomes is consistent with MMEJ between short regions of microhomology (red).

The satellite sequence composition of Y chromosomes differs between species [76, 97, 98]. A high duplication rate may accelerate the birth and turnover of Y-linked satellite sequences. We discovered five new Y-linked satellites in our assemblies and validated their location using FISH (Fig S3–4 and Table S16). These satellites only span a few kilobases of sequences (5,515 to 26,119 bp) and are homogenized. According to its flanking sequence, one new satellite, $(AAACAT)_n$, originated from a DM412B transposable element, which has three tandem copies of AAACAT in its long terminal repeats. The AAACAT repeats expanded to 764 copies on the Y chromosome specifically in *D. mauritiana*. The other four novel satellites are flanked by transposons (< 50 bp) and may derive from non-repetitive sequences. The MMEJ pathway may contribute to the birth of new repeats, as this mechanism is known to generate tandem duplications via template-switching during repair [93]. Short tandem repeats can be further amplified via saltatory replication or unequal crossing-over between sister chromatids.

Consistent with findings in other species [19, 22], we find an enrichment of LTR retrotransposons on the *D. simulans* clade Y chromosomes relative to the rest of the genome (Table S17). Interestingly, we find that the Y-linked LTR retrotransposons also turn over between species (Fig S12 and Table S18). We find a positive correlation

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

between the difference in Y-linked TE abundance between *D. melanogaster* and each of the *D. simulans* clade species versus the rest of the genome (*rho* = 0.45–0.50; Fig S13 and Table S18). This suggests that global changes in transposon activity could explain the differences in Y-linked TEs abundance between species. However, the correlations between species within the *D. simulans* clade are weaker (*rho* < 0.23; Fig S13 and Table S18), consistent with the possibility that some TEs may shift their insertion preference between chromosomes. To test this hypothesis, we estimated the ages of LTR retrotransposons by their length. We find that the recent insertions of LTR transposons are differently distributed across chromosomes between species (Fig S14), suggesting that insertion preferences towards genomic regions may differ for some TEs. For example, we detect many recent DIVER element insertions on the Y chromosome in *D. simulans*, but not in *D. sechellia* (Fig S9).

Discussion

Despite their independent origins, the degenerated Y chromosomes of mammals, fish, and insects have convergently evolved structural features of gene acquisition and amplification, accumulation of repetitive sequences, and gene conversion. Here we consider the mutational processes that contribute to this structure and its consequences for Y chromosome biology. Our assemblies revealed extensive Y chromosome rearrangements between three very closely related *Drosophila* species (Figure 1). These rearrangements may be the consequence of rejoining telomeres after DSBs, as telomere-specific sequences are embedded in non-telomeric regions of *Drosophila* Y chromosomes [58, 99, 100]. We propose that four pieces of evidence suggest DSBs on

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

Y chromosomes may be preferentially repaired using the MMEJ pathway. First, Y-linked sequences are absent from the X chromosome, precluding repair of DSBs by homologous recombination in meiosis. Second, NHEJ on Y chromosomes may be limited because the Ku complex, which is required for NHEJ [94], is excluded from HP1a-rich regions of chromosomes [101]. The Ku complex also binds telomeres and might prevent telomere fusions [102, 103], suggesting that a low concentration of Ku on Y chromosomes could also cause high rates of telomere rejoining. Third, the highly repetitive nature of Y chromosomes may increase the rate of DSB formation, which may also contribute to a higher rate of MMEJ [93, 104]. Fourth, we show that Y chromosomes have high duplication and gene conversion rates, and larger deletion sizes than other genomic regions (Figure 4), consistent with a preference for MMEJ to repair Y-linked DSBs [93].

The exclusion of the Ku complex from heterochromatin could also contribute to an excess of Y-linked duplications we observe in the *D. simulans* clade relative to *D. melanogaster* (Figure 2A and 4). *D. simulans* clade Y chromosomes might harbor relatively more heterochromatin than the *D. melanogaster* Y due to the partial loss of their euchromatic rDNA repeats [57, 61, 62], and *D. simulans* also expresses more heterochromatin-modifying factors, such as *Su(var)*s and *E(var)*s [105], compared to *D. melanogaster*. To explore these hypotheses, the distribution of the Ku complex across chromosomes in the testes of these species should be studied.

If MMEJ is preferentially used to fix DSBs on the Y chromosome, we might expect that the mutations in the MMEJ pathway would preferentially impact Y-bearing sperm. Consistent with this prediction, a previous study showed that male *D. melanogaster* with

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

a deficient MMEJ pathway (*DNApol* mutants) sire female-biased offspring [106]. Moreover, sperm without sex chromosomes that result from X-Y non-disjunction events are not as strongly affected by an MMEJ deficiency as Y-bearing sperm [106], suggesting that sperm with Y chromosomes are more sensitive to defects in MMEJ.

Drosophila Y chromosomes can act as heterochromatin sinks, sequestering heterochromatin marks from pericentromeric regions and suppressing position-effect variegation [54, 107-109]. Therefore, retrotransposons located in heterochromatin might have higher activities in males due to the presence of Y-linked heterochromatin [54, 108], although the genomic distribution of heterochromatin during spermatogenesis is unknown. We find that, like *D. melanogaster* [22], *D. simulans* clade Y chromosomes are enriched for retrotransposons relative to the rest of the genome; however Y chromosomes from even the closely related *D. simulans* clade species harbor distinct retrotransposons (Figure S12 and Table S18), indicating that some TEs may have rapidly shifted their insertion preference. This preference might benefit the TEs because Y-linked TEs might express during spermatogenesis [110]. On the other hand, Y chromosomes can be a significant source of small RNAs that silence repetitive elements during spermatogenesis—*e.g.*, *Su(Ste)* piRNAs in *D. melanogaster* [111, 112]—and thus may also contribute to TE suppression. If Y chromosomes contribute to piRNA or siRNA production (*e.g.*, have piRNA clusters [112, 113]), then the TE insertion preference for the Y chromosome may sometimes be beneficial for the host, as they could provide immunity against active TEs in males. In this sense, Y chromosomes may even act as “TE traps” that incidentally suppress TE activity in the male germline by producing small RNAs.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

Genes may adapt to the Y chromosome after residing there for millions of years [114, 115]. While most genes that move to the Y chromosome quickly degenerate [18, 23], a subset of new Y-linked genes are retained, presumably due to important roles in male fertility or sex chromosome meiotic drive. New Y-linked genes may adapt to this unique genomic environment, evolving structures and regulatory mechanisms that enable optimal expression on the heterochromatic and non-recombining Y chromosome [116]. Here, we describe two new Y-linked ampliconic genes specific to the *D. simulans* clade—*Lhk* and *CK2ßtes-Y*–that show evidence of strong positive evolution and concerted evolution, suggesting that high copy numbers and Y-Y gene conversion are often important for the adaptation of new Y-linked genes.

Many ampliconic genes are taxonomically restricted and are not maintained at high copy numbers over long periods of evolutionary time [14, 17, 20, 24-26]. Some ampliconic gene families are found on both the X and Y chromosomes [24, 89, 117-119]. While we do not know the function of most such co-amplified gene families, the murine example of *Slx/Slxl1* and *Sly* appears to be engaged in an ongoing arms race between the sex chromosomes [117]. We propose that Y-linked gene amplification in the *D. simulans* clade initially occurs due to an arms race and has the added benefit of being preserved by gene conversion.

It is intriguing that the *CK2ßtes-like/CK2ßtes-Y* gene family is homologous to the *Ste/Su(Ste)* system in *D. melanogaster* [82], which is also hypothesized to play a role in sex-chromosome meiotic drive [120]. We speculate that in both the *D. melanogaster* and *D. simulans* clade lineages these gene amplifications have been driven by conflict between the sex chromosomes over transmission through meiosis, but that the conflict

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

involves different molecular mechanisms. In the *CK2ßtes-like/CK2ßtes-Y* system, both X and Y-linked genes are protein-coding genes, which is reminiscent of *Slx*/*Slxl1* and *Sly* which compete for access to the nucleus where they regulate sex-linked gene expression[117, 118]. In contrast, the Y-linked *Su(Ste)* copies in *D. melanogaster* produce small RNAs that suppress the X-linked *Stellate* [84]. We propose that *CK2ßtes-like/CK2ßtes-Y* system in the *D. simulans* clade species may represent the ancestral state because the parental gene *Ssl* is a protein-coding gene. We speculate that systems arising from antagonisms between the sex chromosomes may shift from protein-coding to RNA-based over time because, with RNAi, suppression is maintained at a minimal translation cost.

Distinct Y-linked mutation patterns are described in many species [14-21]. Our analyses provide a link between Y-linked mutation patterns and Y chromosome evolution. While the lack of recombination and male-limited transmission of the Y chromosome reduces the efficacy of selection, the high gene duplication and gene conversion rates may counter these effects and help acquire and maintain new Y-linked genes. The unique Y-linked mutation patterns might be the direct consequence of the heterochromatic environment on sex chromosomes. Therefore, we predict that W chromosomes and non-recombining sex-limited chromosomes (*e.g.*, some B chromosomes), may share similar mutation patterns with Y chromosomes. Indeed, W chromosomes of birds have ampliconic genes and are rich in tandem repeats [86, 121]. However, there seem to be fewer ampliconic gene families on bird W chromosomes compared to Y chromosomes in other animals, suggesting that sexual selection and intragenomic conflict in spermatogenesis are important contributors to Y-linked gene family evolution [122, 123].

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

Materials and Methods

Assembling Y chromosomes using Pacbio reads in *D. simulans* clade

We applied the heterochromatin-sensitive assembling pipeline from [22]. We first extracted 229,464 reads with 2.2-Gbp in *D. mauritiana*, 269,483 reads with 2.3-Gbp in *D. simulans*, and 257,722 reads with 2.6-Gbp in *D. sechellia* using assemblies from [55], respectively. We then assembled these reads using Canu v1.3 and FALCON v0.5.0 combined the parameter tuning method on 2 error rates, eM and eg, in bogart to optimize the assemblies. We first made the Canu assemblies using the parameters "genomeSize=30m stopOnReadQuality=false corMinCoverage=0 corOutCoverage=100 ovlMerSize=31" and "genomeSize=30m stopOnReadQuality=false". For FALCON v0.5.0, we used the parameters "length_cutoff = -1; seed_coverage = 30 or 40; genome_size = 30000000; length_cutoff_pr = 1000". We then picked the assemblies with highest contiguity and completeness without detectable misassemblies from each setting (two Canu settings and one Falcon setting).

After picking the three best assemblies for each species, we tentatively reconciled the assemblies using Quickmerge [124]. We examined and manually curated the merged assemblies. For the *D. mauritiana* assembly, we merged two Canu and one FALCON assemblies, and for our *D. simulans* and *D. sechellia* assemblies, we merged one Canu and one FALCON assemblies independently. We manually curated some conserved Y-linked genes using raw reads and cDNA sequences from NCBI, including *kl-3* of *D. mauritiana*, *kl-3*, *kl-5*, and *PRY* of *D. simulans* and *CCY*, *PRY*, and *Ppr-Y* of *D.*

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

sechellia, due to their low coverage and importance for our phylogenetic analyses. We then merged our heterochromatin restricted assemblies with contigs of the major chromosome arms from [55]. We polished the resulting assemblies once with Quiver using PacBio reads (SMRT Analysis v2.3.0; [125] and ten times with Pilon v1.22 [126] using raw Illumina reads with parameters "--mindepth 3 --minmq 10 --fix bases".

We identified misassemblies and found parts of Y-linked sequences in the contigs from major arms using our female/male coverage assays in *D. sechellia*. We also assembled the total reads (assuming genome size of 180 Mb) and heterochromatin-extracted reads (assuming genome size 40 Mb) using wtdbg v2.4 with parameters "-x rs -t24 -X 100 -e 2" [127] and Flye v2.4.2 [128] with default parameters separately. We polished the resulting wtdbg assemblies with raw Pacbio reads using Flye v2.4.2. We then manually assembled five introns and fixed two misassemblies using sequences from wtdbg whole-genome assemblies (two introns), Flye whole-genome (two introns), and heterochromatin-enriched assemblies (one intron) in *D. sechellia*. We assembled one intron using sequences from wtdbg whole-genome assemblies in *D. simulans*.

We also extracted potential microbial reads (except for *Wolbachia*) that mapped to the *D. sechellia* microbial contigs, and assembled these reads into a 4.5 Mb contig, which represents the whole genome of a *Providencia* species, using Canu v 1.6 (r8426 14520f819a1e5dd221cc16553cf5b5269227b0a3) with parameters "genomeSize=5m useGrid=false stopOnReadQuality=false corMinCoverage=0 corOutCoverage=100". To detect other symbiont-derived sequences in our assemblies, we used Blast v2.7.1+ [129] with blobtools (v1.0; [130]) to search the nt database (parameters "-task megablast -

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

max_target_seqs 1 -max_hsps 1 -evalue 1e-25"). We estimated the Illumina coverage of each contig in males for *D. mauritiana*, *D. simulans* and *D. sechellia*, respectively. We designated and removed contigs homologous to bacteria and fungi in subsequent analyses (Table S19).

Generating DNA-seq from males in the *D. simulans* clade

We extracted DNA from 30 virgin 0-day males using DNeasy Blood & Tissue Kit and diluted it in 100 μL ddH$_2$O. The DNA was then treated with 1 μL 10mg/mL RNaseA (Invitrogen) at 37°C for 1-hr and was re-diluted in 100 μL ddH$_2$O after ethanol precipitation. The size and concentration of DNA were analyzed by gel electrophoresis, Nanodrop, Qubit and Genomic DNA ScreenTape. Finally, we constructed libraries using PCR-free standard Illumina kit and sequenced 125-bp paired-end reads with a 550-bp insert size from the libraries using Hiseq 2500 in UR Genomics Research Center. We deposited the reads in NCBI's SRA under BioProject accession number PRJNA748438.

Identifying Y-linked contigs

To assign contigs to the Y chromosome, we used Illumina reads from male and female PCR-free genomic libraries (except females of *D. mauritiana*) as described in [22]. In short, we mapped the male and female reads separately using BWA (v0.7.15; [131]) and called the coverage of uniquely mapped reads per site with samtools (v1.7; -Q 10 [132]). We further assigned contigs with the median of male-to-female coverage across contigs equal to 0 as Y-linked. We examined the sensitivity and specificity of our methods using all 10-kb regions with known location. Based on our results for 10-kb

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

regions with known location (Table S2) in *D. mauritiana*, we set up an additional criterion for this species—"the average of female-to-male coverage < 0.1"—to reduce the false discovery rate.

Gene and repeat annotations

We used the same pipeline and data to annotate genomes as a previous study [55]. We collected transcripts and translated sequences from *D. melanogaster* (r6.14) and transcript sequences from *D. simulans* [133] using IsoSeq3 [134]. We mapped these sequences to each assembly to generate annotations using maker2 (v2.31.9; [135]. We further mapped the transcriptomes using Star 2.7.3a 2-pass mapping with the maker2 annotation and parameters "-outFilterMultimapNmax 200 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 0.04 --alignIntronMin 20 --alignIntronMax 5000000 --alignMatesGapMax 5000000 --outSAMtype BAM SortedByCoordinate --readFilesCommand zcat --peOverlapNbasesMin 12 --peOverlapMMp 0.1". We then generated the consensus annotations using Stringtie 2.0.3 from all transcriptomes [136]. We further improved the mitochondria annotation using MITOS2. We assigned predicted transcripts to their homologs in *D. melanogaster* using BLAST v2.7.1+ (-evalue 1e-10; [129]).

We used RepeatMasker v4.0.5 [137] with our custom library to annotate the assemblies using parameter "-s." Our custom library is modified from [55], by adding the consensus sequence of *Jockey-3* from *D. melanogaster* to replace its homologs (*G2* in *D.*

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

melanogaster and *Jockey-3* in *D. simulans*; [138]). We extracted the sequences and copies of TEs and other repeats using scripts modified from [139]. To annotate tandem repeats in assemblies, we used TRFinder (v4.09; [140] with parameters “2 7 7 80 10 100 2000 -ngs -h”. We also used kseek to search for tandem repeats in the male Illumina reads.

Transcriptome analyses

We mapped the testes transcriptome to the reference genomes of *D. melanogaster, D. simulans* and *D. mauritiana* (Table S20; no available transcriptome from *D. sechellia*). We used Stringtie 2.0.3 [136] to estimate the expression level using the annotation. However, we applied a different strategy for estimating expression levels of the Y-linked gene families due to the difficulties in precisely annotating multi-copies genes. We constructed a transcript reference using current gene annotation but replaced all transcripts from *Lhk-1, Lhk-2* and *CK2ßtes-Y* with their species-specific reconstructed ancestral copies. We then mapped the transcriptome reads to this reference using Bowtie2 v 2.3.5.1 [141] with parameters “-very-sensitive -p 24 -k 200 -X 1000 --no-discordant --no-mixed”. We then estimated the expression level by salmon v 1.0.0 [142] with parameters “-l A -p 24.” We also mapped small RNA reads from *D. simulans* testes to our custom repeat library and reconstructed ancestral *Lhk-1, Lhk-2* and *CK2ßtes-Y* sequences using Bowtie v 1.2.3 [143] with parameters “-v3 -q -a -m 50 --best –strata.”

To assay the specific expression of different copies, we also mapped transcriptomic and male genomic reads to the same reference using BWA (v0.7.15; [131]. We used ABRA

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

v2.22 [144] to improve the alignments around the indels of these two gene families. We used samtools (v1.7; [132]) to pile up reads that mapped to reconstructed ancestral copies and estimated the frequency of derived SNPs in the reads.

Estimating Y-linked exon copy numbers using Illumina reads

We mapped the Illumina reads from the male individuals of *D. melanogaster* and the *D. simulans* clade species to a genome reference with transcripts of 11 conserved Y-linked genes and the sequences of all non-Y chromosomes (r6.14) in *D. melanogaster*. We called the depth using samtools depth (v1.7; [132]), and estimated the copy number of each exon using the mapped depth. We assumed most Y-linked exons are single-copy, so we divided the depth of each site by the majority of depth across all Y-linked transcripts to estimate the copy number. For the comparison, we simulated the 50X Illumina reads from our assemblies using ART 2.5.8 with the parameter (art_illumina -ss HSXt -m 500 -s 200 -p -l 150 -f 50; [145]). We then mapped the simulated reads to the same reference, called the depth, and divided the depth of each site by 50.

Immunostaining and FISH of mitotic chromosomes

We conducted FISH in brain cells following the protocol from [146] and immunostaining with FISH (immune-FISH) in brain cells following the protocol from [147] and [138]. Briefly, we dissected brains from third instar larva in 1X PBS and treated them for 1-min in hypotonic solution (0.5% sodium citrate). Then, we fixed brain cells in 1.8% paraformaldehyde, 45% acetic acid for 6-min. We subsequently dehydrated in ethanol for the FISH experiments but not for the immune-FISH.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

For immunostaining, we rehydrated the slide using PBS with 0.1% TritonX-100 after removing the coverslip using liquid nitrogen. The slides were blocked with 3% BSA and 1% goat serum/ PBS with 0.1% TritonX-100 for 30-min and hybridized with 1:500 anti-Cenp-C antibody (gift from Dr. Barbara Mellone) overnight at 4°C. We used 1:500 secondary antibodies (Life Technologies Alexa-488, 546, or 647 conjugated, 1:500) in blocking solution with 45-min room temperature incubation to detect the signals. We fixed the slides in 4% paraformaldehyde in 4XSSC for 6-min before doing FISH.

We added probes and denatured the fixed slides at 95°C for 5-min and then hybridized slides at 30°C overnight. For PCR amplified probes with DIG or biotin labels, we blocked the slides for 1-hr using 3% BSA/PBS with 0.1% Tween and incubated slides with 1:200 secondary antibodies (Roche) in 3% BSA/4X SSC with 0.1% Tween and BSA at room temperature for 1 hr. We made *Lhk* and *CK2ßtes-Y* probes using PCR Nick Translation kits (Roche) and ordered oligo probes from IDT. We list probe information in Table S3. We mounted slides in Diamond Antifade Mountant with DAPI (Invitrogen) and visualized them on a Leica DM5500 upright fluorescence microscope, imaged with a Hamamatsu Orca R2 CCD camera and analyzed using Leica's LAX software. We interpreted the binding patterns of Y chromosomes using the density of DAPI staining solely.

Phylogenetic analyses of Y-linked genes

We used BLAST v2.7.1+ [129] to extract the sequences of Y-linked duplications and conserved Y-linked genes from the genome. We only used high-quality sequences

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

polished by Pilon (--mindepth 3 --minmq 10) for our phylogenetic analyses. We aligned and manually inspected sequences with reference transcripts from Flybase using Geneious v8.1.6 [148]. For most Y-linked duplications, except for the genes homologous to *Lhk* and *CK2ßtes-Y*, we constructed neighbor-joining trees using the HKY model with 1,000 replicates using Geneious v8.1.6 [148] to infer their phylogenies. We also measured the length and microhomology in 216 indels from 17 Y-linked duplications using these alignments (Table S14). We also infer the potential mechanisms causing the indels, including tandem duplications and polymerase slippage during DNA replication. We measured the length and microhomology of polymorphic indels in *D. melanogaster* (DGRP [95]) and *D. simulans* [96] populations from [55]. For *Lhk* and *CK2ßtes-Y*, we constructed phylogeny using iqtree 1.6.12 [149, 150] using parameters "-m MFP -nt AUTO -alrt 1000 -bb 1000 -bnni". The node labels in Figure 5 correspond to SH-aLRT support (%) / ultrafast bootstrap support (%). The nodes with SH-aLRT >= 80% and ultrafast bootstrap support >= 95% are strongly supported. Protein evolutionary rates (with CodonFreq = 0/1/2 in PAML) of the bold branches were estimated using PAML with branch models on the reconstructed ancestor sequences (Fig S9 and S11).

Estimating recombination and selection on Y-linked ampliconic genes

Using the phylogenetic trees from iqtree, we infer the most probable sequences for the internal nodes using MEGA 10.1.5 [151, 152] using the maximal likelihood method and G+I model with GTR model. We conducted branch and branch-site models tests in

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

PAML 4.8 using the ancestral sequences of Y-linked and X-linked ampliconic gene families with their homologs on autosomes. We plotted the tree using R package ape 5.3 [153].

We used compute 0.8.4 [154] to calculate Rmin and population recombination rates based on linkage disequilibrium [155, 156] and gene similarity. We included sites with indel polymorphisms in these analyses to increase the sample size (558–1,544 bp alignments). We also reanalyzed data from Chang and Larracuente 2019 [22] to include variant information from these sites. The high similarity between Y-linked ampliconic gene copies may lead us to overestimate gene conversion based on gene similarity [155]. We therefore also reported the lower bound on the gene conversion rate using Rmin [156].

GO term analysis

We used PANTHER (Released 20190711; [157]) with GO Ontology database (Released 2019-10-08) to perform Biological GO term analysis of new Y-linked duplicated genes using Fisher’s exact tests with FDR correction. We input 70 duplicated genes with any known GO terms and used all genes (13767) in *D. melanogaster* as background.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

Data availability

Genomic DNA sequence reads are in NCBI's SRA under BioProject PRJNA748438. All scripts and pipelines are available in GitHub (forthcoming) and the Dryad digital repository (doi forthcoming).

Acknowledgements

This work was funded by the National Institutes of Health (NIH) (R35GM119515 to A.M.L. and R01GM123194 to C.D.M.), National Science Foundation (NSF MCB 1844693) to A.M.L. and funding from the University of Nebraska-Lincoln to C.D.M.. A.M.L. was supported by a Stephen Biggar and Elisabeth Asaro fellowship in Data Science. C.-H.C. was supported by the Messersmith Fellowship from the U of Rochester, the Government Scholarship to Study Abroad from Taiwan, and the Damon Runyon fellowship (DRG: 2438-21). We thank our collaborators, Drs. J.J. Emerson and Mahul Chakraborty, for generating PacBio reads in the *D. simulans* clade, Dr. Barbara Mellone for the antibodies, and Drs. Casey Bergman, Grace YC Lee, Kevin Wei and Anthony Geneva and Larraucente lab members for helpful discussion. We also thank the U of Rochester CIRC for access to computing cluster resources and UR Genomics Research Center for the library construction and sequencing.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

References

1. Rice WR. The Accumulation of Sexually Antagonistic Genes as a Selective Agent Promoting the Evolution of Reduced Recombination between Primitive Sex-Chromosomes. Evolution. 1987;41(4):911-4. doi: Doi 10.2307/2408899. PubMed PMID: WOS:A1987J007200019.

2. Bachtrog D. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nature reviews Genetics. 2013;14(2):113-24. Epub 2013/01/19. doi: 10.1038/nrg3366. PubMed PMID: 23329112; PubMed Central PMCID: PMCPMC4120474.

3. Charlesworth B. Model for evolution of Y chromosomes and dosage compensation. Proceedings of the National Academy of Sciences of the United States of America. 1978;75(11):5618-22. Epub 1978/11/01. PubMed PMID: 281711; PubMed Central PMCID: PMCPMC393018.

4. Rice WR. Genetic hitchhiking and the evolution of reduced genetic activity of the Y sex chromosome. Genetics. 1987;116(1):161-7. Epub 1987/05/01. PubMed PMID: 3596229; PubMed Central PMCID: PMCPMC1203114.

5. Charlesworth D, Charlesworth B, Morgan MT. The pattern of neutral molecular variation under the background selection model. Genetics. 1995;141(4):1619-32. Epub 1995/12/01. PubMed PMID: 8601499; PubMed Central PMCID: PMCPMC1206892.

6. Charlesworth B, Charlesworth D. The degeneration of Y chromosomes. Philos Trans R Soc Lond B Biol Sci. 2000;355(1403):1563-72. Epub 2000/12/29. doi: 10.1098/rstb.2000.0717. PubMed PMID: 11127901; PubMed Central PMCID: PMCPMC1692900.

7. Bergero R, Qiu S, Charlesworth D. Gene loss from a plant sex chromosome system. Curr Biol. 2015;25(9):1234-40. Epub 2015/04/29. doi: 10.1016/j.cub.2015.03.015. PubMed PMID: 25913399.

8. Bergero R, Gardner J, Bader B, Yong L, Charlesworth D. Exaggerated heterochiasmy in a fish with sex-linked male coloration polymorphisms. Proceedings of the National Academy of Sciences of the United States of America. 2019;116(14):6924-31. Epub 2019/03/22. doi: 10.1073/pnas.1818486116. PubMed PMID: 30894479; PubMed Central PMCID: PMCPMC6452659.

9. Lenormand T, Fyon F, Sun E, Roze D. Sex Chromosome Degeneration by Regulatory Evolution. Curr Biol. 2020;30(15):3001-6 e5. Epub 2020/06/20. doi: 10.1016/j.cub.2020.05.052. PubMed PMID: 32559446.

10. Bachtrog D. Protein evolution and codon usage bias on the neo-sex chromosomes of Drosophila miranda. Genetics. 2003;165(3):1221-32. Epub 2003/12/12. PubMed PMID: 14668377; PubMed Central PMCID: PMCPMC1462847.

11. Singh ND, Koerich LB, Carvalho AB, Clark AG. Positive and purifying selection on the Drosophila Y chromosome. Mol Biol Evol. 2014;31(10):2612-23. Epub 2014/06/30. doi: 10.1093/molbev/msu203. PubMed PMID: 24974375; PubMed Central PMCID: PMCPMC4166921.

12. Larracuente AM, Clark AG. Surprising differences in the variability of Y chromosomes in African and cosmopolitan populations of Drosophila melanogaster. Genetics. 2013;193(1):201-14. Epub 2012/10/23. doi: 10.1534/genetics.112.146167. PubMed PMID: 23086221; PubMed Central PMCID: PMCPMC3527246.

13. Bachtrog D. Evidence that positive selection drives Y-chromosome degeneration in Drosophila miranda. Nat Genet. 2004;36(5):518-22. Epub 2004/04/27. doi: 10.1038/ng1347. PubMed PMID: 15107853.

14. Soh YQ, Alfoldi J, Pyntikova T, Brown LG, Graves T, Minx PJ, et al. Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

chromosomes. Cell. 2014;159(4):800-13. Epub 2014/11/25. doi: 10.1016/j.cell.2014.09.052. PubMed PMID: 25417157; PubMed Central PMCID: PMCPMC4260969.

15. Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, et al. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. Nature. 2003;423(6942):873-6. Epub 2003/06/20. doi: 10.1038/nature01723. PubMed PMID: 12815433.

16. Hughes JF, Page DC. The Biology and Evolution of Mammalian Y Chromosomes. Annual review of genetics. 2015;49:507-27. Epub 2015/10/08. doi: 10.1146/annurev-genet-112414-055311. PubMed PMID: 26442847.

17. Bachtrog D, Mahajan S, Bracewell R. Massive gene amplification on a recently formed Drosophila Y chromosome. Nat Ecol Evol. 2019;3(11):1587-97. Epub 2019/11/02. doi: 10.1038/s41559-019-1009-9. PubMed PMID: 31666742.

18. Tobler R, Nolte V, Schlotterer C. High rate of translocation-based gene birth on the Drosophila Y chromosome. Proceedings of the National Academy of Sciences of the United States of America. 2017;114(44):11721-6. Epub 2017/10/29. doi: 10.1073/pnas.1706502114. PubMed PMID: 29078298; PubMed Central PMCID: PMCPMC5676891.

19. Peichel CL, McCann SR, Ross JA, Naftaly AFS, Urton JR, Cech JN, et al. Assembly of a young vertebrate Y chromosome reveals convergent signatures of sex chromosome evolution. bioRxiv. 2019:2019.12.12.874701. doi: 10.1101/2019.12.12.874701.

20. Brashear WA, Raudsepp T, Murphy WJ. Evolutionary conservation of Y Chromosome ampliconic gene families despite extensive structural variation. Genome research. 2018;28(12):1841-51. Epub 2018/11/02. doi: 10.1101/gr.237586.118. PubMed PMID: 30381290; PubMed Central PMCID: PMCPMC6280758.

21. Hall AB, Papathanos PA, Sharma A, Cheng C, Akbari OS, Assour L, et al. Radical remodeling of the Y chromosome in a recent radiation of malaria mosquitoes. Proceedings of the National Academy of Sciences of the United States of America. 2016;113(15):E2114-23. Epub 2016/04/02. doi: 10.1073/pnas.1525164113. PubMed PMID: 27035980; PubMed Central PMCID: PMCPMC4839409.

22. Chang CH, Larracuente AM. Heterochromatin-Enriched Assemblies Reveal the Sequence and Organization of the Drosophila melanogaster Y Chromosome. Genetics. 2019;211(1):333-48. Epub 2018/11/14. doi: 10.1534/genetics.118.301765. PubMed PMID: 30420487; PubMed Central PMCID: PMCPMC6325706.

23. Carvalho AB, Vicoso B, Russo CA, Swenor B, Clark AG. Birth of a new gene on the Y chromosome of Drosophila melanogaster. Proceedings of the National Academy of Sciences of the United States of America. 2015;112(40):12450-5. Epub 2015/09/20. doi: 10.1073/pnas.1516543112. PubMed PMID: 26385968; PubMed Central PMCID: PMCPMC4603513.

24. Ellison C, Bachtrog D. Recurrent gene co-amplification on Drosophila X and Y chromosomes. PLoS Genet. 2019;15(7):e1008251. Epub 2019/07/23. doi: 10.1371/journal.pgen.1008251. PubMed PMID: 31329593; PubMed Central PMCID: PMCPMC6690552.

25. Hughes JF, Skaletsky H, Pyntikova T, Graves TA, van Daalen SK, Minx PJ, et al. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. Nature. 2010;463(7280):536-9. Epub 2010/01/15. doi: 10.1038/nature08700. PubMed PMID: 20072128; PubMed Central PMCID: PMCPMC3653425.

26. Mueller JL, Mahadevaiah SK, Park PJ, Warburton PE, Page DC, Turner JM. The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression. Nat Genet. 2008;40(6):794-9. Epub 2008/05/06. doi: 10.1038/ng.126. PubMed PMID: 18454149; PubMed Central PMCID: PMCPMC2740655.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

27. Connallon T, Clark AG. Gene duplication, gene conversion and the evolution of the Y chromosome. Genetics. 2010;186(1):277-86. doi: 10.1534/genetics.110.116756. PubMed PMID: 20551442; PubMed Central PMCID: PMCPMC2940292.

28. Carlson M, Brutlag D. Cloning and characterization of a complex satellite DNA from Drosophila melanogaster. Cell. 1977;11(2):371-81. Epub 1977/06/01. PubMed PMID: 408008.

29. Lohe AR, Brutlag DL. Identical satellite DNA sequences in sibling species of Drosophila. J Mol Biol. 1987;194(2):161-70. Epub 1987/03/20. PubMed PMID: 3112413.

30. Lohe AR, Brutlag DL. Adjacent satellite DNA segments in Drosophila structure of junctions. J Mol Biol. 1987;194(2):171-9. Epub 1987/03/20. PubMed PMID: 3112414.

31. Mahajan S, Wei KH, Nalley MJ, Gibilisco L, Bachtrog D. De novo assembly of a young Drosophila Y chromosome using single-molecule sequencing and chromatin conformation capture. PLoS Biol. 2018;16(7):e2006348. Epub 2018/07/31. doi: 10.1371/journal.pbio.2006348. PubMed PMID: 30059545; PubMed Central PMCID: PMCPMC6117089.

32. Araripe LO, Tao Y, Lemos B. Interspecific Y chromosome variation is sufficient to rescue hybrid male sterility and is influenced by the grandparental origin of the chromosomes. Heredity (Edinb). 2016;116(6):516-22. Epub 2016/03/17. doi: 10.1038/hdy.2016.11. PubMed PMID: 26980343; PubMed Central PMCID: PMCPMC4868264.

33. Bayes JJ, Malik HS. Altered heterochromatin binding by a hybrid sterility protein in Drosophila sibling species. Science. 2009;326(5959):1538-41. Epub 2009/11/26. doi: 10.1126/science.1181756. PubMed PMID: 19933102; PubMed Central PMCID: PMCPMC2987944.

34. Johnson NA, Perez DE, Cabot EL, Hollocher H, Wu CI. A test of reciprocal X-Y interactions as a cause of hybrid sterility in Drosophila. Nature. 1992;358(6389):751-3. Epub 1992/08/27. doi: 10.1038/358751a0. PubMed PMID: 1508270.

35. Coyne JA. The genetic basis of Haldane's rule. Nature. 1985;314(6013):736-8. Epub 1985/04/01. doi: 10.1038/314736a0. PubMed PMID: 3921852.

36. Bozzetti MP, Massari S, Finelli P, Meggio F, Pinna LA, Boldyreff B, et al. The Ste locus, a component of the parasitic cry-Ste system of Drosophila melanogaster, encodes a protein that forms crystals in primary spermatocytes and mimics properties of the beta subunit of casein kinase 2. Proceedings of the National Academy of Sciences of the United States of America. 1995;92(13):6067-71. Epub 1995/06/20. PubMed PMID: 7597082; PubMed Central PMCID: PMCPMC41643.

37. Courret C, Chang CH, Wei KH, Montchamp-Moreau C, Larracuente AM. Meiotic drive mechanisms: lessons from Drosophila. Proc Biol Sci. 2019;286(1913):20191430. Epub 2019/10/24. doi: 10.1098/rspb.2019.1430. PubMed PMID: 31640520; PubMed Central PMCID: PMCPMC6834043.

38. Tao Y, Araripe L, Kingan SB, Ke Y, Xiao H, Hartl DL. A sex-ratio meiotic drive system in Drosophila simulans. II: an X-linked distorter. PLoS Biol. 2007;5(11):e293. Epub 2007/11/09. doi: 10.1371/journal.pbio.0050293. PubMed PMID: 17988173; PubMed Central PMCID: PMCPMC2062476.

39. Tao Y, Hartl DL, Laurie CC. Sex-ratio segregation distortion associated with reproductive isolation in Drosophila. Proceedings of the National Academy of Sciences of the United States of America. 2001;98(23):13183-8. Epub 2001/11/01. doi: 10.1073/pnas.231478798. PubMed PMID: 11687638; PubMed Central PMCID: PMCPMC60845.

40. Helleu Q, Courret C, Ogereau D, Burnham KL, Chaminade N, Chakir M, et al. Sex-Ratio Meiotic Drive Shapes the Evolution of the Y Chromosome in Drosophila simulans. Mol Biol Evol. 2019;36(12):2668-81. Epub 2019/07/11. doi: 10.1093/molbev/msz160. PubMed PMID: 31290972.

41. Branco AT, Tao Y, Hartl DL, Lemos B. Natural variation of the Y chromosome suppresses sex ratio distortion and modulates testis-specific gene expression in Drosophila

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

simulans. Heredity (Edinb). 2013;111(1):8-15. Epub 2013/04/18. doi: 10.1038/hdy.2013.5. PubMed PMID: 23591516; PubMed Central PMCID: PMCPMC3692315.

42. Montchamp-Moreau C, Ginhoux V, Atlan A. The Y chromosomes of Drosophila simulans are highly polymorphic for their ability to suppress sex-ratio drive. Evolution. 2001;55(4):728-37. Epub 2001/06/08. PubMed PMID: 11392391.

43. Meiklejohn CD, Landeen EL, Gordon KE, Rzatkiewicz T, Kingan SB, Geneva AJ, et al. Gene flow mediates the role of sex chromosome meiotic drive during complex speciation. Elife. 2018;7. Epub 2018/12/14. doi: 10.7554/eLife.35468. PubMed PMID: 30543325; PubMed Central PMCID: PMCPMC6292695.

44. Reijo R, Lee TY, Salo P, Alagappan R, Brown LG, Rosenberg M, et al. Diverse spermatogenic defects in humans caused by Y chromosome deletions encompassing a novel RNA-binding protein gene. Nat Genet. 1995;10(4):383-93. Epub 1995/08/01. doi: 10.1038/ng0895-383. PubMed PMID: 7670487.

45. Vogt PH, Edelmann A, Kirsch S, Henegariu O, Hirschmann P, Kiesewetter F, et al. Human Y chromosome azoospermia factors (AZF) mapped to different subregions in Yq11. Hum Mol Genet. 1996;5(7):933-43. Epub 1996/07/01. PubMed PMID: 8817327.

46. Sun C, Skaletsky H, Rozen S, Gromoll J, Nieschlag E, Oates R, et al. Deletion of azoospermia factor a (AZFa) region of human Y chromosome caused by recombination between HERV15 proviruses. Hum Mol Genet. 2000;9(15):2291-6. Epub 2000/09/26. PubMed PMID: 11001932.

47. Repping S, Skaletsky H, Brown L, van Daalen SK, Korver CM, Pyntikova T, et al. Polymorphism for a 1.6-Mb deletion of the human Y chromosome persists through balance between recurrent mutation and haploid selection. Nat Genet. 2003;35(3):247-51. Epub 2003/10/07. doi: 10.1038/ng1250. PubMed PMID: 14528305.

48. Morgan AP, Pardo-Manuel de Villena F. Sequence and Structural Diversity of Mouse Y Chromosomes. Mol Biol Evol. 2017;34(12):3186-204. Epub 2017/10/14. doi: 10.1093/molbev/msx250. PubMed PMID: 29029271.

49. Lemos B, Branco AT, Hartl DL. Epigenetic effects of polymorphic Y chromosomes modulate chromatin components, immune response, and sexual conflict. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(36):15826-31. Epub 2010/08/28. doi: 10.1073/pnas.1010383107. PubMed PMID: 20798037; PubMed Central PMCID: PMCPMC2936610.

50. Wang M, Branco AT, Lemos B. The Y Chromosome Modulates Splicing and Sex-Biased Intron Retention Rates in Drosophila. Genetics. 2017. Epub 2017/12/22. doi: 10.1534/genetics.117.300637. PubMed PMID: 29263027.

51. Sackton TB, Montenegro H, Hartl DL, Lemos B. Interspecific Y chromosome introgressions disrupt testis-specific gene expression and male reproductive phenotypes in Drosophila. Proceedings of the National Academy of Sciences of the United States of America. 2011;108(41):17046-51. Epub 2011/10/05. doi: 10.1073/pnas.1114690108. PubMed PMID: 21969588; PubMed Central PMCID: PMCPMC3193250.

52. Zhou J, Sackton TB, Martinsen L, Lemos B, Eickbush TH, Hartl DL. Y chromosome mediates ribosomal DNA silencing and modulates the chromatin state in Drosophila. Proceedings of the National Academy of Sciences of the United States of America. 2012;109(25):9941-6. Epub 2012/06/06. doi: 10.1073/pnas.1207367109. PubMed PMID: 22665801; PubMed Central PMCID: PMCPMC3382510.

53. Case LK, Wall EH, Osmanski EE, Dragon JA, Saligrama N, Zachary JF, et al. 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. 2015;16:28. Epub 2015/04/19. doi: 10.1186/s13059-015-0591-7. PubMed PMID: 25886764; PubMed Central PMCID: PMCPMC4396973.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

54. Brown E, Bachtrog D. The Drosophila Y chromosome affects heterochromatin integrity genome-wide. bioRxiv. 2017. doi: 10.1101/156000.

55. Chakraborty M, Chang CH, Khost DE, Vedanayagam J, Adrion JR, Liao Y, et al. Evolution of genome structure in the Drosophila simulans species complex. Genome research. 2021;31(3):380-96. Epub 2021/02/11. doi: 10.1101/gr.263442.120. PubMed PMID: 33563718; PubMed Central PMCID: PMCPMC7919458.

56. Lemeunier F, Ashburner M. Relationships within the melanogaster species subgroup of the genus Drosophila (Sophophora). Chromosoma. 1984;89(5):343-51. doi: 10.1007/bf00331251.

57. Roy V, Monti-Dedieu L, Chaminade N, Siljak-Yakovlev S, Aulard S, Lemeunier F, et al. Evolution of the chromosomal location of rDNA genes in two Drosophila species subgroups: ananassae and melanogaster. Heredity (Edinb). 2005;94(4):388-95. Epub 2005/02/24. doi: 10.1038/sj.hdy.6800612. PubMed PMID: 15726113.

58. Berloco M, Fanti L, Sheen F, Levis RW, Pimpinelli S. Heterochromatic distribution of HeT-A- and TART-like sequences in several Drosophila species. Cytogenetic and genome research. 2005;110(1-4):124-33. Epub 2005/08/12. doi: 10.1159/000084944. PubMed PMID: 16093664.

59. Erhardt S, Mellone BG, Betts CM, Zhang W, Karpen GH, Straight AF. Genome-wide analysis reveals a cell cycle-dependent mechanism controlling centromere propagation. J Cell Biol. 2008;183(5):805-18. Epub 2008/12/03. doi: 10.1083/jcb.200806038. PubMed PMID: 19047461; PubMed Central PMCID: PMCPMC2592830.

60. Paredes S, Branco AT, Hartl DL, Maggert KA, Lemos B. Ribosomal DNA deletions modulate genome-wide gene expression: "rDNA-sensitive" genes and natural variation. PLoS Genet. 2011;7(4):e1001376. Epub 2011/05/03. doi: 10.1371/journal.pgen.1001376. PubMed PMID: 21533076; PubMed Central PMCID: PMCPMC3080856.

61. Lohe AR, Roberts PA. Evolution of DNA in heterochromatin: the Drosophila melanogaster sibling species subgroup as a resource. Genetica. 2000;109(1-2):125-30. Epub 2001/04/11. PubMed PMID: 11293787.

62. Lohe AR, Roberts PA. An unusual Y chromosome of Drosophila simulans carrying amplified rDNA spacer without rRNA genes. Genetics. 1990;125(2):399-406. Epub 1990/06/01. PubMed PMID: 2379820; PubMed Central PMCID: PMCPMC1204028.

63. McKee BD, Karpen GH. Drosophila ribosomal RNA genes function as an X-Y pairing site during male meiosis. Cell. 1990;61(1):61-72. Epub 1990/04/06. PubMed PMID: 2156630.

64. Kopp A, Frank A, Fu J. Historical biogeography of Drosophila simulans based on Y-chromosomal sequences. Mol Phylogenet Evol. 2006;38(2):355-62. Epub 2005/07/30. doi: 10.1016/j.ympev.2005.06.006. PubMed PMID: 16051503.

65. Chakraborty M, Chang C-H, Khost D, Vedanayagam J, Adrion JR, Liao Y, et al. Evolution of genome structure in the Drosophila simulans species complex. bioRxiv. 2020.

66. Bonaccorsi S, Pisano C, Puoti F, Gatti M. Y chromosome loops in Drosophila melanogaster. Genetics. 1988;120(4):1015-34. Epub 1988/12/01. PubMed PMID: 2465201; PubMed Central PMCID: PMCPMC1203565.

67. Bonaccorsi S, Gatti M, Pisano C, Lohe A. Transcription of a satellite DNA on two Y chromosome loops of Drosophila melanogaster. Chromosoma. 1990;99(4):260-6. Epub 1990/08/01. PubMed PMID: 2119983.

68. Meyer GnF. Die Funktionsstrukturen des Y-Chromosoms in den Spermatocytenkernen von Drosophila hydei, D. neohydei, D. repleta und einigen anderen Drosophila-Arten. Chromosoma. 1963;14(3):207-55. doi: 10.1007/bf00326814.

69. Piergentili R. Evolutionary conservation of lampbrush-like loops in drosophilids. BMC Cell Biol. 2007;8:35. Epub 2007/08/19. doi: 10.1186/1471-2121-8-35. PubMed PMID: 17697358; PubMed Central PMCID: PMCPMC1978495.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

70. Fingerhut JM, Moran JV, Yamashita YM. Satellite DNA-containing gigantic introns in a unique gene expression program during Drosophila spermatogenesis. PLoS Genet. 2019;15(5):e1008028. Epub 2019/05/10. doi: 10.1371/journal.pgen.1008028. PubMed PMID: 31071079; PubMed Central PMCID: PMCPMC6508621.

71. Redhouse JL, Mozziconacci J, White RA. Co-transcriptional architecture in a Y loop in Drosophila melanogaster. Chromosoma. 2011;120(4):399-407. Epub 2011/05/11. doi: 10.1007/s00412-011-0321-1. PubMed PMID: 21556802.

72. Pisano C, Bonaccorsi S, Gatti M. The kl-3 loop of the Y chromosome of Drosophila melanogaster binds a tektin-like protein. Genetics. 1993;133(3):569-79. Epub 1993/03/01. PubMed PMID: 8454204; PubMed Central PMCID: PMCPMC1205344.

73. Piergentili R, Bonaccorsi S, Raffa GD, Pisano C, Hackstein JH, Mencarelli C. Autosomal control of the Y-chromosome kl-3 loop of Drosophila melanogaster. Chromosoma. 2004;113(4):188-96. Epub 2004/09/01. doi: 10.1007/s00412-004-0308-2. PubMed PMID: 15338233.

74. Piergentili R, Mencarelli C. Drosophila melanogaster kl-3 and kl-5 Y-loops harbor triple-stranded nucleic acids. J Cell Sci. 2008;121(Pt 10):1605-12. Epub 2008/04/24. doi: 10.1242/jcs.025320. PubMed PMID: 18430782.

75. Chang CH, Larracuente AM. Genomic changes following the reversal of a Y chromosome to an autosome in Drosophila pseudoobscura. Evolution. 2017;71(5):1285-96. Epub 2017/03/23. doi: 10.1111/evo.13229. PubMed PMID: 28322435; PubMed Central PMCID: PMCPMC5485016.

76. Jagannathan M, Warsinger-Pepe N, Watase GJ, Yamashita YM. Comparative Analysis of Satellite DNA in the Drosophila melanogaster Species Complex. G3 (Bethesda). 2017;7(2):693-704. Epub 2016/12/23. doi: 10.1534/g3.116.035352. PubMed PMID: 28007840; PubMed Central PMCID: PMCPMC5295612.

77. Mahajan S, Bachtrog D. Convergent evolution of Y chromosome gene content in flies. Nat Commun. 2017;8(1):785. Epub 2017/10/06. doi: 10.1038/s41467-017-00653-x. PubMed PMID: 28978907; PubMed Central PMCID: PMCPMC5627270.

78. Greil F, Ahmad K. Nucleolar dominance of the Y chromosome in Drosophila melanogaster. Genetics. 2012;191(4):1119-28. Epub 2012/06/01. doi: 10.1534/genetics.112.141242. PubMed PMID: 22649076; PubMed Central PMCID: PMCPMC3415996.

79. Mahadevaraju S, Fear JM, Akeju M, Galletta BJ, Pinheiro M, Avelino CC, et al. Dynamic sex chromosome expression in Drosophila male germ cells. Nat Commun. 2021;12(1):892. Epub 2021/02/11. doi: 10.1038/s41467-021-20897-y. PubMed PMID: 33563972; PubMed Central PMCID: PMCPMC7873209.

80. Hess O, Meyer GF. Genetic activities of the Y chromosome in Drosophila during spermatogenesis. Adv Genet. 1968;14:171-223. Epub 1968/01/01. doi: 10.1016/s0065-2660(08)60427-7. PubMed PMID: 4884781.

81. Loh BJ, Cullen CF, Vogt N, Ohkura H. The conserved kinase SRPK regulates karyosome formation and spindle microtubule assembly in Drosophila oocytes. J Cell Sci. 2012;125(Pt 19):4457-62. Epub 2012/08/03. doi: 10.1242/jcs.107979. PubMed PMID: 22854045; PubMed Central PMCID: PMCPMC3500864.

82. Kogan GL, Usakin LA, Ryazansky SS, Gvozdev VA. Expansion and evolution of the X-linked testis specific multigene families in the melanogaster species subgroup. PLoS One. 2012;7(5):e37738. Epub 2012/06/01. doi: 10.1371/journal.pone.0037738. PubMed PMID: 22649555; PubMed Central PMCID: PMCPMC3359341.

83. Danilevskaya ON, Kurenova EV, Pavlova MN, Bebehov DV, Link AJ, Koga A, et al. He-T family DNA sequences in the Y chromosome of Drosophila melanogaster share homology with the X-linked stellate genes. Chromosoma. 1991;100(2):118-24. Epub 1991/02/01. PubMed PMID: 1672635.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

84. Aravin AA, Klenov MS, Vagin VV, Bantignies F, Cavalli G, Gvozdev VA. Dissection of a natural RNA silencing process in the Drosophila melanogaster germ line. Mol Cell Biol. 2004;24(15):6742-50. Epub 2004/07/16. doi: 10.1128/MCB.24.15.6742-6750.2004. PubMed PMID: 15254241; PubMed Central PMCID: PMCPMC444866.

85. Ohta T. Some models of gene conversion for treating the evolution of multigene families. Genetics. 1984;106(3):517-28. PubMed PMID: 6706111; PubMed Central PMCID: PMCPMC1224254.

86. Backstrom N, Ceplitis H, Berlin S, Ellegren H. Gene conversion drives the evolution of HINTW, an ampliconic gene on the female-specific avian W chromosome. Mol Biol Evol. 2005;22(10):1992-9. doi: 10.1093/molbev/msi198. PubMed PMID: 15972846.

87. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 1997;13(5):555-6. Epub 1997/11/21. PubMed PMID: 9367129.

88. Dover G. Molecular drive: a cohesive mode of species evolution. Nature. 1982;299(5879):111-7. Epub 1982/09/09. doi: 10.1038/299111a0. PubMed PMID: 7110332.

89. Malone CD, Lehmann R, Teixeira FK. The cellular basis of hybrid dysgenesis and Stellate regulation in Drosophila. Curr Opin Genet Dev. 2015;34:88-94. Epub 2015/10/10. doi: 10.1016/j.gde.2015.09.003. PubMed PMID: 26451497; PubMed Central PMCID: PMCPMC4674331.

90. Palumbo G, Bonaccorsi S, Robbins LG, Pimpinelli S. Genetic analysis of Stellate elements of Drosophila melanogaster. Genetics. 1994;138(4):1181-97. Epub 1994/12/01. PubMed PMID: 7896100; PubMed Central PMCID: PMCPMC1206257.

91. Meyer GF, Hess O, Beermann W. Phasenspezifische Funktionsstrukturen in Spermatocytenkernen von Drosophila melanogaster und Ihre Abhängigkeit vom Y-Chromosom. Chromosoma. 1961;12(1):676-716. doi: 10.1007/BF00328946.

92. Chan SH, Yu AM, McVey M. Dual roles for DNA polymerase theta in alternative end-joining repair of double-strand breaks in Drosophila. PLoS Genet. 2010;6(7):e1001005. Epub 2010/07/10. doi: 10.1371/journal.pgen.1001005. PubMed PMID: 20617203; PubMed Central PMCID: PMCPMC2895639.

93. McVey M, Lee SE. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet. 2008;24(11):529-38. Epub 2008/09/24. doi: 10.1016/j.tig.2008.08.007. PubMed PMID: 18809224; PubMed Central PMCID: PMCPMC5303623.

94. Chang HHY, Pannunzio NR, Adachi N, Lieber MR. Non-homologous DNA end joining and alternative pathways to double-strand break repair. Nat Rev Mol Cell Biol. 2017;18(8):495-506. Epub 2017/05/18. doi: 10.1038/nrm.2017.48. PubMed PMID: 28512351.

95. Huang W, Massouras A, Inoue Y, Peiffer J, Ramia M, Tarone AM, et al. Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. Genome research. 2014;24(7):1193-208. Epub 2014/04/10. doi: 10.1101/gr.171546.113. PubMed PMID: 24714809; PubMed Central PMCID: PMCPMC4079974.

96. Signor SA, New FN, Nuzhdin S. A Large Panel of Drosophila simulans Reveals an Abundance of Common Variants. Genome biology and evolution. 2018;10(1):189-206. doi: 10.1093/gbe/evx262. PubMed PMID: 29228179; PubMed Central PMCID: PMCPMC5767965.

97. Wei KHC, Lower SE, Caldas IV, Sless TJ, Barbash DA, Clark AG. Variable rates of simple satellite gains across the Drosophila phylogeny. Molecular Biology and Evolution. 2018:msy005-msy. doi: 10.1093/molbev/msy005.

98. Cechova M, Harris RS, Tomaszkiewicz M, Arbeithuber B, Chiaromonte F, Makova KD. High satellite repeat turnover in great apes studied with short- and long-read technologies. Mol Biol Evol. 2019. Epub 2019/07/06. doi: 10.1093/molbev/msz156. PubMed PMID: 31273383; PubMed Central PMCID: PMCPMC6805231.

99. Abad JP, de Pablos B, Agudo M, Molina I, Giovinazzo G, Martin-Gallardo A, et al. Genomic and cytological analysis of the Y chromosome of Drosophila melanogaster: telomere-

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

derived sequences at internal regions. Chromosoma. 2004;113(6):295-304. Epub 2004/12/24. doi: 10.1007/s00412-004-0318-0. PubMed PMID: 15616866.

100. Agudo M, Losada A, Abad JP, Pimpinelli S, Ripoll P, Villasante A. Centromeres from telomeres? The centromeric region of the Y chromosome of Drosophila melanogaster contains a tandem array of telomeric HeT-A- and TART-related sequences. Nucleic Acids Res. 1999;27(16):3318-24. Epub 1999/08/24. PubMed PMID: 10454639; PubMed Central PMCID: PMCPMC148565.

101. Chiolo I, Minoda A, Colmenares SU, Polyzos A, Costes SV, Karpen GH. Double-strand breaks in heterochromatin move outside of a dynamic HP1a domain to complete recombinational repair. Cell. 2011;144(5):732-44. Epub 2011/03/01. doi: 10.1016/j.cell.2011.02.012. PubMed PMID: 21353298; PubMed Central PMCID: PMCPMC3417143.

102. Melnikova L, Biessmann H, Georgiev P. The Ku protein complex is involved in length regulation of Drosophila telomeres. Genetics. 2005;170(1):221-35. Epub 2005/03/23. doi: 10.1534/genetics.104.034538. PubMed PMID: 15781709; PubMed Central PMCID: PMCPMC1449706.

103. Samper E, Goytisolo FA, Slijepcevic P, van Buul PP, Blasco MA. Mammalian Ku86 protein prevents telomeric fusions independently of the length of TTAGGG repeats and the G-strand overhang. EMBO reports. 2000;1(3):244-52. Epub 2001/03/21. doi: 10.1093/embo-reports/kvd051. PubMed PMID: 11256607; PubMed Central PMCID: PMCPMC1083725.

104. Katsura Y, Sasaki S, Sato M, Yamaoka K, Suzukawa K, Nagasawa T, et al. Involvement of Ku80 in microhomology-mediated end joining for DNA double-strand breaks in vivo. DNA Repair (Amst). 2007;6(5):639-48. Epub 2007/01/24. doi: 10.1016/j.dnarep.2006.12.002. PubMed PMID: 17236818.

105. Lee YCG, Karpen GH. Pervasive epigenetic effects of Drosophila euchromatic transposable elements impact their evolution. Elife. 2017;6. Epub 2017/07/12. doi: 10.7554/eLife.25762. PubMed PMID: 28695823; PubMed Central PMCID: PMCPMC5505702.

106. McKee BD, Hong CS, Das S. On the roles of heterochromatin and euchromatin in meiosis in drosophila: mapping chromosomal pairing sites and testing candidate mutations for effects on X-Y nondisjunction and meiotic drive in male meiosis. Genetica. 2000;109(1-2):77-93. Epub 2001/04/11. PubMed PMID: 11293799.

107. Dimitri P, Pisano C. Position effect variegation in Drosophila melanogaster: relationship between suppression effect and the amount of Y chromosome. Genetics. 1989;122(4):793-800. Epub 1989/08/01. PubMed PMID: 2503420; PubMed Central PMCID: PMCPMC1203755.

108. Henikoff S. Dosage-dependent modification of position-effect variegation in Drosophila. BioEssays : news and reviews in molecular, cellular and developmental biology. 1996;18(5):401-9. Epub 1996/05/01. doi: 10.1002/bies.950180510. PubMed PMID: 8639163.

109. Gatti M, Pimpinelli S. Functional elements in Drosophila melanogaster heterochromatin. Annual review of genetics. 1992;26:239-75. Epub 1992/01/01. doi: 10.1146/annurev.ge.26.120192.001323. PubMed PMID: 1482113.

110. Lawlor MA, Cao W, Ellison CE. A burst of transposon expression accompanies the activation of Y chromosome fertility genes during Drosophila spermatogenesis. bioRxiv. 2021:2021.05.10.443472. doi: 10.1101/2021.05.10.443472.

111. Quenerch'du E, Anand A, Kai T. The piRNA pathway is developmentally regulated during spermatogenesis in Drosophila. RNA. 2016;22(7):1044-54. Epub 2016/05/22. doi: 10.1261/rna.055996.116. PubMed PMID: 27208314; PubMed Central PMCID: PMCPMC4911912.

112. Aravin AA, Naumova NM, Tulin AV, Vagin VV, Rozovsky YM, Gvozdev VA. Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. Curr Biol. 2001;11(13):1017-27. Epub 2001/07/27. PubMed PMID: 11470406.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

113. Chen P, Kotov AA, Godneeva BK, Bazylev SS, Olenina LV, Aravin AA. piRNA-mediated gene regulation and adaptation to sex-specific transposon expression in D. melanogaster male germline. Genes Dev. 2021;35(11-12):914-35. Epub 2021/05/15. doi: 10.1101/gad.345041.120. PubMed PMID: 33985970; PubMed Central PMCID: PMCPMC8168559.

114. Wakimoto BT, Hearn MG. The effects of chromosome rearrangements on the expression of heterochromatic genes in chromosome 2L of Drosophila melanogaster. Genetics. 1990;125(1):141-54. Epub 1990/05/01. PubMed PMID: 2111264; PubMed Central PMCID: PMCPMC1203996.

115. Hearn MG, Hedrick A, Grigliatti TA, Wakimoto BT. The effect of modifiers of position-effect variegation on the variegation of heterochromatic genes of Drosophila melanogaster. Genetics. 1991;128(4):785-97. Epub 1991/08/01. PubMed PMID: 1916244; PubMed Central PMCID: PMCPMC1204552.

116. Dupim EG, Goldstein G, Vanderlinde T, Vaz SC, Krsticevic F, Bastos A, et al. An investigation of Y chromosome incorporations in 400 species of Drosophila and related genera. PLoS Genet. 2018;14(11):e1007770. Epub 2018/11/06. doi: 10.1371/journal.pgen.1007770. PubMed PMID: 30388103; PubMed Central PMCID: PMCPMC6235401.

117. Cocquet J, Ellis PJ, Mahadevaiah SK, Affara NA, Vaiman D, Burgoyne PS. A genetic basis for a postmeiotic X versus Y chromosome intragenomic conflict in the mouse. PLoS Genet. 2012;8(9):e1002900. Epub 2012/10/03. doi: 10.1371/journal.pgen.1002900. PubMed PMID: 23028340; PubMed Central PMCID: PMCPMC3441658.

118. Kruger AN, Brogley MA, Huizinga JL, Kidd JM, de Rooij DG, Hu YC, et al. A Neofunctionalized X-Linked Ampliconic Gene Family Is Essential for Male Fertility and Equal Sex Ratio in Mice. Curr Biol. 2019;29(21):3699-706 e5. Epub 2019/10/22. doi: 10.1016/j.cub.2019.08.057. PubMed PMID: 31630956.

119. Lahn BT, Page DC. A human sex-chromosomal gene family expressed in male germ cells and encoding variably charged proteins. Hum Mol Genet. 2000;9(2):311-9. Epub 1999/12/23. doi: 10.1093/hmg/9.2.311. PubMed PMID: 10607842.

120. Hurst LD. Is Stellate a relict meiotic driver? Genetics. 1992;130(1):229-30. Epub 1992/01/01. PubMed PMID: 1732164; PubMed Central PMCID: PMCPMC1204797.

121. Komissarov AS, Galkina SA, Koshel EI, Kulak MM, Dyomin AG, O'Brien SJ, et al. New high copy tandem repeat in the content of the chicken W chromosome. Chromosoma. 2018;127(1):73-83. Epub 2017/09/28. doi: 10.1007/s00412-017-0646-5. PubMed PMID: 28951974.

122. Bachtrog D. The Y Chromosome as a Battleground for Intragenomic Conflict. Trends Genet. 2020;36(7):510-22. Epub 2020/05/26. doi: 10.1016/j.tig.2020.04.008. PubMed PMID: 32448494.

123. Rogers MJ. Y chromosome copy number variation and its effects on fertility and other health factors: a review. Transl Androl Urol. 2021;10(3):1373-82. Epub 2021/04/15. doi: 10.21037/tau.2020.04.06. PubMed PMID: 33850773; PubMed Central PMCID: PMCPMC8039628.

124. Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. 2016;44(19):e147. Epub 2016/11/02. doi: 10.1093/nar/gkw654. PubMed PMID: 27458204; PubMed Central PMCID: PMCPMC5100563.

125. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10(6):563-9. Epub 2013/05/07. doi: 10.1038/nmeth.2474. PubMed PMID: 23644548.

126. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9(11):e112963. Epub 2014/11/20. doi:

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

10.1371/journal.pone.0112963. PubMed PMID: 25409509; PubMed Central PMCID: PMCPMC4237348.

127. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 2019. Epub 2019/12/11. doi: 10.1038/s41592-019-0669-3. PubMed PMID: 31819265.

128. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37(5):540-6. Epub 2019/04/03. doi: 10.1038/s41587-019-0072-8. PubMed PMID: 30936562.

129. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-10. Epub 1990/10/05. doi: 10.1016/S0022-2836(05)80360-2. PubMed PMID: 2231712.

130. Laetsch D, Blaxter M. BlobTools: Interrogation of genome assemblies [version 1; peer review: 2 approved with reservations]. F1000Research. 2017;6(1287). doi: 10.12688/f1000research.12232.1.

131. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589-95. Epub 2010/01/19. doi: 10.1093/bioinformatics/btp698. PubMed PMID: 20080505; PubMed Central PMCID: PMCPMC2828108.

132. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9. Epub 2009/06/10. doi: 10.1093/bioinformatics/btp352. PubMed PMID: 19505943; PubMed Central PMCID: PMCPMC2723002.

133. Nouhaud P. Long-read based assembly and annotation of a <em>Drosophila simulans</em> genome. bioRxiv. 2018:425710. doi: 10.1101/425710.

134. Gordon SP, Tseng E, Salamov A, Zhang J, Meng X, Zhao Z, et al. Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing. PLoS One. 2015;10(7):e0132628. Epub 2015/07/16. doi: 10.1371/journal.pone.0132628. PubMed PMID: 26177194; PubMed Central PMCID: PMCPMC4503453.

135. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12:491. Epub 2011/12/24. doi: 10.1186/1471-2105-12-491. PubMed PMID: 22192575; PubMed Central PMCID: PMCPMC3280279.

136. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290-5. Epub 2015/02/19. doi: 10.1038/nbt.3122. PubMed PMID: 25690850; PubMed Central PMCID: PMCPMC4643835.

137. Smit A, Hubley R, Green P. RepeatMasker 2013. Available from: http://www.repeatmasker.org.

138. Chang CH, Chavan A, Palladino J, Wei X, Martins NMC, Santinello B, et al. Islands of retroelements are major components of Drosophila centromeres. PLoS Biol. 2019;17(5):e3000241. Epub 2019/05/16. doi: 10.1371/journal.pbio.3000241. PubMed PMID: 31086362; PubMed Central PMCID: PMCPMC6516634.

139. Bailly-Bechet M, Haudry A, Lerat E. “One code to find them all”: a perl tool to conveniently parse RepeatMasker output files. Mobile DNA. 2014;5(1):13. doi: 10.1186/1759-8753-5-13.

140. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573-80. Epub 1998/12/24. PubMed PMID: 9862982; PubMed Central PMCID: PMCPMC148217.

141. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-9. Epub 2012/03/06. doi: 10.1038/nmeth.1923. PubMed PMID: 22388286; PubMed Central PMCID: PMCPMC3322381.

142. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417-9. Epub 2017/03/07.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

doi: 10.1038/nmeth.4197. PubMed PMID: 28263959; PubMed Central PMCID: PMCPMC5600148.

143. Langmead B. Aligning short sequencing reads with Bowtie. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]. 2010;Chapter 11:Unit 11 7. Epub 2010/12/15. doi: 10.1002/0471250953.bi1107s32. PubMed PMID: 21154709; PubMed Central PMCID: PMCPMC3010897.

144. Mose LE, Perou CM, Parker JS. Improved indel detection in DNA and RNA via realignment with ABRA2. Bioinformatics. 2019;35(17):2966-73. Epub 2019/01/17. doi: 10.1093/bioinformatics/btz033. PubMed PMID: 30649250; PubMed Central PMCID: PMCPMC6735753.

145. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformatics. 2012;28(4):593-4. Epub 2011/12/27. doi: 10.1093/bioinformatics/btr708. PubMed PMID: 22199392; PubMed Central PMCID: PMCPMC3278762.

146. Larracuente AM, Ferree PM. Simple method for fluorescence DNA in situ hybridization to squashed chromosomes. JoVE. 2015;95:e52288. doi: doi:10.3791/52288.

147. Pimpinelli S, Bonaccorsi S, Fanti L, Gatti M. Immunostaining of mitotic chromosomes from Drosophila larval brain. Cold Spring Harbor protocols. 2011;2011(9). doi: 10.1101/pdb.prot065524. PubMed PMID: 21880821.

148. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28(12):1647-9. Epub 2012/05/01. doi: 10.1093/bioinformatics/bts199. PubMed PMID: 22543367; PubMed Central PMCID: PMCPMC3371832.

149. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268-74. Epub 2014/11/06. doi: 10.1093/molbev/msu300. PubMed PMID: 25371430; PubMed Central PMCID: PMCPMC4271533.

150. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol. 2018;35(2):518-22. Epub 2017/10/28. doi: 10.1093/molbev/msx281. PubMed PMID: 29077904; PubMed Central PMCID: PMCPMC5850222.

151. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol. 2018;35(6):1547-9. Epub 2018/05/04. doi: 10.1093/molbev/msy096. PubMed PMID: 29722887; PubMed Central PMCID: PMCPMC5967553.

152. Stecher G, Tamura K, Kumar S. Molecular Evolutionary Genetics Analysis (MEGA) for macOS. Mol Biol Evol. 2020. Epub 2020/01/07. doi: 10.1093/molbev/msz312. PubMed PMID: 31904846.

153. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004;20(2):289-90. Epub 2004/01/22. PubMed PMID: 14734327.

154. Thornton K. Libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics. 2003;19(17):2325-7. Epub 2003/11/25. PubMed PMID: 14630667.

155. Hudson RR. Estimating the recombination parameter of a finite population model without selection. Genetical research. 1987;50(3):245-50. PubMed PMID: 3443297.

156. Hudson RR, Kaplan NL. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics. 1985;111(1):147-64. Epub 1985/09/01. PubMed PMID: 4029609; PubMed Central PMCID: PMCPMC1202594.

157. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. 2019;47(D1):D419-D26. Epub 2018/11/09. doi: 10.1093/nar/gky1038. PubMed PMID: 30407594; PubMed Central PMCID: PMCPMC6323939.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

Supplementary text

Validation of variants in Y-linked gene families

We mapped Illumina reads from male genomic DNA and testis RNAseq to the reconstructed ancestral transcript sequences of each gene cluster (*Lhk-1*, *Lhk-2*, *CK2ßtes-Y*) to estimate the expression level of the different Y-linked copies. We first asked if the variants in these two gene families found in our assemblies can be consistently detected in Illumina reads from male genomes. We found that the abundance of derived variants in these two gene families in the DNA-seq data are highly correlated to the frequency of variants in our assemblies (R = 0.89 and 0.98 in *D. mauritiana* and *D. simulans,* respectively). For 559 variants in the *D. simulans* assembly, 33 of them (28 appear once and four appear twice) are missing from the DNA-seq data. For 446 variants in the *D. mauritiana* assembly, 43 of them (32 appear once and six appear twice) are missing from the DNA-seq data. Additionally, nine and eight inconsistent variants are located near (< 100 bp) the start or end of transcripts in *D. simulans* and *D. mauritiana*, respectively. These regions at the edges of transcripts might have fewer Illumina reads coverage than more central regions.

We compared the proportion of synonymous and nonsynonymous changes between copies with high and low expression using transcriptome data to infer selection pressures on different mutations (Fig S10; Table S21).

To reduce the effect of sequencing errors and simplify the phylogenetic analyses on protein evolution rates, we first reconstructed the ancestral sequences of each gene cluster (*Lhk-1*, *Lhk-2*, *CK2ßtes-Y,* and 2 *CK2ßtes-like*; see Fig 5). The reconstructed

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

ancestral sequences should eliminate misassembled bases, which are typically singletons. We conducted branch-model and branch-site-model tests on the reconstructed ancestral sequence using PAML and inferred that both gene families experienced strong positive selection following their duplication to the Y chromosome (from branch model; Tables S17 and S18, Fig 5). The high rate of protein evolution in the Y-linked ampliconic genes suggests that, in addition to subfunctionalization or degeneration, they may also acquire new functions and adapt to being Y-linked.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

Supplementary Figures

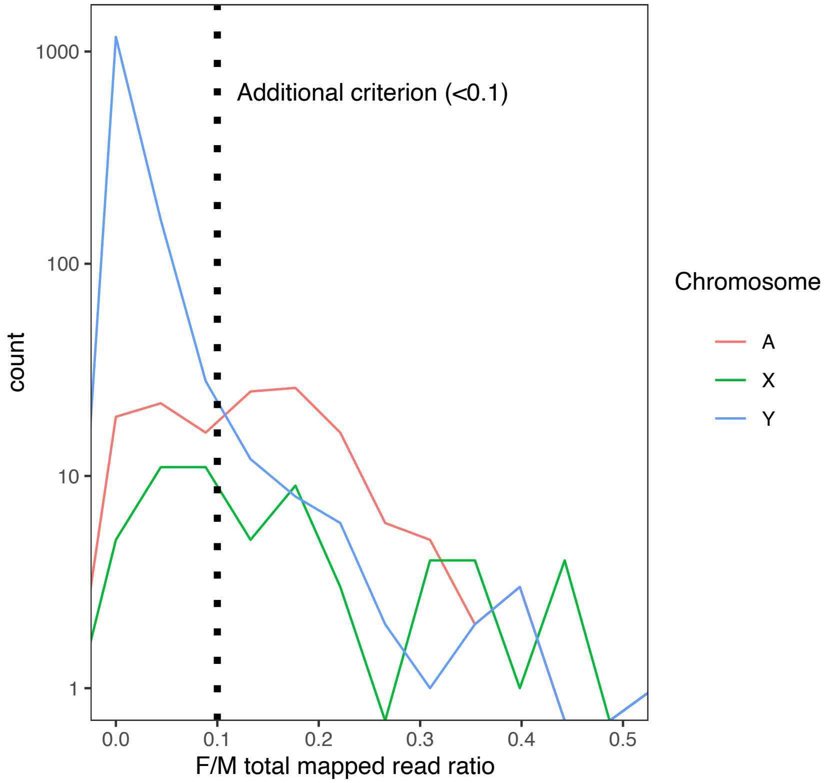


Fig S1. The distribution of female to male total mapped read ratio in each 10-kb window in *D. mauritiana.* Many non-Y regions have median male-to-female coverage 0 in our *D. mauritiana* data. Therefore, we applied an additional criterion based on the female-to-male total mapped reads ratio (<0.1) to reduce the false-positive rate.

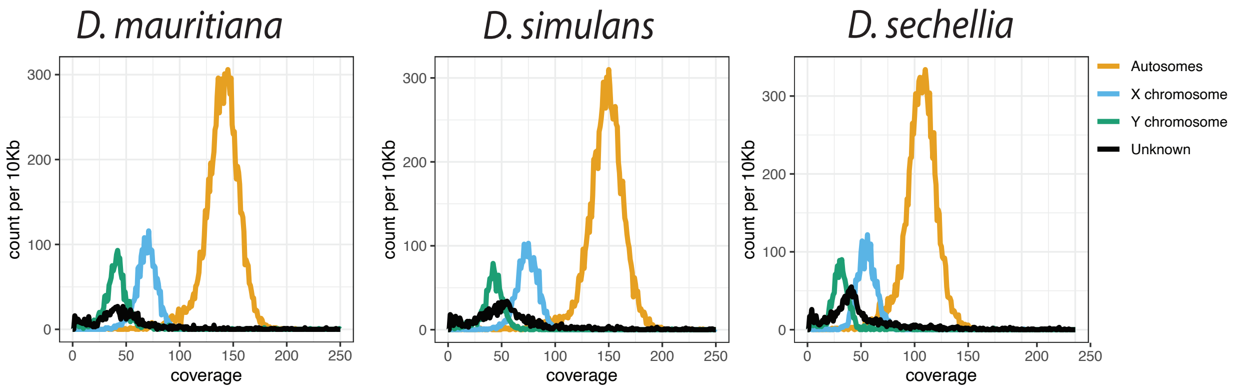


Fig S2. The low Pacbio coverage on the Y chromosome in the *D. simulans* clade. We calculated the median coverage of Pacbio reads every 10-kb and plotted the histogram of depth across genomes based on their chromosome location.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

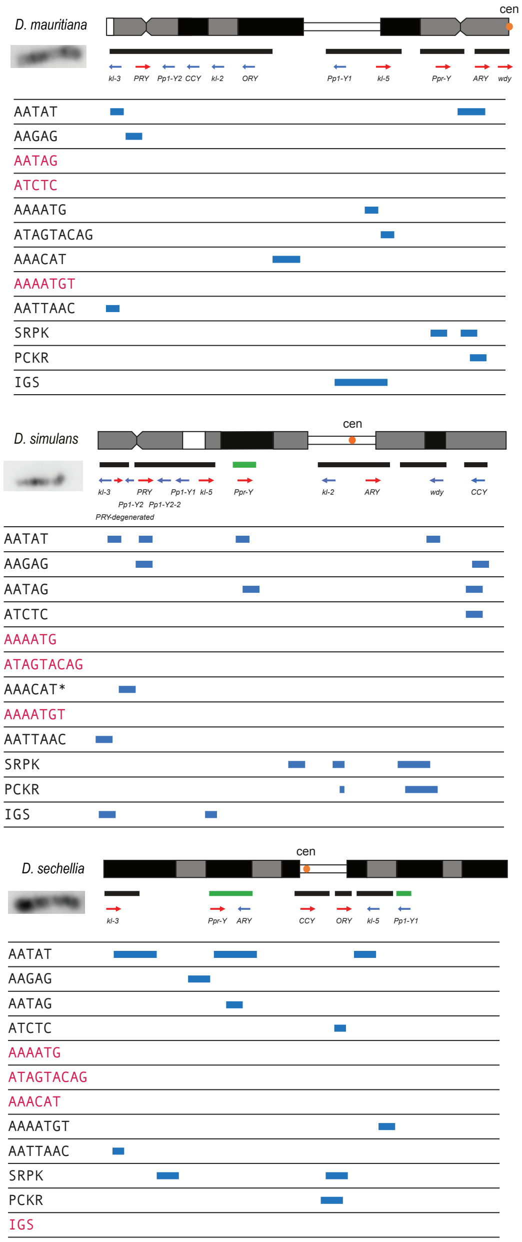


Fig S3. The summarized cytological location of satellite DNA, gene families, and conserved genes on the Y chromosome of the *D. simulans* clade. We used FISH as well as our assemblies to infer the cytological location of Y-linked sequences. The bars represent the location of scaffolds or contigs, and the green bars are scaffolds or contigs without known direction. The satellites in red are sequences we cannot detect on Y chromosomes using FISH.

*Based on the repeat content from the Illumina data (Table S16), the AAACAT signal is probably from the AAACAAT tandem array, instead of AAACAT, in *D. simulans*.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

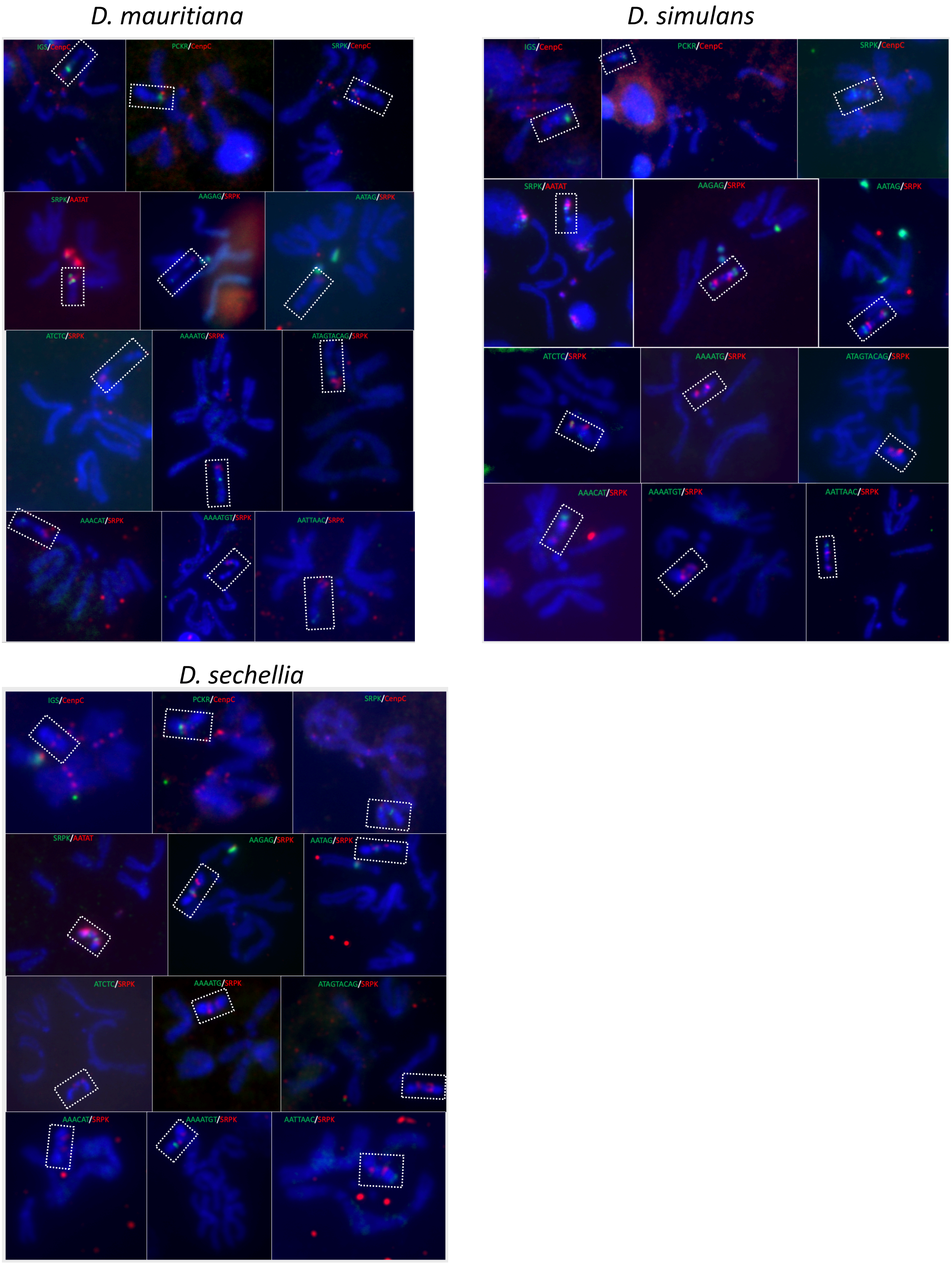


Fig S4. The FISH of satellite and gene families, and conserved genes in the *D. simulans* clade. We surveyed the location of 12 Y-linked sequences using FISH and immunostaining. The colors on the figure represent the probes we used for the experiments.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

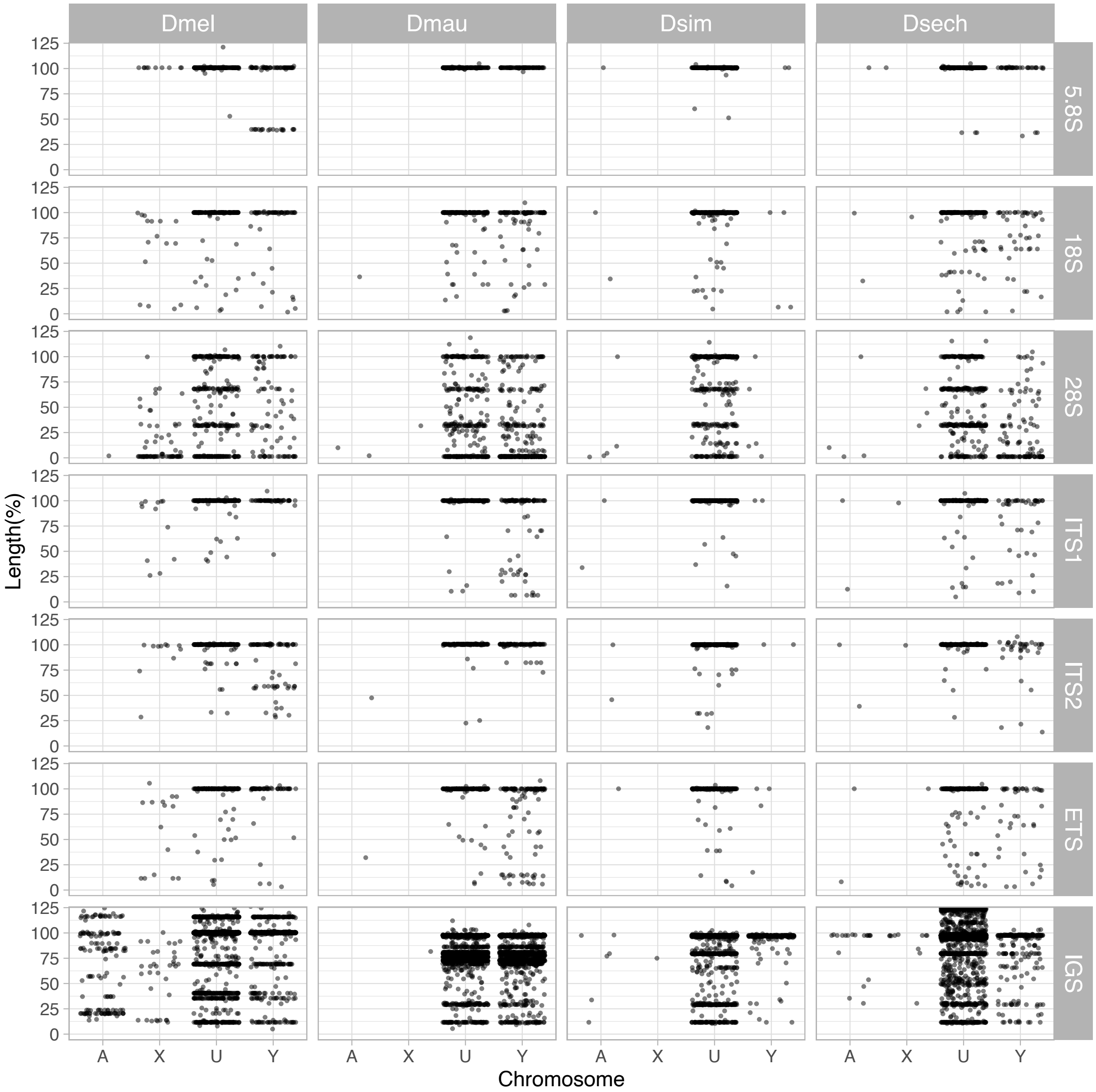
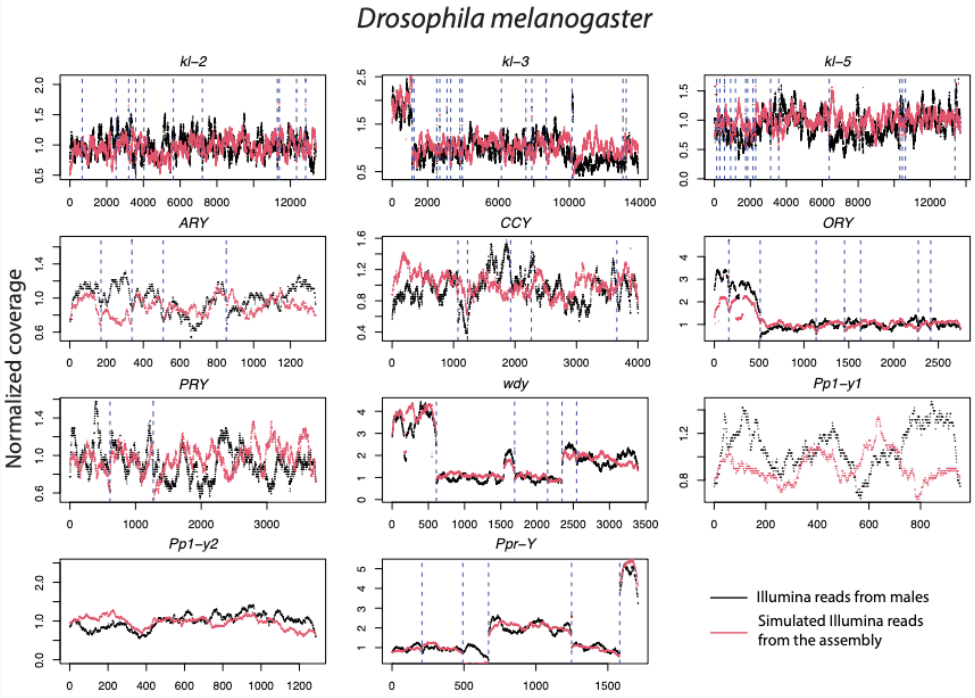
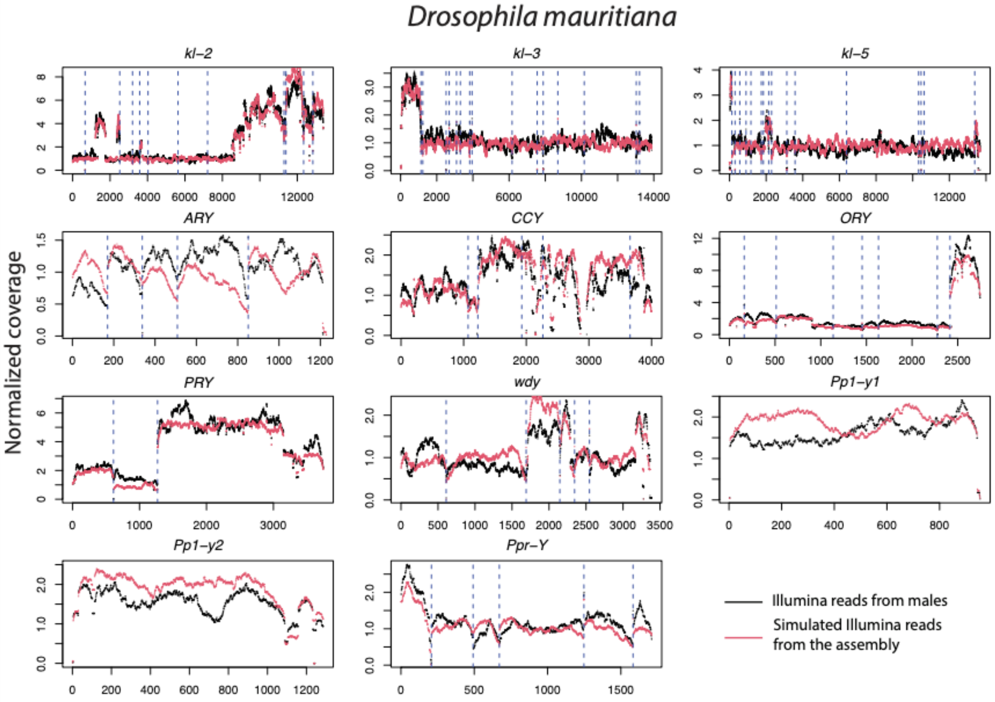


Fig S5. The length of rDNA elements across the chromosomes in *D. melanogaster* and the *D. simulans* clade.We surveyed the length of rDNA elements across chromosomes (A: autosomes, X: X chromosome, U: unknown location and Y: Y chromosome). The length of elements is normalized by the length of consensus from functional elements.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.





bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

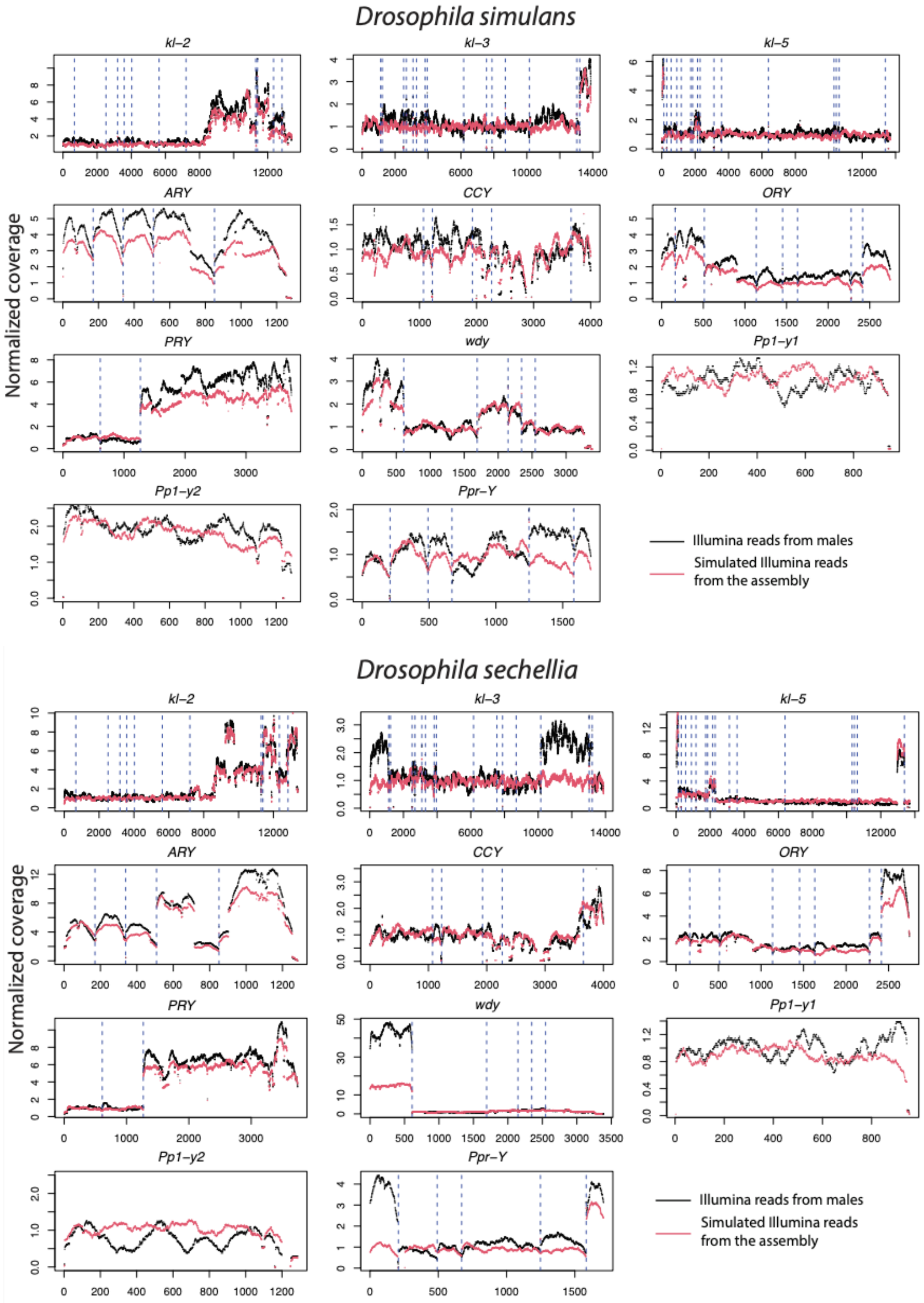
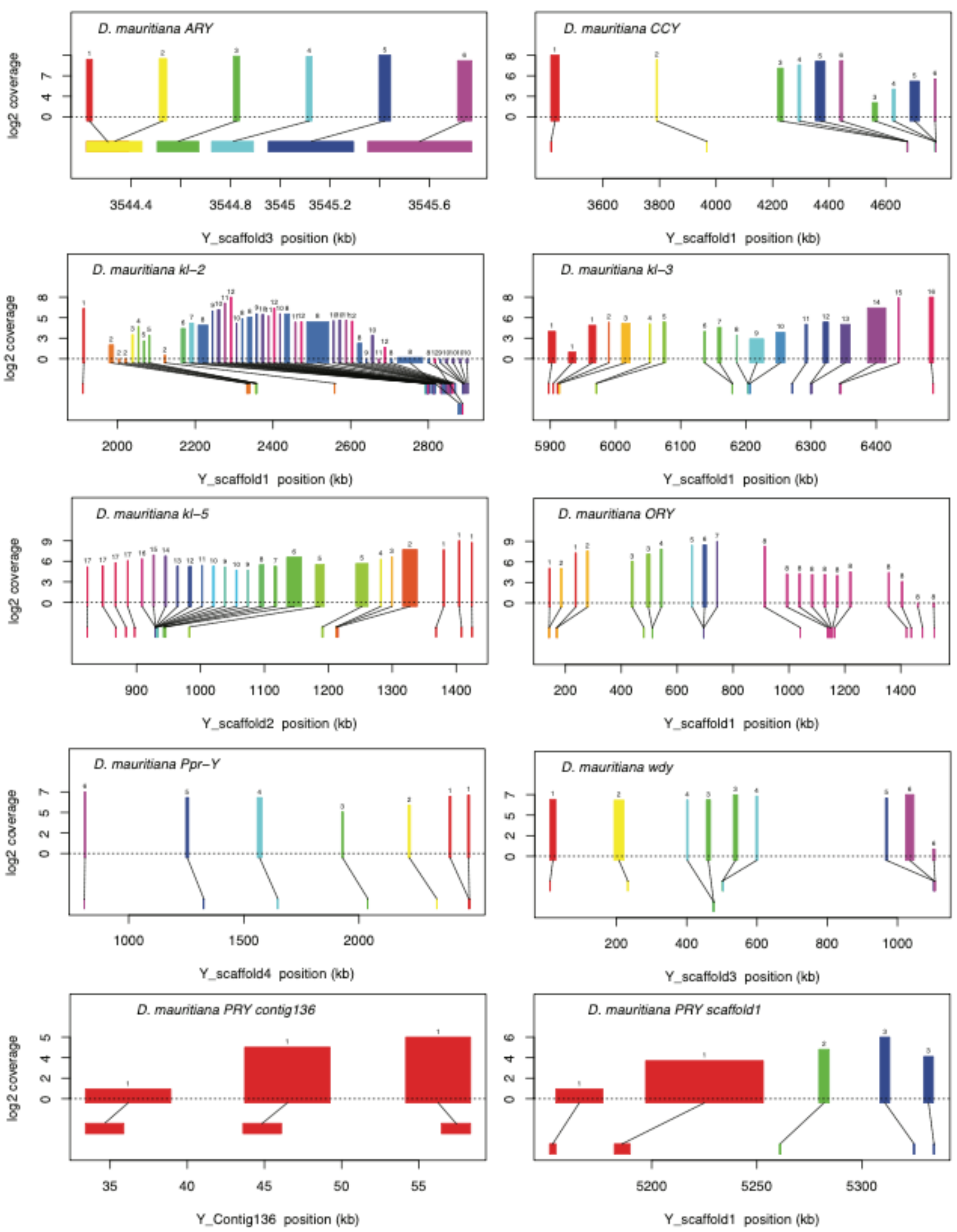
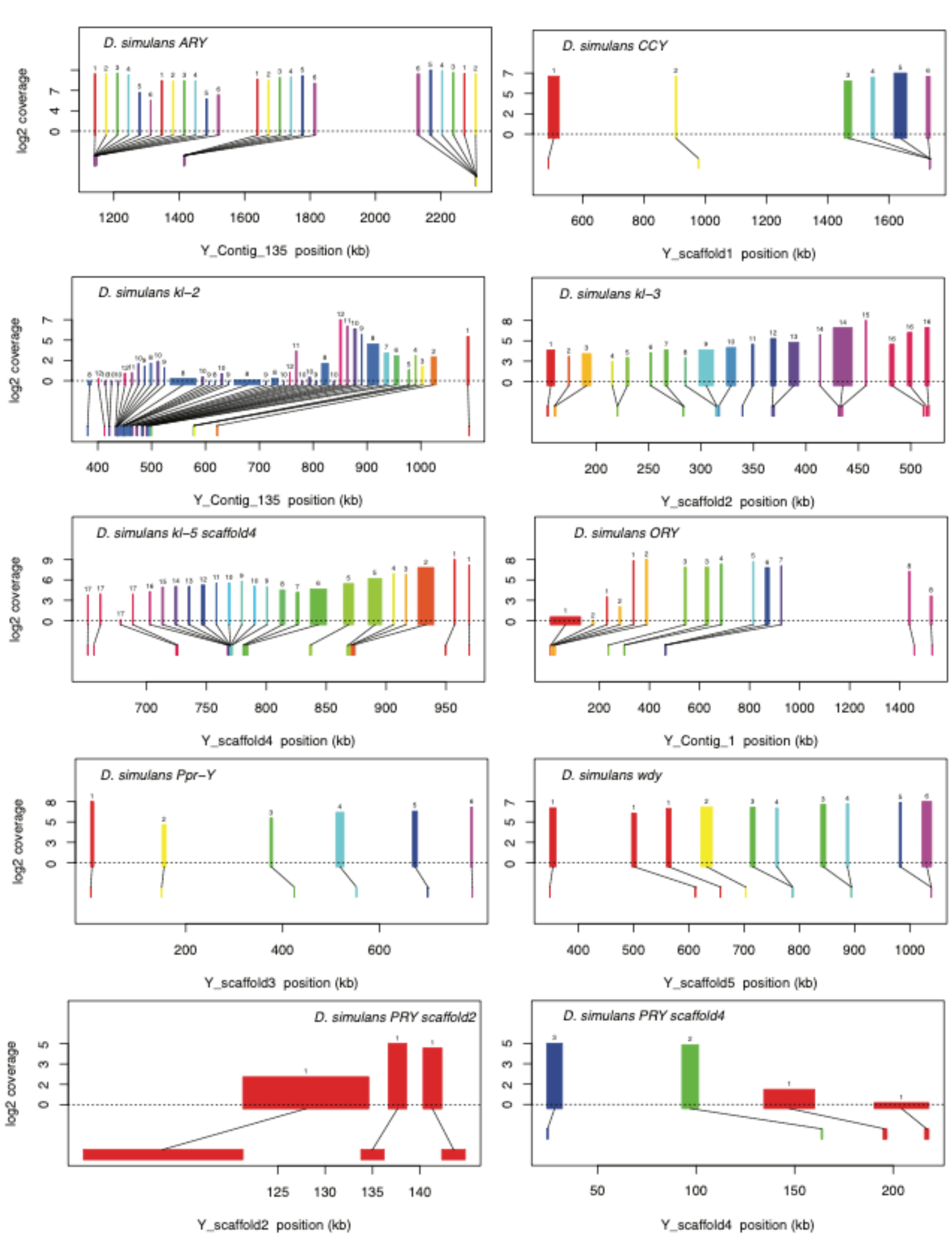


Fig S6. The copy number of male Illumina DNA-seq reads in 11 canonical Y-linked genes. To confirm the copy number of Y-linked genes across species in our assembly, we mapped the Illumina reads from males to a single of *D. melanogaster* Y-linked transcripts and estimated the copy number based on their coverage (black lines). For the comparison, we also simulated Illumina reads from our assemblies and mapped them to the same reference to estimate their copy number (red lines). The dotted lines separate each exon.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

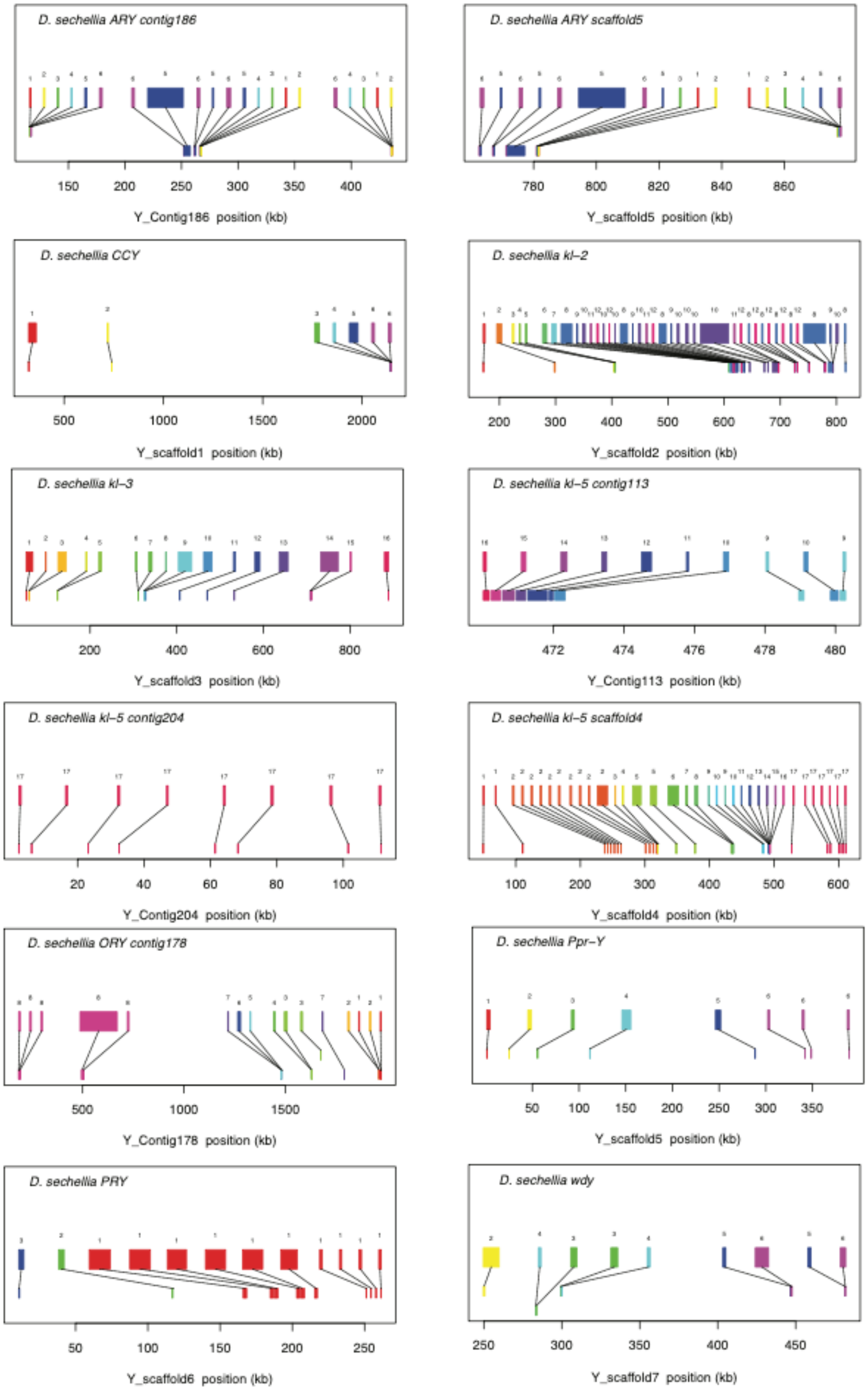

D. sechellia ARY contig186
150
200
250
300
350
400
Y_Contig186 position (kb)
D. sechellia ARY scaffold5
780
800
820
840
860
Y_scaffold5 position (kb)
D. sechellia CCY
500
1000
1500
2000
Y_scaffold1 position (kb)
D. sechellia kl–2
200
300
400
500
600
700
800
Y_scaffold2 position (kb)
D. sechellia kl–3
200
400
600
800
Y_scaffold3 position (kb)
D. sechellia kl–5 contig113
472
474
476
478
480
Y_Contig113 position (kb)
D. sechellia kl–5 contig204
20
40
60
80
100
Y_Contig204 position (kb)
D. sechellia kl–5 scaffold4
100
200
300
400
500
600
Y_scaffold4 position (kb)
D. sechellia ORY contig178
500
1000
1500
Y_Contig178 position (kb)
D. sechellia Ppr–Y
50
100
150
200
250
300
350
Y_scaffold5 position (kb)
D. sechellia PRY
50
100
150
200
250
Y_scaffold6 position (kb)
D. sechellia wdy
250
300
350
400
450
Y_scaffold7 position (kb)

Fig S7. Gene structure of 11 conserved Y-linked genes inferred from assemblies and RNA-seq data. Upper bars indicate exons that are colored and numbered, with their height indicating average read depth from sequenced testes RNA (*D. simulans* and *D. mauritiana* only). Lower bars indicate exon positions on the assembly and position on the Y-axis indicates coding strand.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

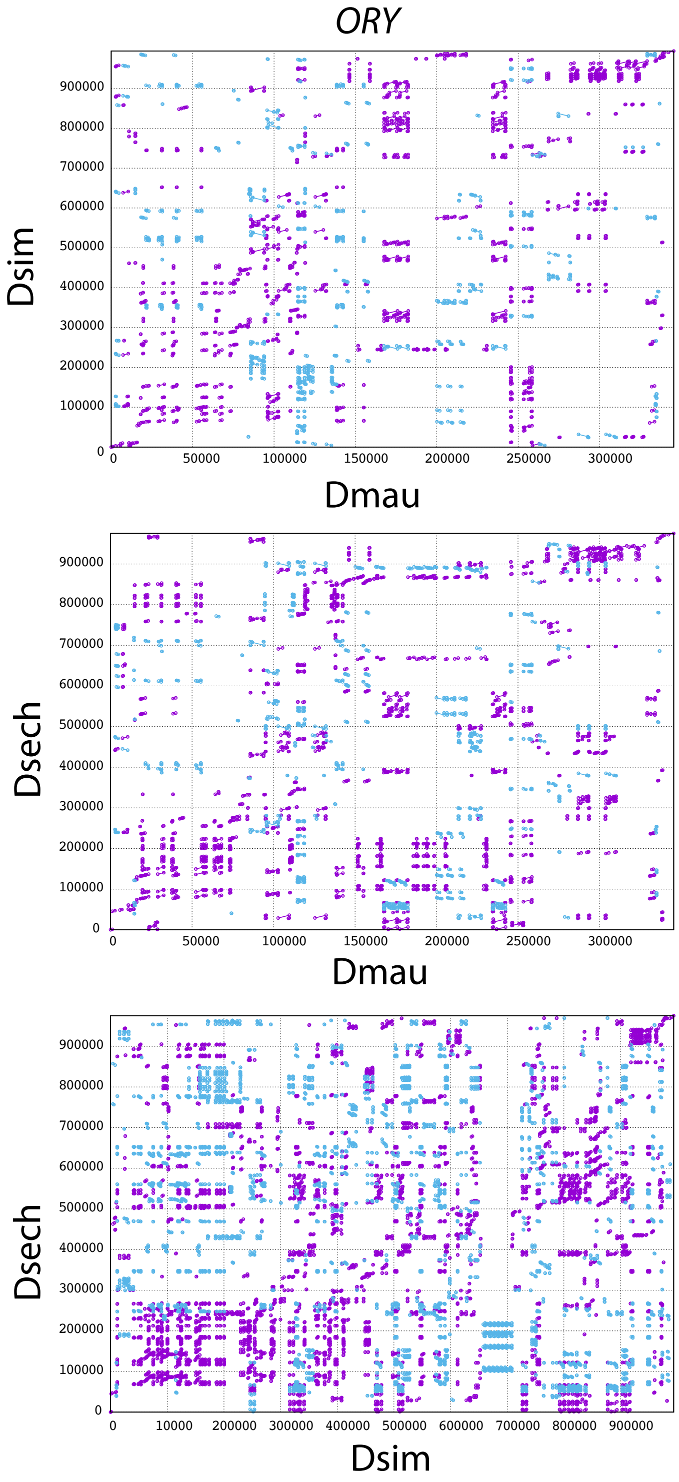


Fig S8. The mummerplot of the *ORY* alignment in the *D. simulans* clade. We used MUMMER to align *ORY* from different species and plot the figure. Purple lines and dots represent forward matches, and blue lines and dots represent reverse matches.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

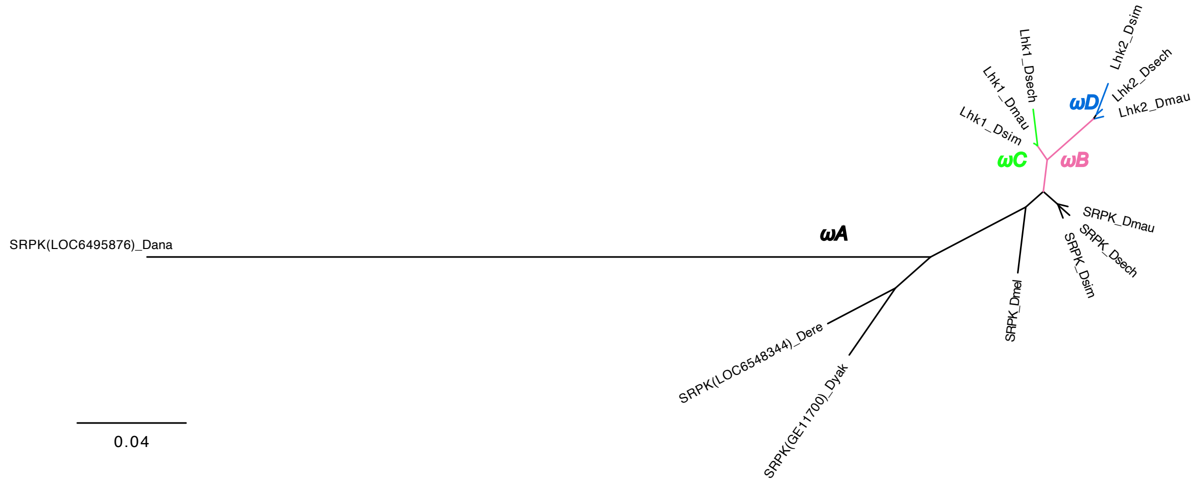


Fig S9. The phylogeny of *Lhk* used in PAML analyses. We marked the branches used in branch-model and branch-site model tests. We did all comparisons using the branch with different colors in likelihood-ratio tests. Please see the detailed results in Table S17.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

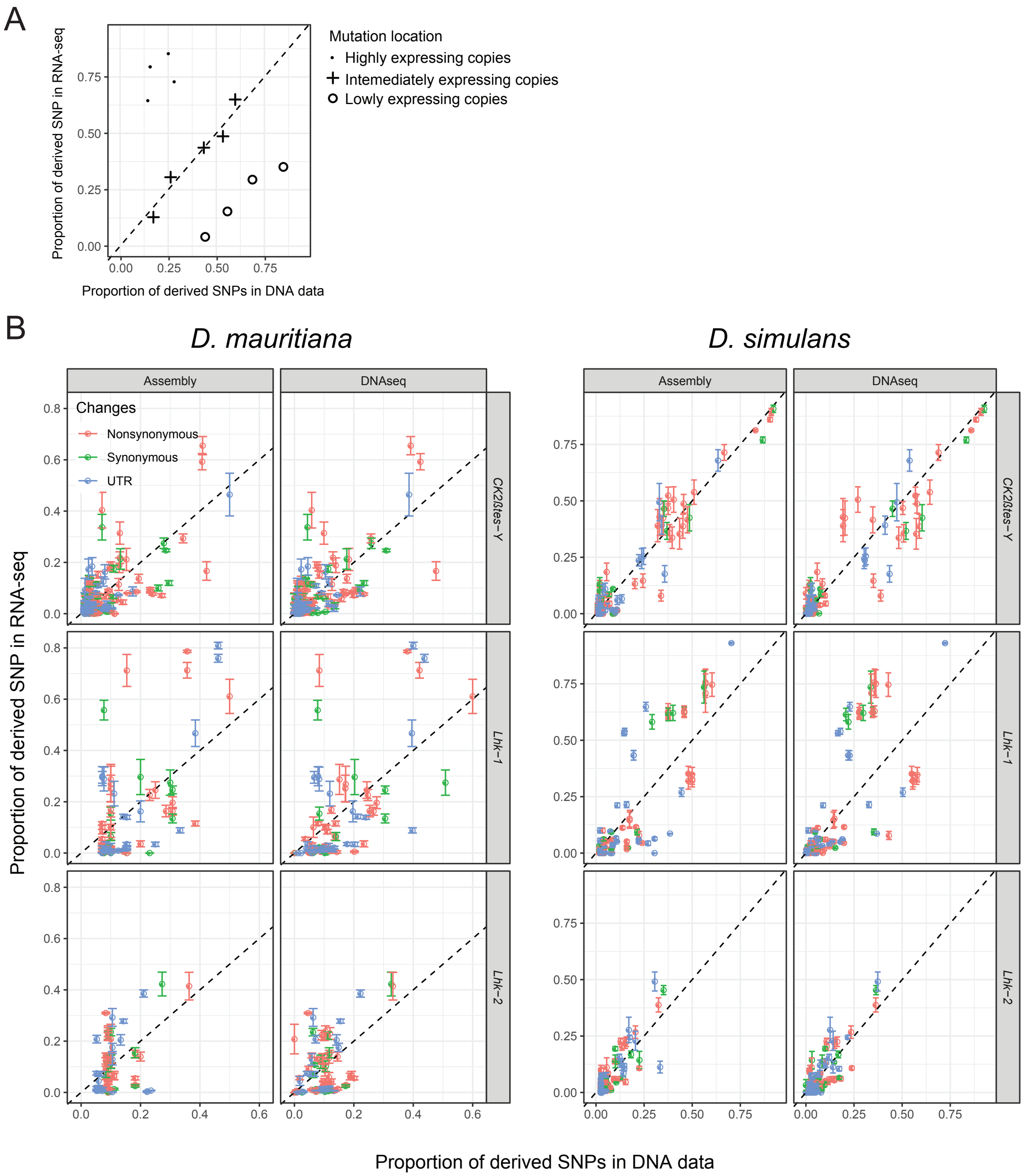


Fig S10. The expression of different copies from *Lhk* and *CK2ßtes-Y* gene families. (A) We quantify the frequency of each derived SNP within the genome using DNA-seq and the expression level of each allele using RNA-seq. We cataloged each SNP as synonymous, nonsynonymous or UTR. (B) We found that across three Y-linked gene families, only highly expressed *Lhk-1* copies have fewer nonsynonymous mutations than lowly expressed copies in *D. simulans*, consistent with purifying selection (Table S12 and S21; Chi-square test's P=0.01). We did not detect other significant changes in other comparisons (Table S12 and S21; Chi-square test's P > 0.01).

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

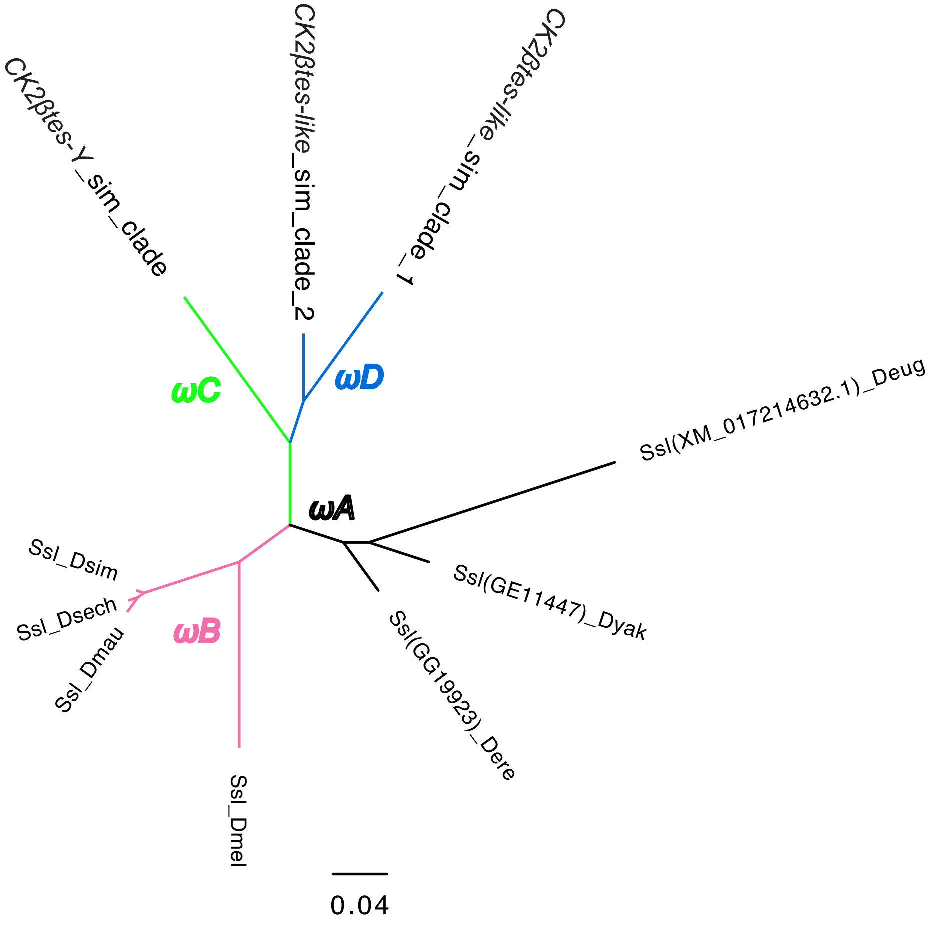
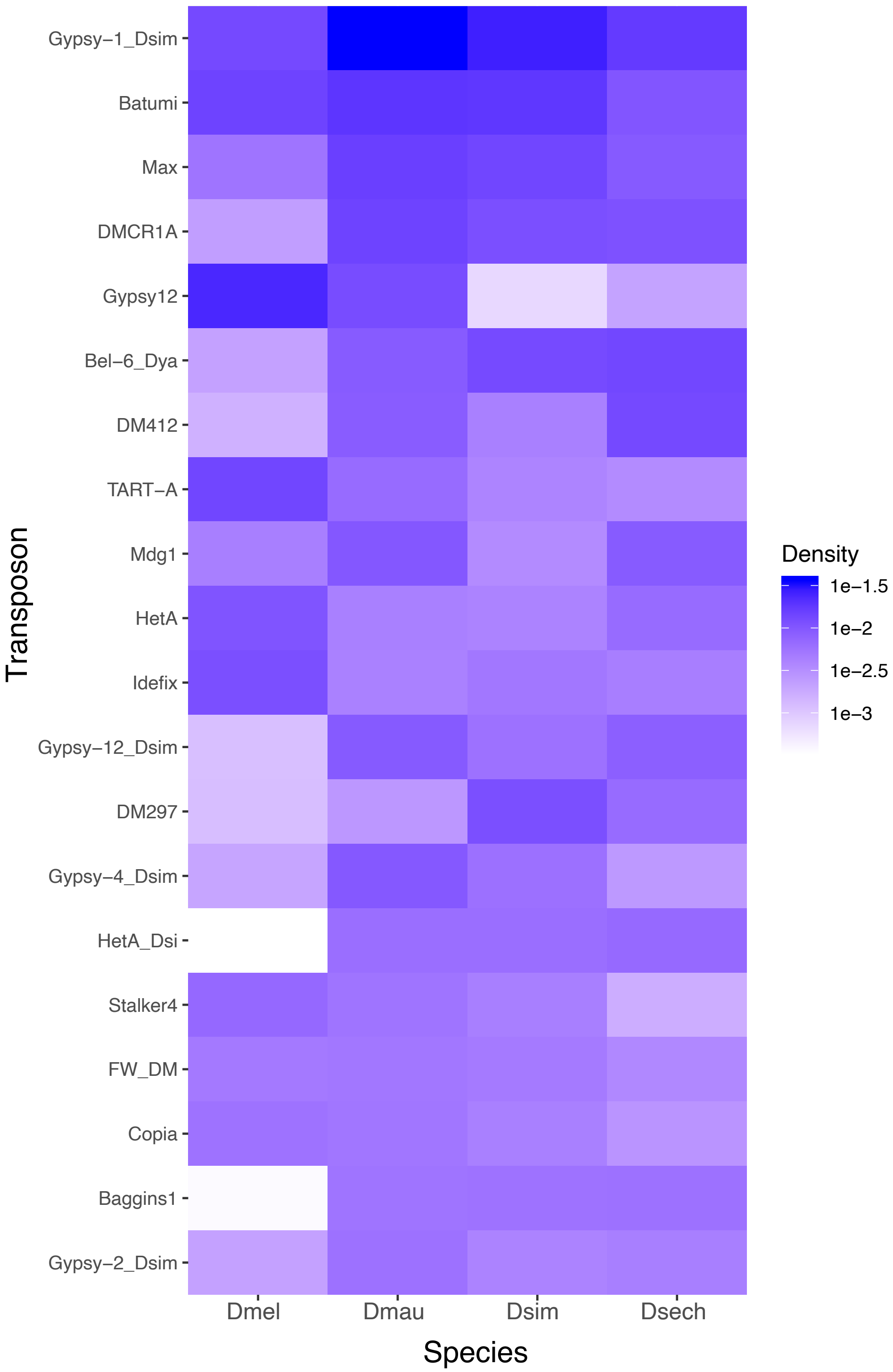


Fig S11. The phylogeny of *CK2ßtes-Y* used in PAML analyses. We marked the branches used in branch-model and branch-site model tests. We did all comparisons using the branch with different colors in likelihood-ratio tests. Please see the detailed results in Table S18.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.



1421

1422 **Fig S12. The abundance of repetitive elements on Y chromosomes of *D.*
1423 *melanogaster* and the *D. simulans* clade species.** We plotted the density of 20 most
1424 enriched (by total occupying sequences) repetitive elements on Y chromosomes across
1425 four species. The colors represent the proportion of repetitive sequences in all
1426 assembled Y-linked sequences.

1427

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

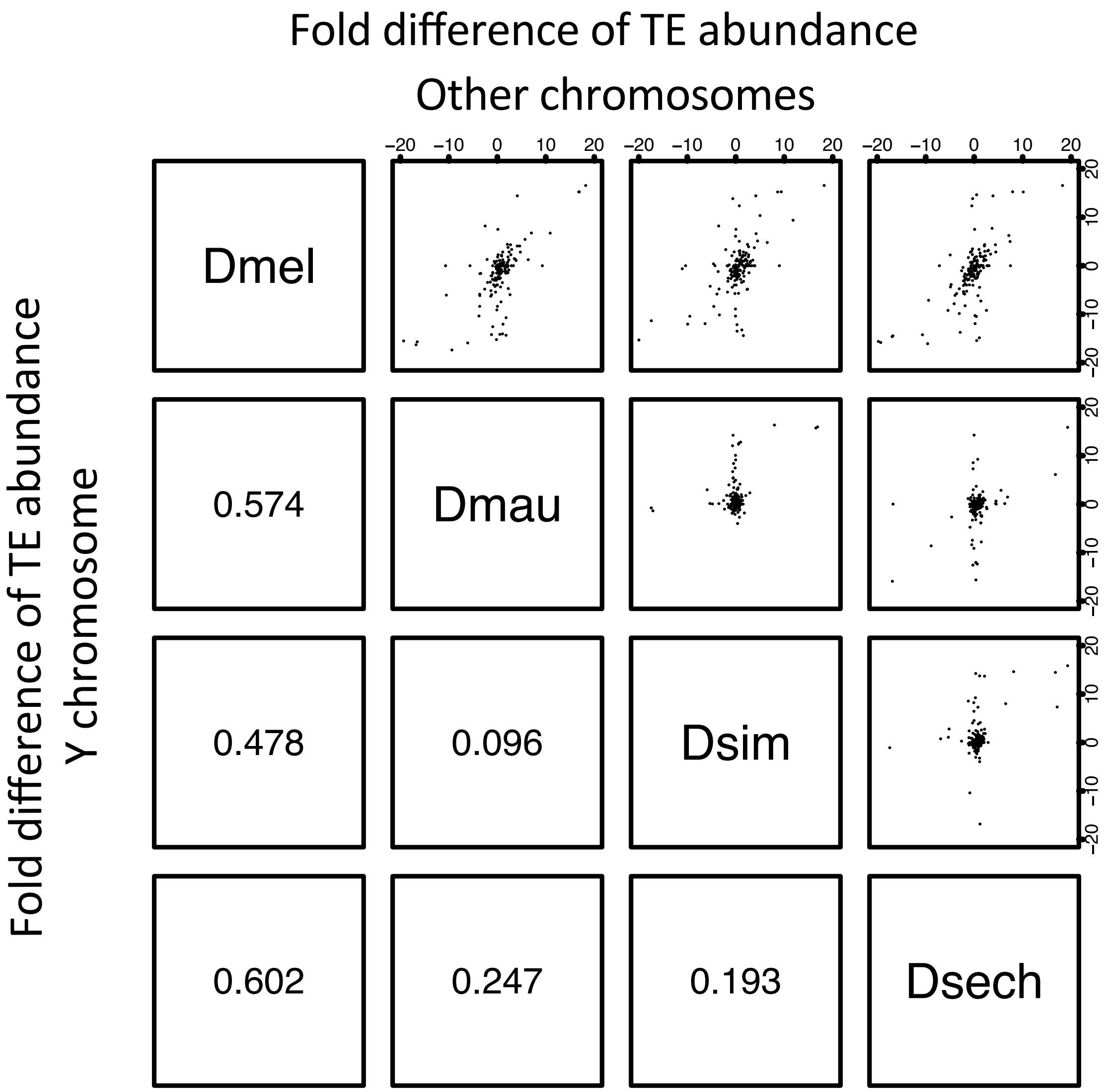


Fig S13. The correlation of TE abundance between Y chromosomes and other chromosomes of *D. melanogaster* and the *D. simulans* clade. We calculated the fold changes of TE occupying sites (bp) between species by chromosomes. Each point from the figures above the diagonal represents the changes of a TE element on the Y chromosome and the other (non-Y) chromosomes. The number below the diagonal shows Spearman's rank correlation coefficient for each comparison.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

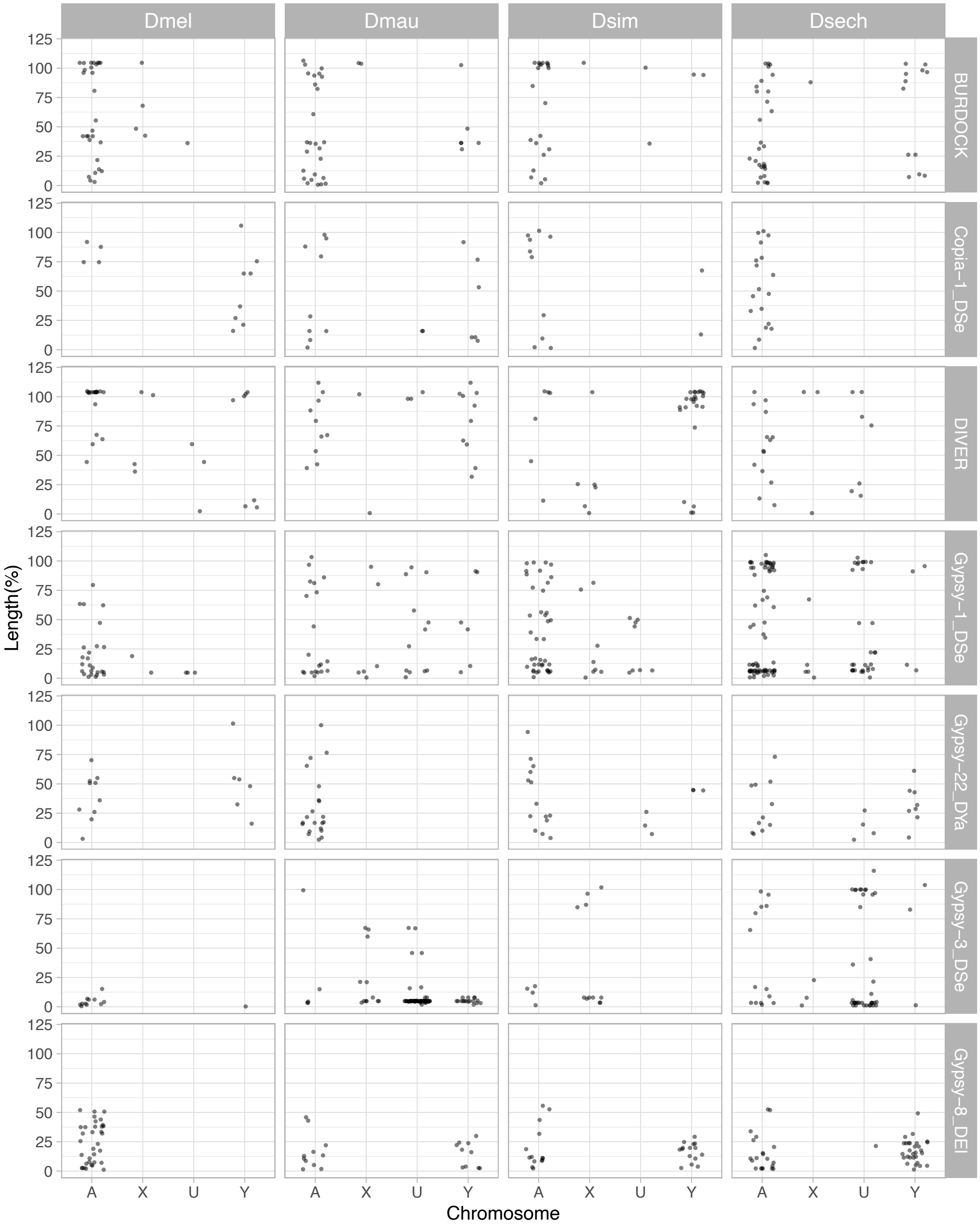


Fig S14. The length of LTR retrotransposons between Y chromosomes and other chromosomes of *D. melanogaster* and the *D. simulans* clade. We surveyed the length of LTR retrotransposons across chromosomes (A: autosomes, X: X chromosome, U: unknown location and Y: Y chromosome). The length of elements is normalized by the length of consensus from full-length elements and represents the ages of each LTR retrotransposon.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

Supplementary Table legend

Table S1. The copy number of exons in conserved Y-linked genes. We listed the copy number of each exon in conserved Y-linked genes based on BLAST results.

Table S2. The estimates of sensitivity and specificity of our Y-linked sequence assignment methods using 10-kb regions with known chromosomal location. We calculated the median female-over-male coverage in our Illumina data in every 10-kb region with known chromosomal location. We then estimated the sensitivity and specificity of our methods using these data.

Table S3. Probe and primer information.

Table S4. The genomic location of duplicated exons in conserved Y-linked genes. We listed the genomic location of each exon in conserved Y-linked genes in our assemblies based on BLAST results.

Table S5. The intron length of all conserved Y-linked genes across species. We showed the length of each Y-linked exon in all conserved Y-linked genes based on BLAST results. If there are multiple copies of an exon, we choose the copy with a complete open reading frame and the highest expression level.

Table S6. Recent Y-linked duplications in *D. melanogaster* and species in the *D. simulans* clade. We list information on the recent Y-linked duplications and genes, including copy numbers, expression levels, phylogenies, and open reading frames. We also included some duplications from repetitive regions where we can date their origins.

Table S7. Enriched GO terms in Y-linked duplicated genes in *D. melanogaster* and the *D. simulans* clade. We searched the enriched GO term from recently duplicated Y-linked genes from Table S6 using PANTHER (Released 20190711; [157]). We listed all GO terms significantly enriched in the duplication (FDR < 0.05).

Table S8. The summary of conserved Y-linked genes and ampliconic genes expression. We summarized the expression level of conserved Y-linked genes and ampliconic genes. We sum up the gene expression for genes with multiple duplicated copies on Y chromosomes.

Table S9. The number of small RNA reads mapped to the repetitive sequences and Y-linked gene families in the *D. simulans* clade.

Table S10. Gene conversion rates for Y-linked ampliconic genes in the *D. simulans* clade. We listed the gene conversion rates and gene similarities on each Y-linked ampliconic gene family (*e.g.*, *Lhk-1, Lhk-2,* and *CK2ßtes-Y*). We estimated gene conversion rates using both gene similarities (p) and population recombination rates (Rmin and rho).

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

Table S11. PAML results for branch and branch-site model analyses of *Lhk* in the *D. simulans* clade. We showed raw results and LRT tests for branch and branch-site model analyses from PAML. We also report rates of protein evolution for each branch in each model and sites under positive selection in the branch-site model analyses.

Table S12. The number of new mutations observed in highly and lowly expressed copies of Y-linked gene families. We list the number of synonymous, nonsynonymous and UTR changes in highly and lowly expressed copies of Y-linked genes families. We suggest that highly expressed copies evolve under stronger selection (positive or purifying) than other copies. Therefore, we compared the number of synonymous changes over nonsynonymous changes in highly expressing copies to the other copies. See Table S21 for detailed information.

Table S13. PAML results for branch and branch-site model analyses of *CK2ßtes-Y* in the *D. simulans* clade. We showed raw results and LRT tests for branch and branch-site model analyses from PAML. We also report rates of protein evolution for each branch in each model and sites under positive selection in the branch-site model analyses.

Table S14. Indels in Y-linked duplications in *D. melanogaster* and the *D. simulans* clade. We listed the position and sizes of all indels we found in Y-linked duplications. We also inferred the potential microhomologies used for MHEJ repairing. We also infer other DSB repairing mechanisms, including tandem duplications and replication slippages, based on the sequence information.

Table S15. Polymorphic indels in *D. melanogaster* and *D. simulans* populations. We listed the position and sizes of polymorphic indels from *D. melanogaster* and *D. simulans* populations. We also inferred the potential microhomologies causing the deletions.

Table S16. The abundance of simple repeats in Illumina reads from male flies estimated with kseek and from our genome assemblies. We used kseek to measure the relative abundance of simple repeats in our Illumina reads. We also used TRF finder to calculate repeat contents in our assemblies. We compared the two results and picked probes for our FISH experiments.

Table S17. Repeat composition across chromosomes in *D. melanogaster* and the *D. simulans* clade. We list the composition of LTR retrotransposon, LINE, DNA transposons, satellite, simple repeats, rRNA, and other repeats across every chromosome in our assemblies.

Table S18. The detail of repetitive sequences across chromosomes in *D. melanogaster* and the *D. simulans* clade. We list the total sequence length from each transposon or complex repeat on Y-linked contigs/scaffolds and other contigs/scaffolds in our assemblies.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.16.456461; this version posted August 16, 2021. 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 4.0 International license.

Table S19. The Illumina coverage and blast result for each contig in the *D. simulans* clade. We used Blast v2.7.1+ [129] with blobtools (v1.0; [130]) to search the nt database (parameters "-task megablast -max_target_seqs 1 -max_hsps 1 -evalue 1e-25"). We estimated the Illumina coverage of each contig in males of *D. mauritiana*, *D. simulans* and *D. sechellia*, respectively.

Table S20. The summary of reads data used in this study

Table S21. The information and read coverage of each SNP in Y-linked gene families from Illumina reads. We listed the coverage of each SNP in Y-linked gene from each RNA-seq replicate and DNA-seq. We also recorded their frequency in our assembly and their translated amino acid. We estimated the expression level of each variant based on the SNP frequency in the genome. We also performed Welch's t-test to compare SNP frequency from DNA-seq and assemblies to it from RNA-seq. We further identify the SNPs associated with the allele that change more than 5 TPM compared to its estimated expression level from its frequency. The SNPs significant in the Welch's t-test and located in lowly or highly expressing alleles are chosen to perform the Chi-square test in Table S12.