RESEARCH ARTICLE

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], Amanda M Larracuente[1]*

[1]Department of Biology, University of Rochester, Rochester, United States; [2]School of Biological Sciences, University of Nebraska-Lincoln, Lincoln, United States

Abstract Y chromosomes across diverse species convergently evolve a gene-poor, heterochromatic organization enriched for duplicated genes, LTR retrotransposons, 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 contributes to the convergent evolution of Y chromosome organization across organisms.

Editor's evaluation

This manuscript by Chang et al. reports the evolutionary patterns of Y-chromosome evolution in *Drosophila*, providing perhaps the most comprehensive interspecific comparison of Y chromosomes available to date. They focus on four species of the *melanogaster* species subgroup and do extensive sequencing and assembly. The manuscript describes the pattern of divergence in these chromosomes, and uses comparative approaches to compare the drivers of evolution in flies and mammals. The authors suggest that the Y chromosome uses a different mechanism to repair double strand breaks than on autosomes. We were impressed by the novelty and rigor of the work as well as the overall presentation of the results.

***For correspondence:** cchang2@fredhutch.org (C-HC); alarracu@UR.Rochester.edu (AML)

Present address: †Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, United States; ‡Department of Molecular Biology and Genetics, Field of Genetics, Genomics and Development, Cornell University, Ithaca, United States

Competing interest: The authors declare that no competing interests exist.

Funding: See page 21

Received: 24 November 2021
Accepted: 18 December 2021
Published: 06 January 2022

Reviewing Editor: Daniel R Matute, University of North Carolina, Chapel Hill, United States

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 (***Rice, 1987a***), which limits the efficacy of natural selection, and causes the accumulation of deleterious mutations through Muller's ratchet, background selection, and hitchhiking effects (***Bachtrog, 2013***; ***Charlesworth, 1978***;

Rice, 1987b; *Charlesworth et al., 1995*; *Charlesworth and Charlesworth, 2000*). 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 (*Bachtrog, 2013*; *Bergero et al., 2015*). 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 (*Bergero et al., 2019*) and new models propose that regulatory evolution is the initial trigger for recombination suppression (*Lenormand et al., 2020*). Regardless of its role in initiating recombination suppression, on degenerating Y chromosomes, 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 (*Bachtrog, 2003*; *Singh et al., 2014*). 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 at least some functional constraint (*Singh et al., 2014*). Although there is evidence suggesting that some Y chromosomes have experienced recent selective sweeps (*Larracuente and Clark, 2013*; *Bachtrog, 2004*), the relative importance of positive selection in shaping Y chromosome evolution remains unclear.

Y chromosomes harbor extensive structural divergence between species, in part through the acquisition of genes from other genomic regions (*Soh et al., 2014*; *Rozen et al., 2003*; *Hughes and Page, 2015*; *Bachtrog et al., 2019*; *Tobler et al., 2017*; *Peichel et al., 2019*; *Brashear et al., 2018*; *Hall et al., 2016*). However, the functions of most Y-linked genes are unknown (*Tobler et al., 2017*; *Hall et al., 2016*; *Chang and Larracuente, 2019*; *Carvalho et al., 2015*). 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 (*Soh et al., 2014*; *Bachtrog et al., 2019*; *Brashear et al., 2018*; *Ellison and Bachtrog, 2019*; *Hughes et al., 2010*; *Mueller et al., 2008*). 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 (*Rozen et al., 2003*; *Connallon and Clark, 2010*).

Detailed analyses of old Y chromosomes have been restricted to a few species with reference-quality assemblies, for example, 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 (*Carlson and Brutlag, 1977*; *Lohe and Brutlag, 1987a*; *Lohe and Brutlag, 1987b*). Recent advances in long-read sequencing make it feasible to assemble large parts of Y chromosomes (*Peichel et al., 2019*; *Hall et al., 2016*; *Chang and Larracuente, 2019*; *Mahajan et al., 2018*) 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 divergent, contribute to hybrid sterility (*Araripe et al., 2016*; *Bayes and Malik, 2009*; *Johnson et al., 1992*; *Coyne, 1985*), and at least four X-linked meiotic drive systems likely shape Y chromosome evolution in these species (*Bozzetti et al., 1995*; *Courret et al., 2019*; *Tao et al., 2007*; *Tao et al., 2001*; *Helleu et al., 2019*; *Branco et al., 2013*; *Montchamp-Moreau et al., 2001*; *Meiklejohn et al., 2018*). Previous genetic and transcriptomic studies suggest that Y chromosome variation can impact male fitness and gene regulation (*Reijo et al., 1995*; *Vogt et al., 1996*; *Sun et al., 2000*; *Repping et al., 2003*; *Morgan and Pardo-Manuel de Villena, 2017*; *Lemos et al., 2010*; *Wang et al., 2018*; *Sackton et al., 2011*). Since there is minimal nucleotide variation and divergence in Y-linked protein-coding sequences within and between these *Drosophila* species (*Singh et al., 2014*; *Larracuente and Clark, 2013*; *Helleu et al., 2019*), 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 (*Zhou et al., 2012*). The chromatin on *Drosophila* Y chromosomes has genome-wide effects on expression level and chromatin states (*Brown and Bachtrog, 2017*), but aside from the rDNA, the molecular basis of Y chromosome divergence and variation in these species remains elusive.

Table 1. Contiguity statistics for heterochromatin-enriched assemblies.

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

*Chang and Larracuente, 2019.
†This paper.

To study the factors and forces shaping the evolution of Y chromosome structure, we assembled the Y chromosomes of the three species in the *D. simulans* clade to reveal their structure and evolution and compared them to *D. melanogaster*. We find 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.

Results

Improving 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 (***Bachtrog et al., 2019***; ***Peichel et al., 2019***; ***Hall et al., 2016***; ***Chang and Larracuente, 2019***). To improve Y chromosome assemblies for comparative genomic analyses, we applied our heterochromatin-sensitive assembly pipeline (***Chang and Larracuente, 2019***) with long reads that we previously generated (***Chakraborty et al., 2021***) to de novo 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 (***Supplementary file 1***; ***Gepner and Hays, 1993***; ***Bernardo Carvalho et al., 2009***). Based on the median male-to-female coverage (***Chang and Larracuente, 2019***), we assigned 13.7–18.9 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 is comparable to *D. melanogaster* (***Table 1***; ***Chang and Larracuente, 2019***). We evaluated our methods by comparing our assignments for every 10 kb window of assembled sequence 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 (***Supplementary file 2***). 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 the false-positive rate from 13% to 5% in regions with known chromosomal location (***Supplementary file 2***; ***Figure 1—figure supplement 1***). We can detect potential misassemblies by looking for discordant assignments between 10 kb windows on the same contigs. Because we do not find any obviously discordant F/M ratios for any contigs, we make chromosome assignments based on median male-to-female coverage and the ratio of female-to-male total mapped reads across whole contigs. Based on these chromosome assignments, we find 40–44% lower PacBio coverage on Y than X chromosomes in all three species (***Figure 1—figure supplement 2***; see Appendix 1).

The cytological organization of the *D. simulans* clade Y chromosomes is not well-described (***Lemeunier and Ashburner, 1984***; ***Roy et al., 2005***; ***Berloco et al., 2005***). 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 (***Figure 1*** and ***Figure 1—figure supplements 3–4***; ***Supplementary file 3***) to determine Y chromosome organization at the cytological level. We also determined the location of the centromeres using immunostaining with a Cenp-C antibody (***Figure 1—figure supplement 4***; ***Erhardt et al., 2008***). These cytological

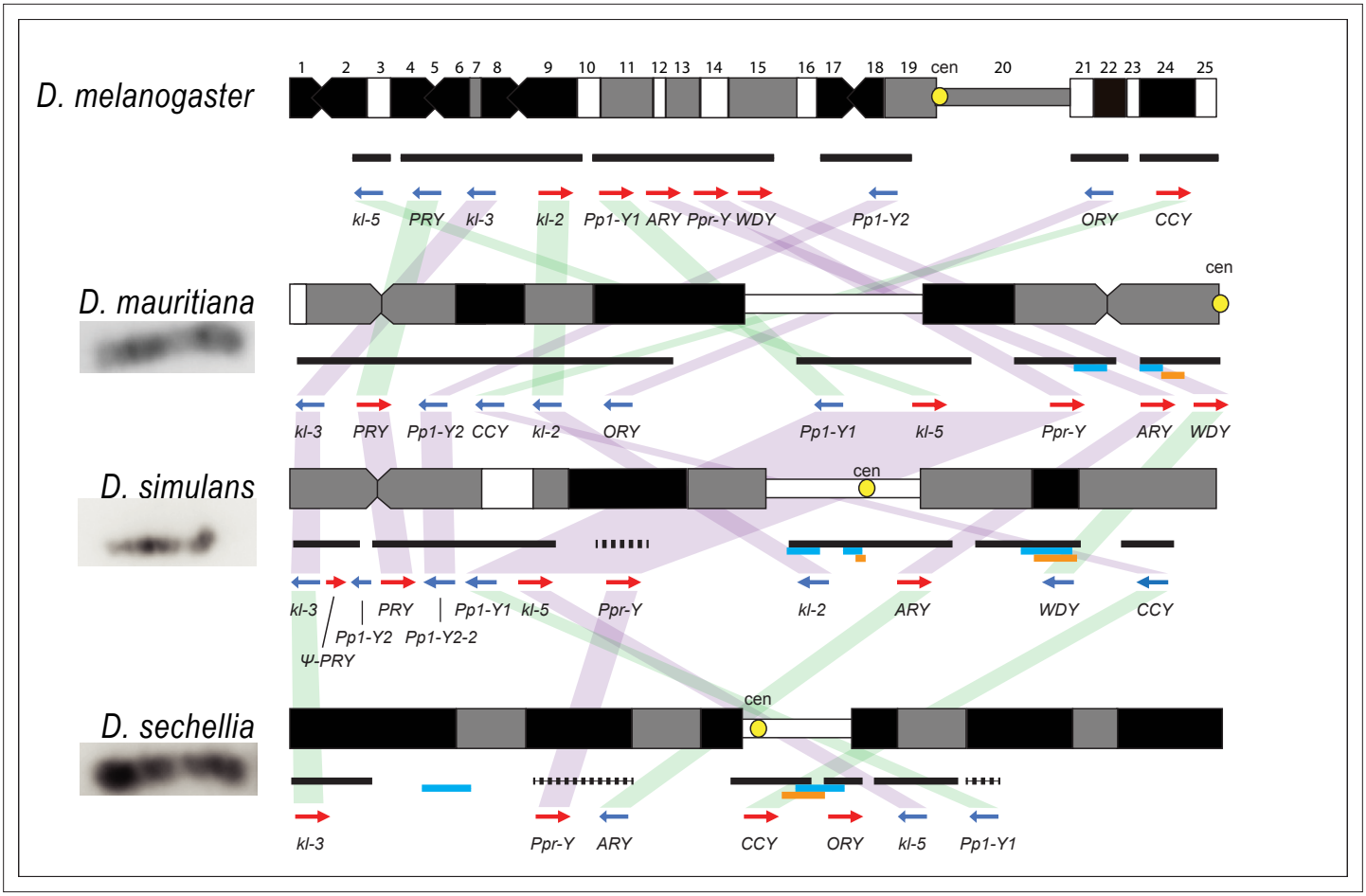

Figure 1. Y chromosome organization in *D. melanogaster* and the three *D. simulans* clade species. Schematics of the cytogenetic maps note the locations of Y-linked genes in *D. melanogaster* and *D. simulans* clade species. The bars show the relative placement of the scaffolds on the cytological bands 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). Yellow circles denote centromere locations (cen). The blocks connecting genes between species highlight the structural rearrangements between species (purple for same, and green for inverted, orientation).

The online version of this article includes the following figure supplement(s) for figure 1:

Figure supplement 1. The distribution of female-to-male total mapped read ratio in each 10-kb window in *D. mauritiana*.

Figure supplement 2. The low Pacbio coverage on Y chromosomes in the *D. simulans* clade.

Figure supplement 3. Summarized cytological location of satellite DNA, gene families, and conserved genes on the Y chromosome of the *D. simulans* clade.

Figure supplement 4. FISH for satellite and gene families, and conserved genes in the *D. simulans* clade.

Figure supplement 5. The length of rDNA elements across chromosomes in *D. melanogaster* and the *D. simulans* clade.

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 (***Figure 1*** and ***Figure 1—figure supplement 3***). Although we cannot examine the detailed organization as a complete contiguous Y-linked sequence, we did not find any discordance between our scaffolds and cytological data. We find evidence for extensive Y chromosomal structural rearrangements, including changes in satellite distribution, gene order, and centromere position. These rearrangements are dramatic even among the *D. simulans* clade species, which diverged less than 250 KYA (***Figure 1*** and ***Figure 1—figure supplement 3***). 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* (**Berloco et al., 2005**). 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.

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 (**Araripe et al., 2016**; **Johnson et al., 1992**). One candidate locus that may contribute to functional divergence and possibly hybrid lethality is the Y-linked rDNA (**Zhou et al., 2012**; **Paredes et al., 2011**). Y-linked rDNA, specifically 28 S rDNA, were lost in *D. simulans* and *D. sechellia*, but not in *D. mauritiana* (**Roy et al., 2005**; **Lohe and Roberts, 2000**; **Lohe and Roberts, 1990**). However, the intergenic spacer (IGS) repeats between rDNA genes, which are responsible for X-Y pairing in *D. melanogaster* males (**McKee and Karpen, 1990**), are retained on both sex chromosomes in all three species (**Roy et al., 2005**; **Lohe and Roberts, 2000**; **Lohe and Roberts, 1990**). Consistent with previous cytological studies (**Roy et al., 2005**; **Lohe and Roberts, 2000**; **Lohe and Roberts, 1990**), we find that *D. simulans* and *D. sechellia* lost most Y-linked 18 S and 28 S rDNA sequences (**Figure 1—figure supplement 5**). 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* (**Figure 1—figure supplements 3–4**; **Roy et al., 2005**; **Lohe and Roberts, 2000**; **Lohe and Roberts, 1990**), 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* (**Helleu et al., 2019**; **Chakraborty et al., 2021**; **Kopp et al., 2006**). 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 **Figure 2—figure supplements 1–2**; **Supplementary file 1**; **Chakraborty, 2020**). Using Illumina reads, we confirm the copy number variation in our assemblies and reveal some duplicated Y-linked exons, particularly in *kl-3*, *WDY,* and *Ppr-Y*, that are not assembled in *D. sechellia* (**Figure 2—figure supplement 1**). Some duplicates may be 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 **Supplementary file 4**), 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 **Figure 2—figure supplement 2**, **Supplementary file 4**). 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. Although the duplications of Y-linked genes can influence their expression, especially for genes with short introns (*e.g. ARY*, *Ppr-Y1* and *Ppr-Y2*), 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, but intron lengths vary between these species (**Figure 3**). The length of longer introns (> 100 bp in any species) is more dynamic than that of short introns (**Figure 3**; **Supplementary file 5**). The dramatic size differences in most introns cannot be attributed to a single deletion or duplication (see *ORY* example in **Figure 2—figure supplement 3**). 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 (**Bonaccorsi et al., 1988**; **Bonaccorsi et al., 1990**). While Y-loops are found across the *Drosophila* genus (**Meyer, 1963**; **Piergentili, 2007**), their potential functions are unknown (**Fingerhut et al., 2019**; **Redhouse et al., 2011**; **Pisano et al., 1993**; **Piergentili et al., 2004**; **Piergentili and Mencarelli, 2008**) and the genes/introns that produce Y-loops differs among species

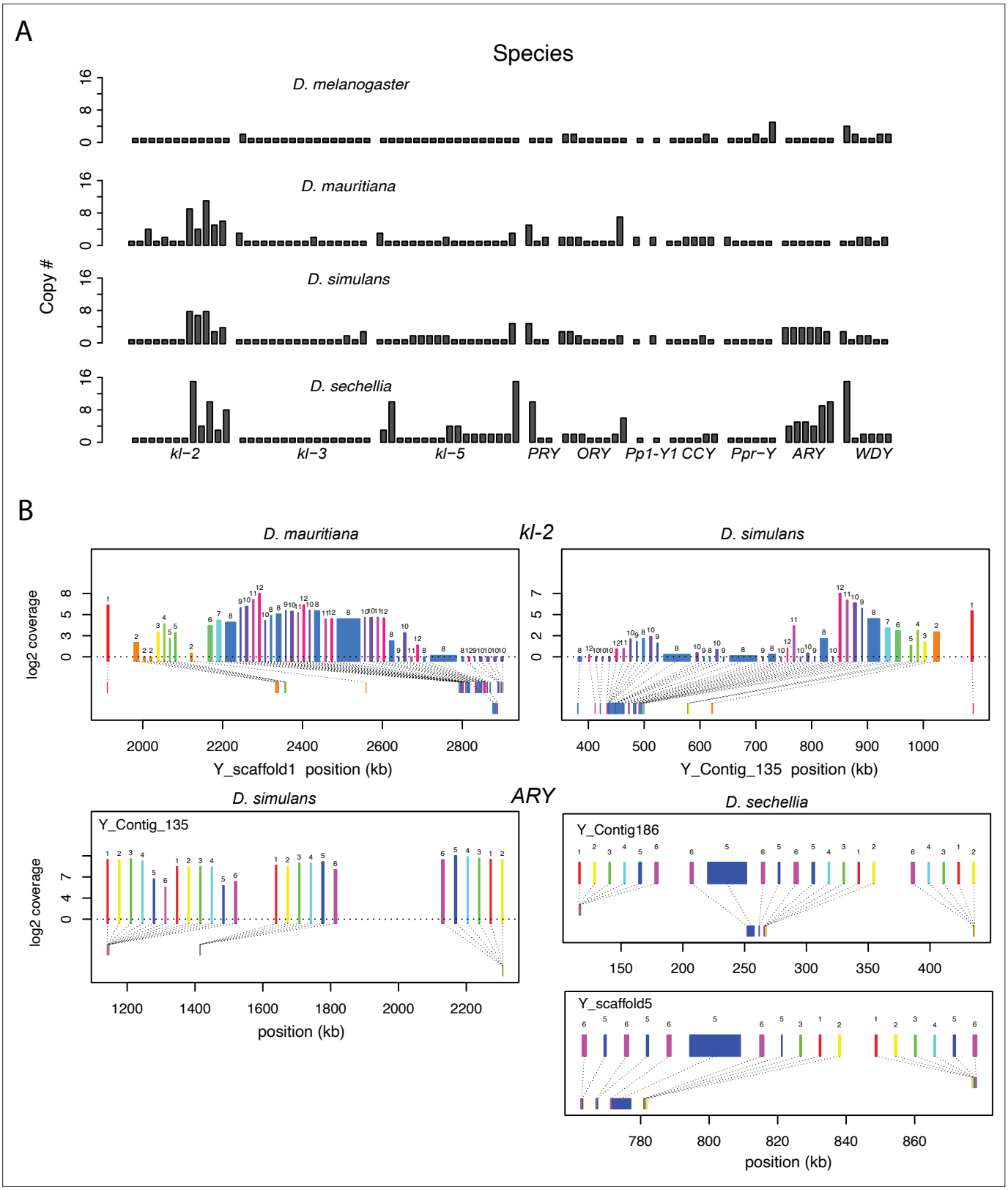

Figure 2. Duplication of canonical Y-linked exons. (**A**) Exon copy number is highly variable across the three *D. simulans* clade species and generally greater than in *D. melanogaster*. (**B**) Gene structure of *kl-2* and *ARY* inferred from assemblies and RNA-seq data. Upper bars indicate exons 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. Some of the duplicated exons are expressed. For short

Figure 2 continued on next page

Figure 2 continued

genes (*e.g., ARY*), the duplicates may be functional and influence protein expression level, unlike duplicated exons of long genes (*e.g., kl-2*).

The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. The coverage of male Illumina DNA-seq reads in 11 canonical Y-linked genes.

Figure supplement 2. Gene structure of 11 conserved Y-linked genes inferred from assemblies and RNA-seq data.

Figure supplement 3. The mummerplot of the *ORY* alignment in the *D. simulans* clade.

(***Chang and Larracuente, 2017***). *D. melanogaster* has three Y-loops transcribed from introns of *ORY* (*ks-1* in previous literature), *kl-3*, and *kl-5* (***Bonaccorsi et al., 1988***). Based on cytological evidence, *D. simulans* has three Y-loops, whereas *D. mauritiana* and *D. sechellia* only have two (***Piergentili, 2007***). 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 (***Supplementary file 6***), and published FISH results (***Jagannathan et al., 2017***). 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).

Consistent with previous studies (***Tobler et al., 2017***; ***Chakraborty et al., 2021***), 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 (***Figure 4***; ***Supplementary file 7***), including eight newly discovered duplications (***Tobler et al., 2017***; ***Chakraborty et al., 2021***). Twenty-eight duplications are DNA-based, 13 are RNA-based, and the rest are unknown due to limited sequence information (***Supplementary file 7***). The rate of transposition to the Y chromosome is about three to four times higher in the *D. simulans* clade compared to *D. melanogaster* (***Chang and Larracuente, 2019***). We also infer that 17 duplicated genes were independently deleted from *D. simulans* clade Y chromosomes. Some of these Y-linked duplications, including *Fdy, Mst77Y* and *pirate*, are known to be functional and/or under purifying selection (***Tobler et al., 2017***; ***Krsticevic et al., 2015***; ***Russell and Kaiser, 1993***; ***Chen et al., 2021***). However, 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

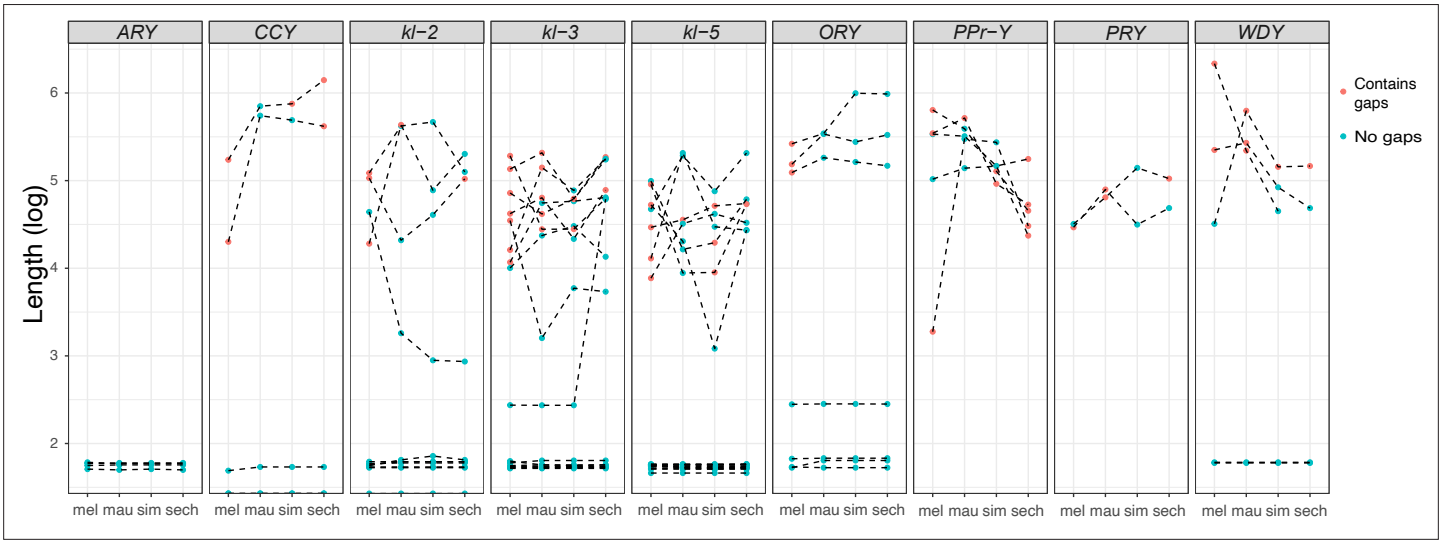

Figure 3. Evolution of intron lengths in canonical Y-linked genes. The intron length in canonical Y-linked genes is different between *D. melanogaster* and the three *D.* simulans clade species. Orthologous introns are connected by dotted lines. Completely assembled introns are in blue and introns with gaps in the assembly are in red, and are therefore minimum intron lengths.

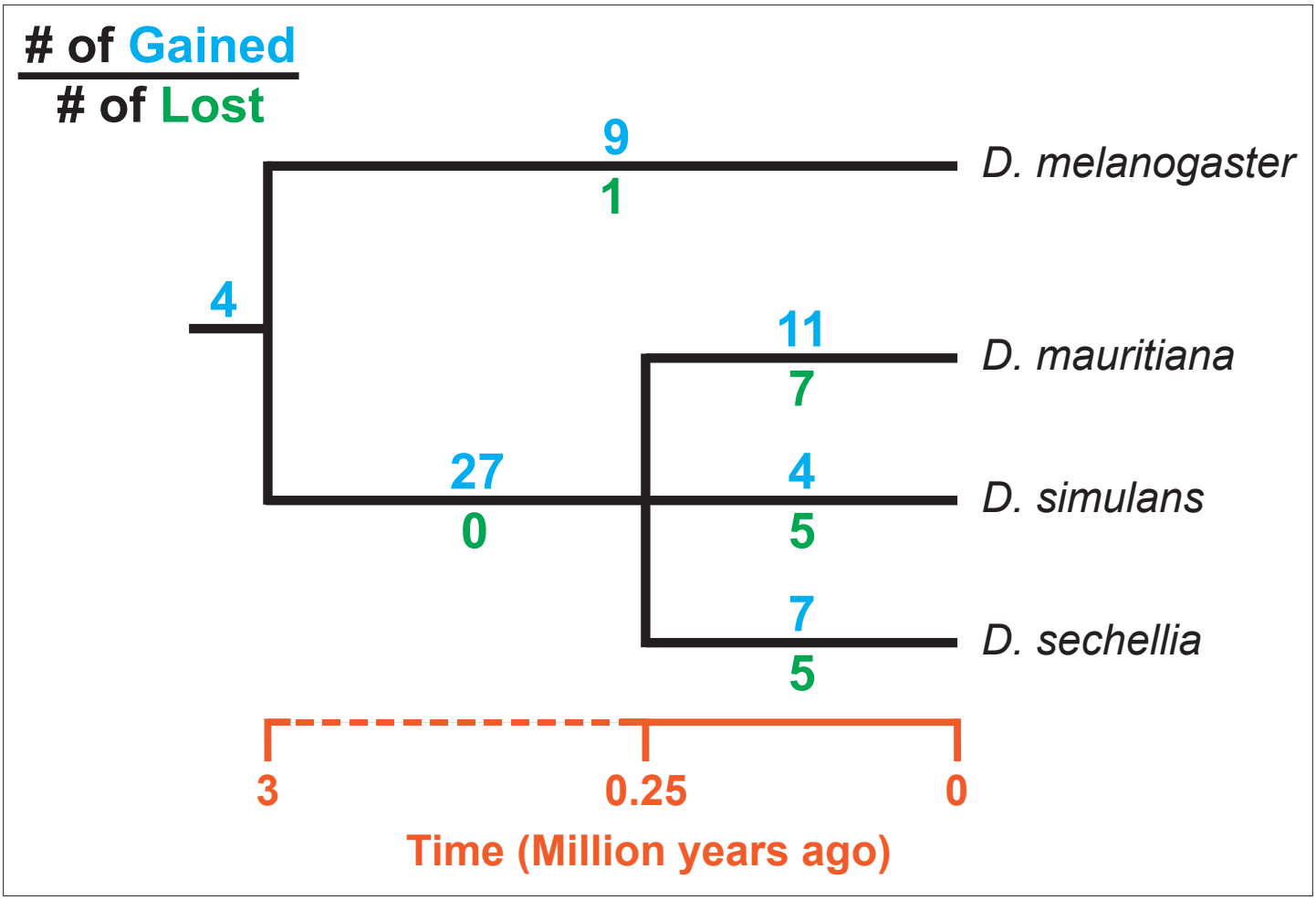

Figure 4. Turnover of new duplications to Y chromosomes in *D. melanogaster* and three species in the *D. simulans* clade. Using phylogenetic analyses, we inferred the evolutionary histories of new Y-linked duplications. The blue and green numbers represent the number of independent duplications and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

frames (< 100 amino acids; ***Supplementary file 7***). We also detect intrachromosomal duplications of these Y-linked pseudogenes (***Supplementary file 7***), suggesting a high duplication rate within these Y chromosomes.

Most new Y-linked duplications in *D. melanogaster* and the *D. simulans* clade are from genes with presumed functions in chromatin modification, cell division, and sexual reproduction (***Supplementary file 8***), consistent with other *Drosophila* species (***Bachtrog et al., 2019***; ***Mahajan and Bachtrog, 2017***). While Y-linked duplicates of genes with these functions may be selectively beneficial, 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 (***Greil and Ahmad, 2012***; ***Mahadevaraju et al., 2021***; ***Hess and Meyer, 1968***).

The evolution of new Y-linked gene families

Ampliconic gene families are found on Y chromosomes in multiple *Drosophila* species (***Ellison and Bachtrog, 2019***). We discovered two new gene families that have undergone extensive amplification on *D. simulans* clade Y chromosomes (***Figure 5***). Both families appear to encode functional protein-coding genes with complete open reading frames and high expression in mRNA-seq data (***Supplementary file 9***) and have 36–146 copies in each species' Y chromosome. We also confirm that >90%

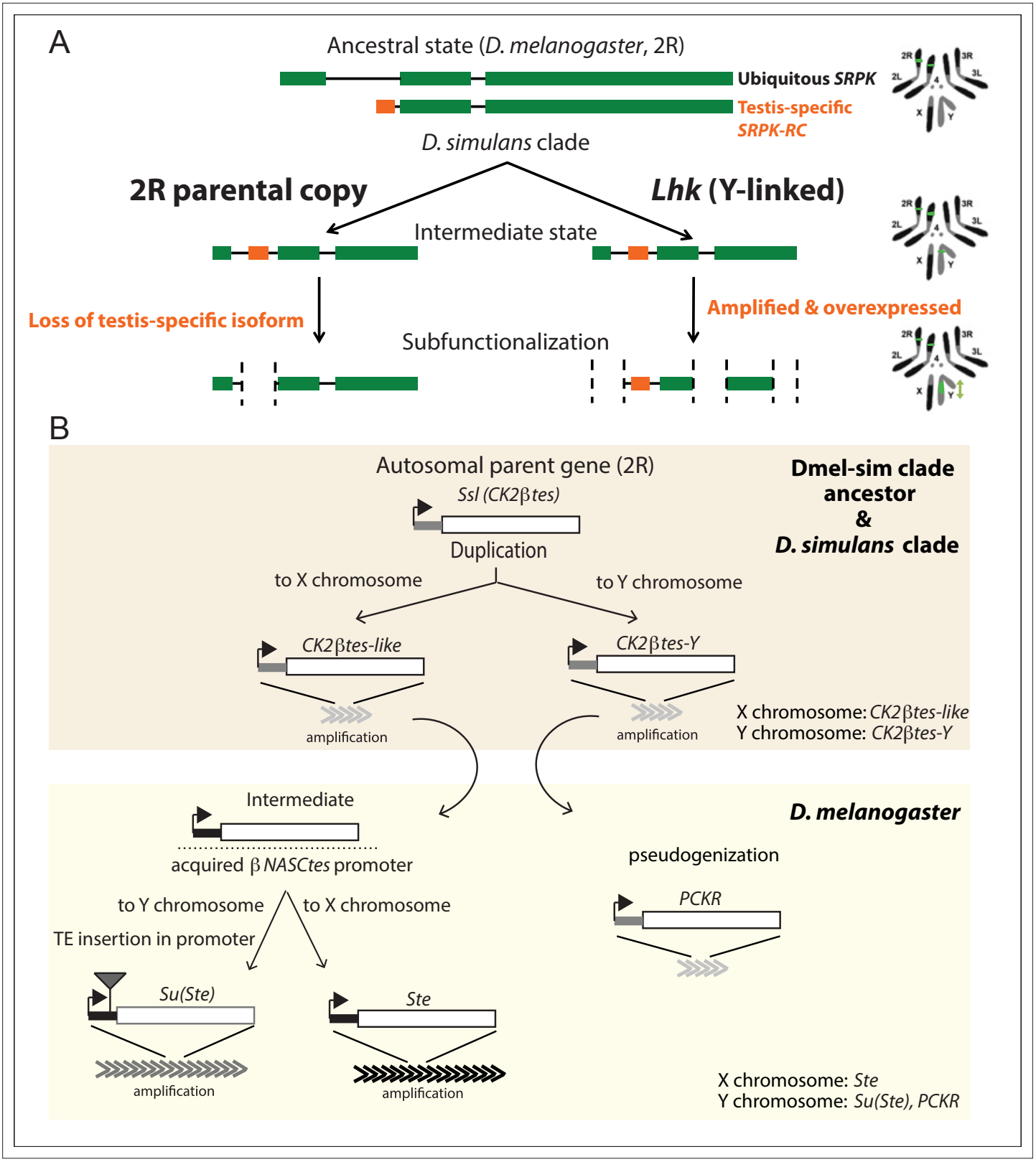

Figure 5. The history of Y-linked ampliconic genes. (**A**) Schematic showing the inferred evolutionary history of *SRPK-Y*. *SRPK* duplicated to the ancestral Y chromosome in the *D. simulans* clade. The Y-linked copy (*Lhk*) retained an exon with testis-specific expression, which was lost in the parental copy on 2R. The Y-linked copy (*Lhk*) further duplicated and increased their expression in testes. (**B**) Schematic showing the inferred evolutionary history of sex-linked *Ssl/CK2ßtes* paralogs. In the *D. melanogaster – D. simulans* clade ancestor, the autosomal gene *Ssl/CK2ßtes* duplicated from chromosome

Figure 5 continued on next page

Figure 5 continued

2R to the sex chromosome and independently amplified into the multi-copy gene families *CK2ßtes-like* on the X chromosome and *CK2ßtes-Y* on the Y chromosomes (shaded orange box). The gene structures are maintained in the *D. simulans* clade species, but not in *D. melanogaster*. In the *D. melanogaster* lineage (shaded yellow box), *CK2ßtes-Ys* became pseudogenes (*PCKR*) and *CK2ßtes-like* acquired a promoter from *ßNASCtes* to create a chimeric gene. Subsequent duplication of the chimeric gene to the X chromosome gave rise to the X-linked *Ste* loci in *D. melanogaster*. Duplication of the chimeric gene to the Y chromosome, with a subsequent TE insertion in the promoter and amplification event, gave rise to the Y-linked *Su(Ste)* loci in *D. melanogaster*.

of the variants in our assembled Y-linked gene families are represented in Illumina DNA-seq data (Appendix 1).

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 (***Loh et al., 2012***). We therefore hypothesize that the relocation of the testis-specific isoform to the *D. simulans* clade Y chromosomes may have resolved intralocus sexual antagonism over these two functions. Our phylogenetic analysis identified two subfamilies of *Lhk* that we designate *Lhk-1* and *Lhk-2* (***Figure 6A***). 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 with this observation, we detect two to three *Lhk* foci on Y chromosomes in the *D. simulans* clade using FISH (***Figure 6A and C*** and ***Figure 1—figure supplements 3–4***).

The second amplified gene family comprises both X-linked and Y-linked duplicates of the *Ssl* gene located on chromosome 2 R; it is unclear whether the X- or Y-linked copies originated first (***Figure 5B***). The X-linked copies are known as *CK2ßtes-like* in *D. simulans* (***Kogan et al., 2012***). The Y-linked copies are also found in *D. melanogaster* but are degenerated and have little or no expression (***Chang and Larracuente, 2019***; ***Danilevskaya et al., 1991***), 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, ***Supplementary file 9***) 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* (***Supplementary file 9***; ***Kogan et al., 2012***). The Y-linked copies in *D. melanogaster, Su(Ste),* are known to be a source of piRNAs (***Aravin et al., 2004***). 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 (***Supplementary file 10***).

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 (***Rozen et al., 2003***; ***Chang and Larracuente, 2019***; ***Ohta, 1984***; ***Backström et al., 2005***). 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* (***Chang and Larracuente, 2019***; ***Supplementary file 11***). These higher gene conversion rates compared to the other chromosomes might be a shared feature of Y chromosomes across taxa (***Rozen et al., 2003***).

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 (***Figure 6A and B***; ***Table 2***; ***Yang, 1997***). 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 (***Figure 6A***; ***Table 2***; ***Figure 6—figure supplement 1***; ***Supplementary file 12***). 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; ***Figure 6—figure supplement 2*** and ***Supplementary file 13***). Both

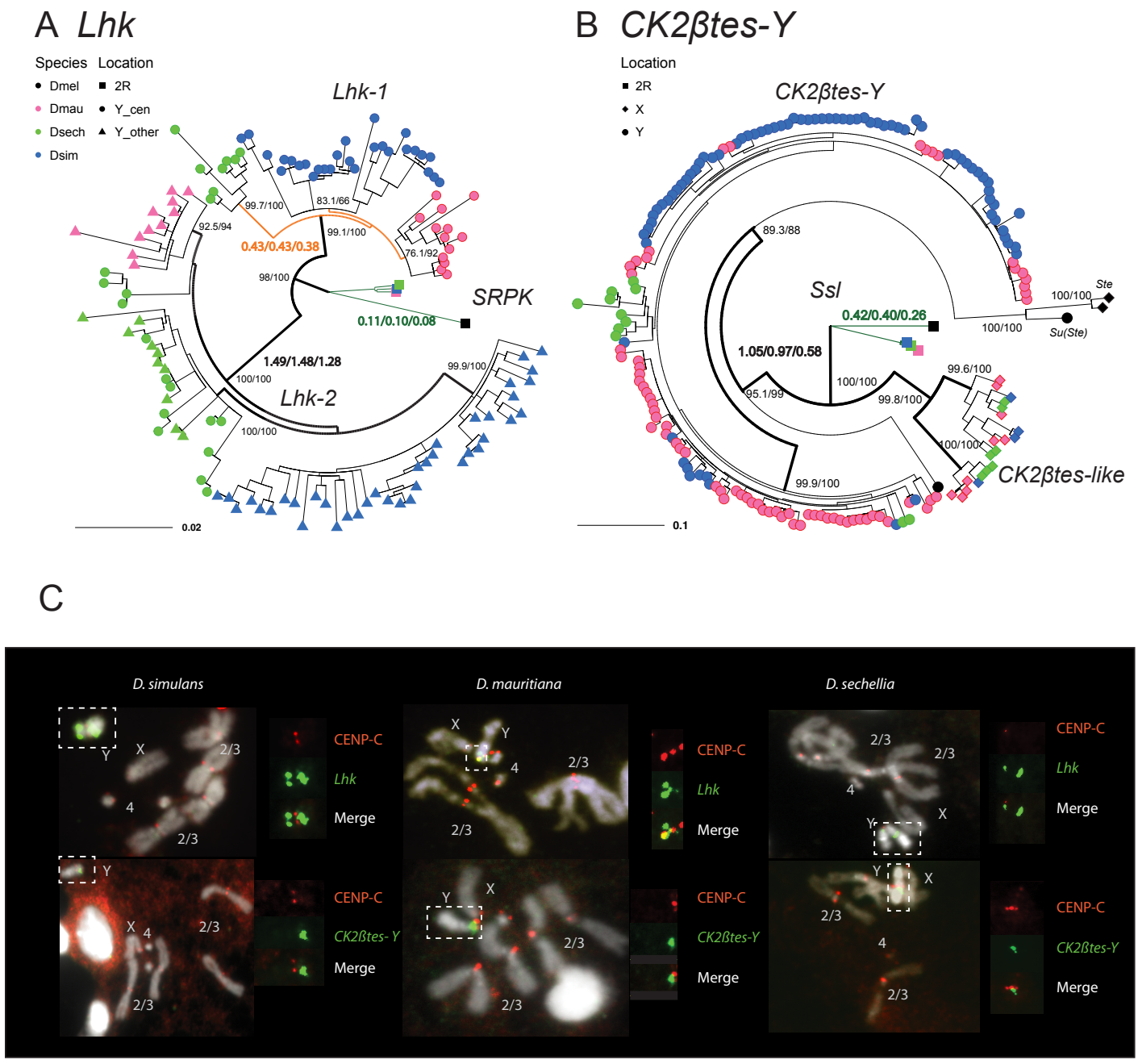

Figure 6. The rapid evolution and gene conversion of Y-linked ampliconic genes. (**A**) The inferred maximum likelihood phylogeny for *Lhk*. Node labels indicate SH-aLRT and ultrafast bootstrap (*e.g.* 100/100) or rates of protein evolution from PAML with CodonFreq = 0,1, or 2 (*e.g.* 1.01/1.02/1.03) (***Figure 6—figure supplement 1*** and ***Figure 6—figure supplement 3***). *Lhk* shows evidence for positive selection (branch tests and branch-site tests with $\omega > 1$) 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. (**B**) Same as A 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, P0.05; ***Supplementary file 12*** and ***Supplementary file 14***). (**C**) Cytological location of Y-linked gene families detected using Immunolabeling with fluorescence in situ hybridization (immunoFISH) for the centromere (CENP-C antibody, red signal). On the Y chromosomes, *Lhk* FISH signals suggest that this gene family occurs in 2–3 cytological locations (green signal), with one near the centromere. *CK2ßtes-Y* FISH signals are only located near centromeres. Based on our analysis of sequence information, we suggest that most *Lhk-1* copies are located near *CK2ßtes-Y* and the centromere.

The online version of this article includes the following figure supplement(s) for figure 6:

Figure 6 continued on next page

Figure 6 continued

Figure supplement 1. The phylogeny of *Lhk* used in PAML analyses.

Figure supplement 2. The expression of different copies from *Lhk* and *CK2ßtes-Y* gene families.

Figure supplement 3. The phylogeny of *CK2ßtes-Y* used in PAML analyses.

Lhk gene families are expressed two- to seven-fold higher than the ancestral copy on 2R in the same species, and 1.9–64-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* (***Supplementary file 9***). 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 (***Dover, 1982***).

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; ***Figure 6B***; ***Table 2***; ***Figure 6—figure supplement 3***; ***Supplementary file 14***). 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; ***Figure 6B***; ***Table 2***; ***Figure 6—figure supplement 3***; ***Supplementary file 14***). 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 (***Malone et al., 2015***; ***Palumbo et al., 1994***; ***Meyer et al., 2004***). 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 (***Chan et al., 2010***). Compared to NHEJ, MMEJ is more error-prone and can result in translocations and duplications (***McVey and Lee, 2008***). Preferential use of MMEJ instead of NHEJ could contribute to the high duplication rate and extensive genome rearrangements that we observe 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 21 putative pseudogenes. Both NHEJ and MMEJ can generate indels, but NHEJ usually produces smaller indels (1–3 bp) compared to MMEJ (> 3 bp) (***McVey and Lee, 2008***; ***Chang et al., 2017***). We also cataloged short stretches of homology 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; ***Huang et al., 2014***) and *D. simulans* (***Signor et al., 2018***). 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 (25.1% of Y-linked indels are ≥10 bp deletions; ***Supplementary file 15***) 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; ***Figure 4A***; ***Supplementary file 15***). The pattern of excess large deletions is shared in the three *D. simulans* clade species Y chromosomes but is not obvious in *D. melanogaster* (***Figure 7B***). However, because most (36/41) *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 17 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), consistent with a previous study using TE sequences

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)***
one ω	0.17			–3250.74									
two ω†	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‡	0.11	1.49	0.43	–3216.30	3.92	0.05							
CK2βtes													
one ω	0.35			–3295.01									
two ω§	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‡	0.20	0.42	1.05	–3266.33	11.35	7.56E-04							

*See ***Supplementary files 12 and 14*** for all sites.
†Autosomal and Y lineage have protein evolution of ω1 and ω2, respectively.
‡See ***Supplementary files 12 and 14***, Figure 6—figure supplement 1 and Figure 6—figure supplement 3 for the assignment of lineages.
§Autosomal and sex chromosomal (X and Y) have protein evolution of ω1 and ω2, respectively.

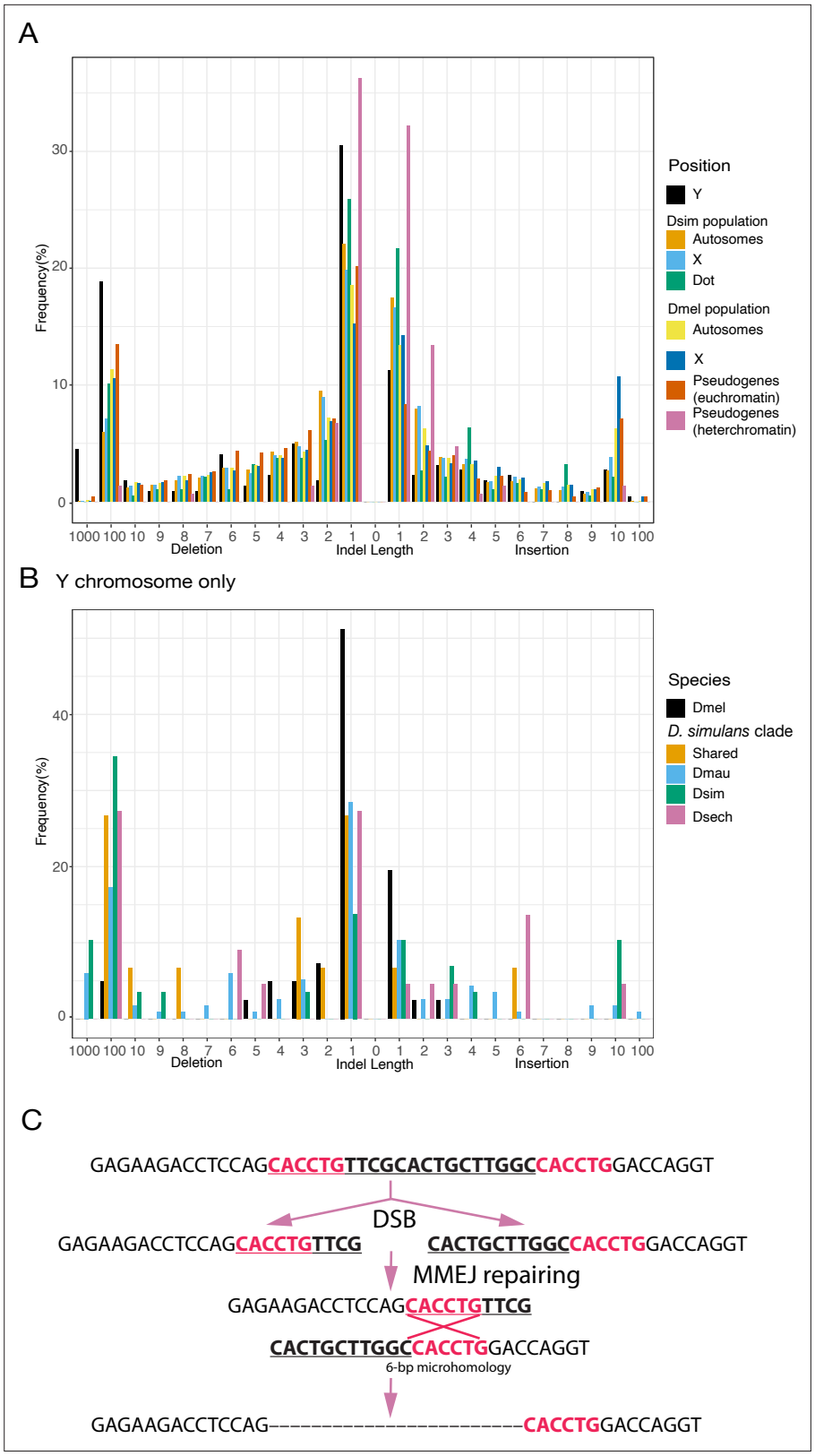

Figure 7. An excess of large deletions on Y chromosomes compared to population data suggests a preference for MMEJ. (**A**) We compared the size of 223 indels on 21 recently duplicated Y-linked genes in *D. melanogaster* and the *D. simulans* clade species to the indels polymorphic in the *D. melanogaster* and *D. simulans* populations. For the indels in *D. melanogaster* and *D. simulans* populations, we separated them based on their location,

Figure 7 continued on next page

Figure 7 continued

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). We also surveyed indel polymorphism in pseudogenes in *D. melanogaster* using population data. (**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 online version of this article includes the following figure supplement(s) for figure 7:

Figure supplement 1. The abundance of repetitive elements on Y chromosomes of *D. melanogaster* and the *D. simulans* clade species.

Figure supplement 2. The correlation of TE abundance between Y chromosomes and other chromosomes of *D. melanogaster* and the *D. simulans* clade.

Figure supplement 3. The length of LTR retrotransposons between Y chromosomes and other chromosomes of *D. melanogaster* and the *D. simulans* clade.

across different chromatin domains (***Blumenstiel et al., 2002***). We also found fewer large deletions (2/149 indels are ≥10 bp in 400 kb alignments; ***Figure 7A***) in heterochromatic pseudogenes using 19 long-read (Pacbio or nanopore) assemblies. The enrichment of 1 bp indels (101/149; ***Figure 7A***) in heterochromatic pseudogenes might represent sequencing errors in long-read assemblies (***Weirather et al., 2017***). 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 (> 7 bp) 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; ***Supplementary files 15–16***), consistent with most being a consequence of MMEJ-mediated repair.

The satellite sequence composition of Y chromosomes differs between species (***Jagannathan et al., 2017***; ***Wei et al., 2018***; ***Cechova et al., 2019***). 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 (***Figure 1—figure supplements 3–4*** and ***Supplementary file 6***). These satellites only span a few kilobases of sequences (5,515–26,119 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*. This is consistent with other reports of novel satellites arising from TEs (***Dias et al., 2014***). 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 (***McVey and Lee, 2008***). Short-tandem repeats can be further amplified via saltatory replication or unequal crossing-over between sister chromatids.

Consistent with findings in other species (***Peichel et al., 2019***; ***Chang and Larracuente, 2019***), we find an enrichment of LTR retrotransposons on the *D. simulans* clade Y chromosomes relative to the rest of the genome (***Supplementary file 17***). Interestingly, we find that the Y-linked LTR retrotransposons also turn over between species (***Figure 7—figure supplement 1*** and ***Supplementary file 18***). We find a positive correlation 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; ***Figure 7—figure supplement 2*** and ***Supplementary file 18***). 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; ***Figure 7—figure supplement 2*** and ***Supplementary file 18***), 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 (***Figure 7—figure supplement 3***), 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* (***Figure 7—figure supplement 3***).

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 (***Berloco et al., 2005***; ***Abad et al., 2004***; ***Agudo et al., 1999***). We propose that four pieces of evidence suggest DSBs on Y chromosomes may be preferentially repaired using the MMEJ pathway. First, Y-linked sequences are generally 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 (***Chang et al., 2017***), is excluded from HP1a-rich regions of chromosomes (***Chiolo et al., 2011***). The Ku complex also binds telomeres and might prevent telomere fusions (***Melnikova et al., 2005***; ***Samper et al., 2000***), 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 (***McVey and Lee, 2008***; ***Katsura et al., 2007***). Fourth, we show that Y chromosomes have high duplication and gene conversion rates, and larger deletion sizes than other genomic regions (***Figure 7***), consistent with a preference for MMEJ to repair Y-linked DSBs (***McVey and Lee, 2008***).

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* (***Figures 2A and 7***). *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 (***Roy et al., 2005***; ***Lohe and Roberts, 2000***; ***Lohe and Roberts, 1990***), and *D. simulans* also expresses more heterochromatin-modifying factors, such as *Su(var)*s and *E(var)*s (***Lee and Karpen, 2017***), 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 disproportionately impact Y-bearing sperm. Consistent with this prediction, a previous study showed that male *D. melanogaster* with a deficient MMEJ pathway (*DNApol theta* mutants) sire female-biased offspring (***McKee et al., 2000***). 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 (***McKee et al., 2000***), 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 (***Brown and Bachtrog, 2017***; ***Dimitri and Pisano, 1989***; ***Henikoff, 1996***; ***Gatti and Pimpinelli, 1992***). Therefore, retrotransposons located in heterochromatin might have higher activities in males due to the presence of Y-linked heterochromatin (***Brown and Bachtrog, 2017***; ***Henikoff, 1996***), although the genomic distribution of heterochromatin during spermatogenesis is unknown. We find that, like *D. melanogaster* (***Chang and Larracuente, 2019***), *D. simulans* clade Y chromosomes are enriched in retrotransposons relative to the rest of the genome; however, Y chromosomes from even the closely related *D. simulans* clade species harbor distinct retrotransposons (***Figure 7—figure supplement 1*** and ***Supplementary file 18***), indicating that some TEs may have rapidly shifted their insertion preference. This preference might benefit the TEs because Y-linked TEs might be expressed during spermatogenesis (***Lawlor et al., 2021***). On the other hand, Y chromosomes can be a significant source of small RNAs that silence repetitive elements during spermatogenesis—for example, *Su(Ste)* piRNAs in *D. melanogaster* (***Quénerch'du et al., 2016***; ***Aravin et al., 2001***) —and thus may also contribute to TE suppression. If Y chromosomes contribute to piRNA or siRNA production (*e.g.* have piRNA clusters ***Chen et al., 2021***; ***Aravin et al., 2001***), 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.

Genes may adapt to the Y chromosome after residing there for millions of years (**Wakimoto and Hearn, 1990**; **Hearn et al., 1991**). While most genes that move to the Y chromosome quickly degenerate (**Tobler et al., 2017**; **Carvalho et al., 2015**), 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 (**Dupim et al., 2018**). We identified many Y-linked duplicates in the ~15 Mb of Y chromosome that we surveyed in each species. Future improvements in genomic sequence data and assemblies may recover additional Y-linked duplicates among the unassembled satellite-rich sequences. 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 (**Soh et al., 2014**; **Bachtrog et al., 2019**; **Brashear et al., 2018**; **Ellison and Bachtrog, 2019**; **Hughes et al., 2010**; **Mueller et al., 2008**). Some ampliconic gene families are found on both the X and Y chromosomes (**Ellison and Bachtrog, 2019**; **Malone et al., 2015**; **Cocquet et al., 2012**; **Kruger et al., 2019**; **Lahn and Page, 2000**). 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 (**Cocquet et al., 2012**). We propose that Y-linked gene amplification in the *D. simulans* clade initially occurred due to an arms race and was 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* (**Kogan et al., 2012**), which is also hypothesized to play a role in sex-chromosome meiotic drive (**Hurst, 1992**). 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 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 (**Cocquet et al., 2012**; **Kruger et al., 2019**). In contrast, the Y-linked *Su(Ste)* copies in *D. melanogaster* produce small RNAs that suppress the X-linked *Stellate* (**Aravin et al., 2004**). 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 (**Soh et al., 2014**; **Rozen et al., 2003**; **Hughes and Page, 2015**; **Bachtrog et al., 2019**; **Tobler et al., 2017**; **Peichel et al., 2019**; **Brashear et al., 2018**; **Hall et al., 2016**). 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 (**Backström et al., 2005**; **Komissarov et al., 2018**). 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 (**Bachtrog, 2020**; **Rogers, 2021**).

Materials and methods

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

We applied the heterochromatin-sensitive assembly pipeline from **Chang and Larracuente, 2019**. 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 **Chakraborty et al.,**

2021, respectively. We then assembled these reads using Canu v1.3 and FALCON v0.5.0 combined the parameter tuning method on two error rates, eM and eg, in bogart to optimize the assemblies. We first made the Canu assemblies using the parameters 'genomeSize = 30 m stopOnReadQuality = false corMinCoverage = 0 corOutCoverage = 100 ovlMerSize = 31' and 'genomeSize = 30 m 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 (***Chakraborty et al., 2016***). 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. 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 ***Chakraborty et al., 2021***. We polished the resulting assemblies once with Quiver using PacBio reads (SMRT Analysis v2.3.0; ***Chin et al., 2013***) and ten times with Pilon v1.22 (***Walker et al., 2014***) 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' (***Ruan and Li, 2020***) and Flye v2.4.2 (***Kolmogorov et al., 2019***) 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 14,520f819a1e5dd221cc16553cf5b5269227b0a3) with parameters 'genomeSize = 5 m useGrid = false stopOnReadQuality = false corMinCoverage = 0 corOutCoverage = 100'. To detect other symbiont-derived sequences in our assemblies, we used Blast v2.7.1+ (***Altschul et al., 1990***) with blobtools (v1.0; ***Laetsch and Blaxter, 2017***) 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 for *D. mauritiana*, *D. simulans*, and *D. sechellia*, respectively. We designated and removed contigs homologous to bacteria and fungi in subsequent analyses (***Supplementary file 19***).

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_2O. The DNA was then treated with 1 µL 10 mg/mL RNaseA (Invitrogen) at 37 °C for 1 hr and was re-diluted in 100 µL ddH_2O 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 ***Chang and Larracuente, 2019***. In short, we mapped the male and female reads separately using BWA (v0.7.15; ***Li and Durbin, 2010***) and called the coverage of uniquely mapped reads per site with samtools (v1.7; -Q 10 ***Li et al., 2009***). 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 regions with known location (***Supplementary file 2***) 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 (***Chakraborty et al., 2021***). We collected transcripts and translated sequences from *D. melanogaster* (r6.14) and transcript sequences from *D. simulans* ***Nouhaud, 2018*** using IsoSeq3 (***Gordon et al., 2015***). We mapped these sequences to each assembly to generate annotations using maker2 (v2.31.9; ***Holt and Yandell, 2011***). We further mapped the transcriptomes using Star 2.7.3 a 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 (***Pertea et al., 2015***). 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; ***Altschul et al., 1990***).

We used RepeatMasker v4.0.5 (***Smit et al., 2013***) with our custom library to annotate the assemblies using parameter '-s.' Our custom library is modified from ***Chakraborty et al., 2021***, by adding the consensus sequence of *Jockey-3* from *D. melanogaster* to replace its homologs (*G2* in *D. melanogaster* and *Jockey-3* in *D. simulans*; ***Chang et al., 2019***). We extracted the sequences and copies of TEs and other repeats using scripts modified from ***Bailly-Bechet et al., 2014***. To annotate tandem repeats in assemblies, we used TRFinder (v4.09; ***Benson, 1999***) with parameters '2 7 7 80 10 100 2000 -ngs -h'. We also used kseek (***Wei et al., 2018***) 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* (***Supplementary file 20***; no available transcriptome from *D. sechellia*). We used Stringtie 2.0.3 (***Pertea et al., 2015***) 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 (***Langmead and Salzberg, 2012***) 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 (***Patro et al., 2017***) 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 (***Langmead, 2010***) 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; ***Li and Durbin, 2010***). We used ABRA v2.22 (***Mose et al., 2019***) to improve the alignments around the indels of these two gene families. We used samtools (v1.7; ***Li et al., 2009***) 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; ***Li et al., 2009***), 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 50 X 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; ***Huang et al., 2012***). 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 ***Larracuente and Ferree, 2015*** and immunostaining with FISH (immune-FISH) in brain cells following the protocol from ***Pimpinelli et al., 2011*** and ***Chang et al., 2019***. Briefly, we dissected brains from third instar larva in 1 X 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.

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/4 X 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 ***Supplementary file 3***. 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+ (***Altschul et al., 1990***) to extract the sequences of Y-linked duplications and conserved Y-linked genes from the genome. We only used high-quality sequences 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 (***Kearse et al., 2012***). 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 1000 replicates using Geneious v8.1.6 (***Kearse et al., 2012***) to infer their phylogenies. We also measured the length and microhomology in 223 indels from 21 Y-linked duplications using these alignments (***Supplementary file 15***). 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 ***Huang et al., 2014***) and *D. simulans* (***Signor et al., 2018***) populations from ***Chakraborty et al., 2021***. For *Lhk* and *CK2ßtes-Y*, we constructed phylogeny using iqtree 1.6.12 (***Nguyen et al., 2015***; ***Hoang et al., 2018***) 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 (***Figure 6—figure supplement 1*** and ***Figure 6—figure supplement 3***).

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 (***Kumar et al., 2018***; ***Stecher et al., 2020***) using the maximal likelihood method and G + I model with GTR model. We conducted branch and branch-site models tests in 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 (***Paradis et al., 2004***).

We used compute 0.8.4 (***Thornton, 2003***) to calculate Rmin and population recombination rates based on linkage disequilibrium (***Hudson, 1987***; ***Hudson and Kaplan, 1985***) and gene similarity. We included sites with indel polymorphisms in these analyses to increase the sample size (558–1544 bp alignments). We also reanalyzed data from ***Chang and Larracuente, 2019*** 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 (***Hudson, 1987***). We therefore also reported the lower bound on the gene conversion rate using Rmin (***Hudson and Kaplan, 1985***).

GO term analysis

We used PANTHER (Released 20190711; ***Mi et al., 2019***) 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 (13,767) in *D. melanogaster* as background.

Indel analyses

We downloaded the SNP calls (vcf files) from population genomic data in North Carolina of *D. melanogaster* (DGRP ***Huang et al., 2014***) and California of *D. simulans* (***Signor et al., 2018***). We then used vcftools (***Danecek et al., 2011***) to remove the low-quality SNPs using parameters '--maf 0.1 --keep-only-indels --min-alleles 2 --max-alleles 2 --recode'. We additionally filtered out the potential mismapped regions with '--max-missing-count 20' in *D. melanogaster* or '--max-missing-count 17' in *D. simulans*. Lastly, we analyzed the SNPs in the specific regions using bedtools intersect (***Quinlan and Hall, 2010***) with gene annotation files (dmel-r5.57 or dsim annotation from maker2 v2.31.9; ***Holt and Yandell, 2011***). For the heterochromatic pseudogenes, we download 18 long-read polished assemblies from NCBI (***Supplementary file 20***). We then used blastn to get sequences of pseudogenes from the population, aligned, and surveyed their indel lengths. All the alignments for our indel assignment are available in the GitHub repository (https://github.com/LarracuenteLab/simclade_Y; ***Chang, 2022***; copy archived at swh:1:rev:b1939db576cb1616094a59775a38014a7d61eb7f) and the Dryad digital repository (https://doi.org/10.5061/dryad.280gb5mr6).

Acknowledgements

This work was funded by the National Institutes of Health (NIH) (R35GM119515 to AML and R01GM123194 to CDM), National Science Foundation (NSF MCB 1844693) to AML and funding from the University of Nebraska-Lincoln to CDM. AML was supported by a Stephen Biggar and Elisabeth Asaro fellowship in Data Science. C-HC 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. JJ 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 Larracuente 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.

Additional information

Funding

Funder	Grant reference number	Author
National Institute of General Medical Sciences	R35GM119515	Amanda M Larracuente
National Institute of General Medical Sciences	R01GM123194	Colin D Meiklejohn
National Science Foundation	MCB 1844693	Amanda M Larracuente
Damon Runyon Cancer Research Foundation	DRG: 2438-21	Ching-Ho Chang
College of Arts and Sciences, University of Nebraska-Lincoln		Colin D Meiklejohn

Funder	Grant reference number	Author
University of Rochester		Amanda M Larracuente Ching-Ho Chang
Ministry of Education, Taiwan		Ching-Ho Chang

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Ching-Ho Chang, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing; Lauren E Gregory, Investigation, Validation, Writing – review and editing; Kathleen E Gordon, Investigation, Writing – review and editing; Colin D Meiklejohn, Data curation, Formal analysis, Funding acquisition, Investigation, Validation, Visualization, Writing – review and editing; Amanda M Larracuente, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing

Author ORCIDs

Ching-Ho Chang http://orcid.org/0000-0001-9361-1190
Colin D Meiklejohn http://orcid.org/0000-0003-2708-8316
Amanda M Larracuente http://orcid.org/0000-0001-5944-5686

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.75795.sa1
Author response https://doi.org/10.7554/eLife.75795.sa2

Additional files

Supplementary files

- Supplementary file 1. 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.
- Supplementary file 2. 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.
- Supplementary file 3. Probe and primer information.
- Supplementary file 4. 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.
- Supplementary file 5. 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.
- Supplementary file 6. 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.
- Supplementary file 7. 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.
- Supplementary file 8. Enriched GO terms in Y-linked duplicated genes in *D. melanogaster* and the *D. simulans* clade. We identified GO terms associated with genes that recently duplicated to the Y chromosome listed in ***Supplementary file 7*** using PANTHER (Released 20190711; [163]). We listed

all GO terms significantly enriched in the duplication (FDR < 0.05).

- Supplementary file 9. 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.
- Supplementary file 10. The number of small RNA reads mapped to the repetitive sequences and Y-linked gene families in the *D. simulans* clade.
- Supplementary file 11. 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).
- Supplementary file 12. 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.
- Supplementary file 13. 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 ***Supplementary file 21*** for detailed information.
- Supplementary file 14. 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.
- Supplementary file 15. 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.
- Supplementary file 16. 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.
- Supplementary file 17. 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.
- Supplementary file 18. 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.
- Supplementary file 19. The Illumina coverage and blast result for each contig in the *D. simulans* clade. We used Blast v2.7.1+ [135] with blobtools (v1.0; [136]) 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.
- Supplementary file 20. The summary of reads data used in this study.
- Supplementary file 21. 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.
- Transparent reporting form

Data availability

Genomic DNA sequence reads are in NCBI's SRA under BioProject PRJNA748438. All scripts and pipelines are available in GitHub (https://github.com/LarracuenteLab/simclade_Y; copy archived

at swh:1:rev:b1939db576cb1616094a59775a38014a7d61eb7f) and the Dryad digital repository (doi:https://doi.org/10.5061/dryad.280gb5mr6).

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Chang C, Gregory L, Gordon K, Meiklejohn C, Larracuente A	2021	Genome sequencing of males in *Drosophila* simulans clade	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA748438	NCBI BioProject, PRJNA748438
Chang C, Gregory L, Gordon K, Meiklejohn CD, Larracuente A	2021	Unique structure and positive selection promote the rapid divergence of *Drosophila* Y chromosomes	https://doi.org/10.5061/dryad.280gb5mr6	Dryad Digital Repository, 10.5061/dryad.280gb5mr6

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Garrigan et al.	2012	*Drosophila* mauritiana Genome sequencing	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA158675	NCBI BioProject, PRJNA158675
Modencode S	2012	*D. melanogaster* Dissected Tissue RNASeq	https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP003905	NCBI study, SRP003905
Gerstein et al.	2014	modENCODE *D. melanogaster* Developmental Total RNA-Seq	https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP001696	NCBI study, SRP001696
Chakraborty et al.	2017	DSPR Founder Genomes	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA418342/	NCBI BioProject, PRJNA418342
Wei et al.	2018	D. melanogaster, D. simulans, D. sechellia, D. erecta, D. ananassae, D. pseudoobscura, D. persimilis, D mojavensis, and D. virilis Raw sequence reads	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA423291	NCBI BioProject, PRJNA423291
Laktionov et. al.	2018	Genome-wide profiling of gene expression and transcription factors binding reveals new insights into the mechanisms of gene regulation during *Drosophila* spermatogenesis [RNA-Seq]	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA380909	NCBI BioProject, PRJNA380909
Lin et al.	2018	*Drosophila* simulans Raw sequence reads	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA477366	NCBI BioProject, PRJNA477366
Shah et al.	2020	Novel quality metrics identify high-quality assemblies of piRNA clusters	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA618654/	NCBI BioProject, PRJNA618654
Kim BY	2021	Nanopore-based assembly of many drosophilid genomes	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA675888/	NCBI BioProject, PRJNA675888
Chakraborty et al.	2021	Transcriptome sequencing of *Drosophila* simulans clade	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA541548	NCBI BioProject, PRJNA541548

References

Abad JP, de Pablos B, Agudo M, Molina I, Giovinazzo G, Martín-Gallardo A, Villasante A. 2004. Genomic and cytological analysis of the Y chromosome of *Drosophila melanogaster*: telomere-derived sequences at internal regions. *Chromosoma* **113**:295–304. DOI: https://doi.org/10.1007/s00412-004-0318-0, PMID: 15616866

Agudo M, Losada A, Abad JP, Pimpinelli S, Ripoll P, Villasante A. 1999. 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 Research* **27**:3318–3324. DOI: https://doi.org/10.1093/nar/27.16.3318, PMID: 10454639

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**:403–410. DOI: https://doi.org/10.1016/S0022-2836(05)80360-2, PMID: 2231712

Araripe LO, Tao Y, Lemos B. 2016. Interspecific Y chromosome variation is sufficient to rescue hybrid male sterility and is influenced by the grandparental origin of the chromosomes. *Heredity* **116**:516–522. DOI: https://doi.org/10.1038/hdy.2016.11, PMID: 26980343

Aravin AA, Naumova NM, Tulin AV, Vagin VV, Rozovsky YM, Gvozdev VA. 2001. Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the *D. melanogaster* germline. *Current Biology* **11**:1017–1027. DOI: https://doi.org/10.1016/s0960-9822(01)00299-8, PMID: 11470406

Aravin AA, Klenov MS, Vagin VV, Bantignies F, Cavalli G, Gvozdev VA. 2004. Dissection of a natural RNA silencing process in the *Drosophila melanogaster* germ line. *Molecular and Cellular Biology* **24**:6742–6750. DOI: https://doi.org/10.1128/MCB.24.15.6742-6750.2004, PMID: 15254241

Bachtrog D. 2003. Protein evolution and codon usage bias on the neo-sex chromosomes of *Drosophila* miranda. *Genetics* **165**:1221–1232. DOI: https://doi.org/10.1093/genetics/165.3.1221, PMID: 14668377

Bachtrog D. 2004. Evidence that positive selection drives Y-chromosome degeneration in *Drosophila* miranda. *Nature Genetics* **36**:518–522. DOI: https://doi.org/10.1038/ng1347, PMID: 15107853

Bachtrog D. 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nature Reviews. Genetics* **14**:113–124. DOI: https://doi.org/10.1038/nrg3366, PMID: 23329112

Bachtrog D, Mahajan S, Bracewell R. 2019. Massive gene amplification on a recently formed *Drosophila* Y chromosome. *Nature Ecology & Evolution* **3**:1587–1597. DOI: https://doi.org/10.1038/s41559-019-1009-9, PMID: 31666742

Bachtrog D. 2020. The Y Chromosome as a Battleground for Intragenomic Conflict. *Trends in Genetics* **36**:510–522. DOI: https://doi.org/10.1016/j.tig.2020.04.008, PMID: 32448494

Backström N, Ceplitis H, Berlin S, Ellegren H. 2005. Gene conversion drives the evolution of HINTW, an ampliconic gene on the female-specific avian W chromosome. *Molecular Biology and Evolution* **22**:1992–1999. DOI: https://doi.org/10.1093/molbev/msi198, PMID: 15972846

Bailly-Bechet M, Haudry A, Lerat E. 2014. "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. *Mobile DNA* **5**:13. DOI: https://doi.org/10.1186/1759-8753-5-13

Bayes JJ, Malik HS. 2009. Altered heterochromatin binding by a hybrid sterility protein in *Drosophila* sibling species. *Science* **326**:1538–1541. DOI: https://doi.org/10.1126/science.1181756, PMID: 19933102

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**:573–580. DOI: https://doi.org/10.1093/nar/27.2.573, PMID: 9862982

Bergero R, Qiu S, Charlesworth D. 2015. Gene loss from a plant sex chromosome system. *Current Biology* **25**:1234–1240. DOI: https://doi.org/10.1016/j.cub.2015.03.015, PMID: 25913399

Bergero R, Gardner J, Bader B, Yong L, Charlesworth D. 2019. Exaggerated heterochiasmy in a fish with sex-linked male coloration polymorphisms. *PNAS* **116**:6924–6931. DOI: https://doi.org/10.1073/pnas.1818486116, PMID: 30894479

Berloco M, Fanti L, Sheen F, Levis RW, Pimpinelli S. 2005. Heterochromatic distribution of HeT-A- and TART-like sequences in several *Drosophila* species. *Cytogenetic and Genome Research* **110**:124–133. DOI: https://doi.org/10.1159/000084944, PMID: 16093664

Bernardo Carvalho A, Koerich LB, Clark AG. 2009. Origin and evolution of Y chromosomes: *Drosophila* tales. *Trends in Genetics* **25**:270–277. DOI: https://doi.org/10.1016/j.tig.2009.04.002, PMID: 19443075

Blumenstiel JP, Hartl DL, Lozovsky ER. 2002. Patterns of insertion and deletion in contrasting chromatin domains. *Molecular Biology and Evolution* **19**:2211–2225. DOI: https://doi.org/10.1093/oxfordjournals.molbev.a004045, PMID: 12446812

Bonaccorsi S, Pisano C, Puoti F, Gatti M. 1988. Y chromosome loops in *Drosophila melanogaster*. *Genetics* **120**:1015–1034. DOI: https://doi.org/10.1093/genetics/120.4.1015, PMID: 2465201

Bonaccorsi S, Gatti M, Pisano C, Lohe A. 1990. Transcription of a satellite DNA on two Y chromosome loops of *Drosophila melanogaster*. *Chromosoma* **99**:260–266. DOI: https://doi.org/10.1007/BF01731701, PMID: 2119983

Bozzetti MP, Massari S, Finelli P, Meggio F, Pinna LA, Boldyreff B, Issinger OG, Palumbo G, Ciriaco C, Bonaccorsi S. 1995. 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. *PNAS* **92**:6067–6071. DOI: https://doi.org/10.1073/pnas.92.13.6067, PMID: 7597082

Branco AT, Tao Y, Hartl DL, Lemos B. 2013. Natural variation of the Y chromosome suppresses sex ratio distortion and modulates testis-specific gene expression in *Drosophila* simulans. *Heredity* **111**:8–15. DOI: https://doi.org/10.1038/hdy.2013.5, PMID: 23591516

Brashear WA, Raudsepp T, Murphy WJ. 2018. Evolutionary conservation of Y Chromosome ampliconic gene families despite extensive structural variation. *Genome Research* **28**:1841–1851. DOI: https://doi.org/10.1101/gr.237586.118, PMID: 30381290

Brown E, Bachtrog D. 2017. The *Drosophila* Y Chromosome Affects Heterochromatin Integrity Genome-Wide. [bioRxiv]. DOI: https://doi.org/10.1101/156000

Carlson M, Brutlag D. 1977. Cloning and characterization of a complex satellite DNA from *Drosophila melanogaster*. *Cell* **11**:371–381. DOI: https://doi.org/10.1016/0092-8674(77)90054-x, PMID: 408008

Carvalho AB, Vicoso B, Russo CAM, Swenor B, Clark AG. 2015. Birth of a new gene on the Y chromosome of *Drosophila melanogaster*. *PNAS* **112**:12450–12455. DOI: https://doi.org/10.1073/pnas.1516543112, PMID: 26385968

Cechova M, Harris RS, Tomaszkiewicz M, Arbeithuber B, Chiaromonte F, Makova KD. 2019. High satellite repeat turnover in great apes studied with short- and long-read technologies. *Molecular Biology and Evolution* **36**:2415–2431. DOI: https://doi.org/10.1093/molbev/msz156, PMID: 31273383

Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. 2016. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Research* **44**:e147. DOI: https://doi.org/10.1093/nar/gkw654, PMID: 27458204

Chakraborty M. 2020. Evolution of Genome Structure in the *Drosophila* Simulans Species Complex. [bioRxiv]. DOI: https://doi.org/10.1101/2020.02.27.968743

Chakraborty M, Chang CH, Khost DE, Vedanayagam J, Adrion JR, Liao Y, Montooth KL, Meiklejohn CD, Larracuente AM, Emerson JJ. 2021. Evolution of genome structure in the *Drosophila simulans* species complex. *Genome Research* **31**:380–396. DOI: https://doi.org/10.1101/gr.263442.120, PMID: 33563718

Chan SH, Yu AM, McVey M. 2010. Dual roles for DNA polymerase theta in alternative end-joining repair of double-strand breaks in *Drosophila*. *PLOS Genetics* **6**:e1001005. DOI: https://doi.org/10.1371/journal.pgen.1001005, PMID: 20617203

Chang CH, Larracuente AM. 2017. Genomic changes following the reversal of a Y chromosome to an autosome in *Drosophila* pseudoobscura. *Evolution; International Journal of Organic Evolution* **71**:1285–1296. DOI: https://doi.org/10.1111/evo.13229, PMID: 28322435

Chang HHY, Pannunzio NR, Adachi N, Lieber MR. 2017. Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nature Reviews. Molecular Cell Biology* **18**:495–506. DOI: https://doi.org/10.1038/nrm.2017.48, PMID: 28512351

Chang CH, Chavan A, Palladino J, Wei X, Martins NMC, Santinello B, Chen CC, Erceg J, Beliveau BJ, Wu CT, Larracuente AM, Mellone BG. 2019. Islands of retroelements are major components of *Drosophila* centromeres. *PLOS Biology* **17**:e3000241. DOI: https://doi.org/10.1371/journal.pbio.3000241, PMID: 31086362

Chang CH, Larracuente AM. 2019. Heterochromatin-Enriched Assemblies Reveal the Sequence and Organization of the *Drosophila melanogaster* Y Chromosome. *Genetics* **211**:333–348. DOI: https://doi.org/10.1534/genetics.118.301765, PMID: 30420487

Chang CH. 2022. simclade_Y. swh:1:rev:b1939db576cb1616094a59775a38014a7d61eb7f. Software Heritage. https://archive.softwareheritage.org/swh:1:dir:73ec96265042d04d5c1c7497fe2276bd83309c6b;origin=https://github.com/LarracuenteLab/simclade_Y;visit=swh:1:snp:a9c367d00c0109078ac14c44d3c97515ce040ec4;anchor=swh:1:rev:b1939db576cb1616094a59775a38014a7d61eb7f

Charlesworth B. 1978. Model for evolution of Y chromosomes and dosage compensation. *PNAS* **75**:5618–5622. DOI: https://doi.org/10.1073/pnas.75.11.5618, PMID: 281711

Charlesworth D, Charlesworth B, Morgan MT. 1995. The pattern of neutral molecular variation under the background selection model. *Genetics* **141**:1619–1632. DOI: https://doi.org/10.1093/genetics/141.4.1619, PMID: 8601499

Charlesworth B, Charlesworth D. 2000. The degeneration of Y chromosomes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **355**:1563–1572. DOI: https://doi.org/10.1098/rstb.2000.0717, PMID: 11127901

Chen P, Kotov AA, Godneeva BK, Bazylev SS, Olenina LV, Aravin AA. 2021. piRNA-mediated gene regulation and adaptation to sex-specific transposon expression in *D. melanogaster* male germline. *Genes & Development* **35**:914–935. DOI: https://doi.org/10.1101/gad.345041.120, PMID: 33985970

Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* **10**:563–569. DOI: https://doi.org/10.1038/nmeth.2474, PMID: 23644548

Chiolo I, Minoda A, Colmenares SU, Polyzos A, Costes SV, Karpen GH. 2011. Double-strand breaks in heterochromatin move outside of a dynamic HP1a domain to complete recombinational repair. *Cell* **144**:732–744. DOI: https://doi.org/10.1016/j.cell.2011.02.012, PMID: 21353298

Cocquet J, Ellis PJI, Mahadevaiah SK, Affara NA, Vaiman D, Burgoyne PS. 2012. A genetic basis for a postmeiotic X versus Y chromosome intragenomic conflict in the mouse. *PLOS Genetics* **8**:e1002900. DOI: https://doi.org/10.1371/journal.pgen.1002900, PMID: 23028340

Connallon T, Clark AG. 2010. Gene duplication, gene conversion and the evolution of the Y chromosome. *Genetics* **186**:277–286. DOI: https://doi.org/10.1534/genetics.110.116756, PMID: 20551442

Courret C, Chang CH, Wei KHC, Montchamp-Moreau C, Larracuente AM. 2019. Meiotic drive mechanisms: lessons from Drosophila Proceedings. *Biological Sciences* **286**:20191430. DOI: https://doi.org/10.1098/rspb.2019.1430

Coyne JA. 1985. The genetic basis of Haldane's rule. *Nature* **314**:736–738. DOI: https://doi.org/10.1038/314736a0, PMID: 3921852

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. *Bioinformatics* **27**:2156–2158. DOI: https://doi.org/10.1093/bioinformatics/btr330, PMID: 21653522

Danilevskaya ON, Kurenova EV, Pavlova MN, Bebehov DV, Link AJ, Koga A, Vellek A, Hartl DL. 1991. He-T family DNA sequences in the Y chromosome of *Drosophila melanogaster* share homology with the X-linked stellate genes. *Chromosoma* **100**:118–124. DOI: https://doi.org/10.1007/BF00418245, PMID: 1672635

Dias GB, Svartman M, Delprat A, Ruiz A, Kuhn GCS. 2014. Tetris is a foldback transposon that provided the building blocks for an emerging satellite DNA of *Drosophila* virilis. *Genome Biology and Evolution* **6**:1302–1313. DOI: https://doi.org/10.1093/gbe/evu108, PMID: 24858539

Dimitri P, Pisano C. 1989. Position effect variegation in *Drosophila melanogaster*: relationship between suppression effect and the amount of Y chromosome. *Genetics* **122**:793–800. DOI: https://doi.org/10.1093/genetics/122.4.793, PMID: 2503420

Dover G. 1982. Molecular drive: a cohesive mode of species evolution. *Nature* **299**:111–117. DOI: https://doi.org/10.1038/299111a0, PMID: 7110332

Dupim EG, Goldstein G, Vanderlinde T, Vaz SC, Krsticevic F, Bastos A, Pinhão T, Torres M, David JR, Vilela CR, Carvalho AB. 2018. An investigation of Y chromosome incorporations in 400 species of *Drosophila* and related genera. *PLOS Genetics* **14**:e1007770. DOI: https://doi.org/10.1371/journal.pgen.1007770, PMID: 30388103

Ellison C, Bachtrog D. 2019. Recurrent gene co-amplification on *Drosophila* X and Y chromosomes. *PLOS Genetics* **15**:e1008251. DOI: https://doi.org/10.1371/journal.pgen.1008251, PMID: 31329593

Erhardt S, Mellone BG, Betts CM, Zhang W, Karpen GH, Straight AF. 2008. Genome-wide analysis reveals a cell cycle-dependent mechanism controlling centromere propagation. *The Journal of Cell Biology* **183**:805–818. DOI: https://doi.org/10.1083/jcb.200806038, PMID: 19047461

Fingerhut JM, Moran JV, Yamashita YM. 2019. Satellite DNA-containing gigantic introns in a unique gene expression program during *Drosophila* spermatogenesis. *PLOS Genetics* **15**:e1008028. DOI: https://doi.org/10.1371/journal.pgen.1008028, PMID: 31071079

Flynn JM, Long M, Wing RA, Clark AG. 2020. Evolutionary Dynamics of Abundant 7-bp Satellites in the Genome of *Drosophila* virilis. *Molecular Biology and Evolution* **37**:1362–1375. DOI: https://doi.org/10.1093/molbev/msaa010, PMID: 31960929

Gatti M, Pimpinelli S. 1992. Functional elements in *Drosophila melanogaster* heterochromatin. *Annual Review of Genetics* **26**:239–275. DOI: https://doi.org/10.1146/annurev.ge.26.120192.001323, PMID: 1482113

Gepner J, Hays TS. 1993. A fertility region on the Y chromosome of *Drosophila melanogaster* encodes a dynein microtubule motor. *PNAS* **90**:11132–11136. DOI: https://doi.org/10.1073/pnas.90.23.11132, PMID: 8248219

Gordon SP, Tseng E, Salamov A, Zhang J, Meng X, Zhao Z, Kang D, Underwood J, Grigoriev IV, Figueroa M, Schilling JS, Chen F, Wang Z. 2015. Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing. *PLOS ONE* **10**:e0132628. DOI: https://doi.org/10.1371/journal.pone.0132628, PMID: 26177194

Greil F, Ahmad K. 2012. Nucleolar dominance of the Y chromosome in *Drosophila melanogaster*. *Genetics* **191**:1119–1128. DOI: https://doi.org/10.1534/genetics.112.141242, PMID: 22649076

Hall AB, Papathanos PA, Sharma A, Cheng C, Akbari OS, Assour L, Bergman NH, Cagnetti A, Crisanti A, Dottorini T, Fiorentini E, Galizi R, Hnath J, Jiang X, Koren S, Nolan T, Radune D, Sharakhova MV, Steele A, Timoshevskiy VA, et al. 2016. Radical remodeling of the Y chromosome in a recent radiation of malaria mosquitoes. *PNAS* **113**:E2114–E2123. DOI: https://doi.org/10.1073/pnas.1525164113, PMID: 27035980

Hearn MG, Hedrick A, Grigliatti TA, Wakimoto BT. 1991. The effect of modifiers of position-effect variegation on the variegation of heterochromatic genes of *Drosophila melanogaster*. *Genetics* **128**:785–797. DOI: https://doi.org/10.1093/genetics/128.4.785, PMID: 1916244

Helleu Q, Courret C, Ogereau D, Burnham KL, Chaminade N, Chakir M, Aulard S, Montchamp-Moreau C. 2019. Sex-Ratio Meiotic Drive Shapes the Evolution of the Y Chromosome in *Drosophila simulans*. *Molecular Biology and Evolution* **36**:2668–2681. DOI: https://doi.org/10.1093/molbev/msz160, PMID: 31290972

Henikoff S. 1996. Dosage-dependent modification of position-effect variegation in *Drosophila* BioEssays: news and reviews in molecular, cellular and developmental biology. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* **18**:401–409. DOI: https://doi.org/10.1002/bies.950180510, PMID: 8639163

Hess O, Meyer GF. 1968. Genetic activities of the Y chromosome in *Drosophila* during spermatogenesis. *Advances in Genetics* **14**:171–223. DOI: https://doi.org/10.1016/s0065-2660(08)60427-7, PMID: 4884781

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* **35**:518–522. DOI: https://doi.org/10.1093/molbev/msx281, PMID: 29077904

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**:491. DOI: https://doi.org/10.1186/1471-2105-12-491, PMID: 22192575

Huang W, Li L, Myers JR, Marth GT. 2012. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**:593–594. DOI: https://doi.org/10.1093/bioinformatics/btr708, PMID: 22199392

Huang W, Massouras A, Inoue Y, Peiffer J, Ràmia M, Tarone AM, Turlapati L, Zichner T, Zhu D, Lyman RF, Magwire MM, Blankenburg K, Carbone MA, Chang K, Ellis LL, Fernandez S, Han Y, Highnam G, Hjelmen CE, Jack JR, et al. 2014. Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic

Reference Panel lines. *Genome Research* **24**:1193–1208. DOI: https://doi.org/10.1101/gr.171546.113, PMID: 24714809

Hudson RR, Kaplan NL. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**:147–164. DOI: https://doi.org/10.1093/genetics/111.1.147, PMID: 4029609

Hudson RR. 1987. Estimating the recombination parameter of a finite population model without selection. *Genetical Research* **50**:245–250. DOI: https://doi.org/10.1017/s0016672300023776, PMID: 3443297

Hughes JF, Skaletsky H, Pyntikova T, Graves TA, van Daalen SKM, Minx PJ, Fulton RS, McGrath SD, Locke DP, Friedman C, Trask BJ, Mardis ER, Warren WC, Repping S, Rozen S, Wilson RK, Page DC. 2010. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463**:536–539. DOI: https://doi.org/10.1038/nature08700, PMID: 20072128

Hughes JF, Page DC. 2015. The Biology and Evolution of Mammalian Y Chromosomes. *Annual Review of Genetics* **49**:507–527. DOI: https://doi.org/10.1146/annurev-genet-112414-055311, PMID: 26442847

Hurst LD. 1992. Is Stellate a relict meiotic driver? *Genetics* **130**:229–230. DOI: https://doi.org/10.1093/genetics/130.1.229, PMID: 1732164

Jagannathan M, Warsinger-Pepe N, Watase GJ, Yamashita YM. 2017. Comparative Analysis of Satellite DNA in the *Drosophila melanogaster* Species Complex. *G3: Genes, Genomes, Genetics* **7**:693–704. DOI: https://doi.org/10.1534/g3.116.035352, PMID: 28007840

Johnson NA, Perez DE, Cabot EL, Hollocher H, Wu CI. 1992. A test of reciprocal X-Y interactions as a cause of hybrid sterility in *Drosophila*. *Nature* **358**:751–753. DOI: https://doi.org/10.1038/358751a0, PMID: 1508270

Katsura Y, Sasaki S, Sato M, Yamaoka K, Suzukawa K, Nagasawa T, Yokota J, Kohno T. 2007. Involvement of Ku80 in microhomology-mediated end joining for DNA double-strand breaks in vivo. *DNA Repair* **6**:639–648. DOI: https://doi.org/10.1016/j.dnarep.2006.12.002, PMID: 17236818

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**:1647–1649. DOI: https://doi.org/10.1093/bioinformatics/bts199, PMID: 22543367

Kim KE, Peluso P, Babayan P, Yeadon PJ, Yu C, Fisher WW, Chin CS, Rapicavoli NA, Rank DR, Li J, Catcheside DEA, Celniker SE, Phillippy AM, Bergman CM, Landolin JM. 2014. Long-read, whole-genome shotgun sequence data for five model organisms. *Scientific Data* **1**:140045. DOI: https://doi.org/10.1038/sdata.2014.45, PMID: 25977796

Kogan GL, Usakin LA, Ryazansky SS, Gvozdev VA. 2012. Expansion and evolution of the X-linked testis specific multigene families in the melanogaster species subgroup. *PLOS ONE* **7**:e37738. DOI: https://doi.org/10.1371/journal.pone.0037738, PMID: 22649555

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* **37**:540–546. DOI: https://doi.org/10.1038/s41587-019-0072-8, PMID: 30936562

Komissarov AS, Galkina SA, Koshel EI, Kulak MM, Dyomin AG, O'Brien SJ, Gaginskaya ER, Saifitdinova AF. 2018. New high copy tandem repeat in the content of the chicken W chromosome. *Chromosoma* **127**:73–83. DOI: https://doi.org/10.1007/s00412-017-0646-5, PMID: 28951974

Kopp A, Frank A, Fu J. 2006. Historical biogeography of *Drosophila simulans* based on Y-chromosomal sequences. *Molecular Phylogenetics and Evolution* **38**:355–362. DOI: https://doi.org/10.1016/j.ympev.2005.06.006, PMID: 16051503

Krsticevic FJ, Schrago CG, Carvalho AB. 2015. Long-Read Single Molecule Sequencing to Resolve Tandem Gene Copies: The Mst77Y Region on the *Drosophila melanogaster* Y Chromosome. *G3: Genes, Genomes, Genetics* **5**:1145–1150. DOI: https://doi.org/10.1534/g3.115.017277, PMID: 25858959

Kruger AN, Brogley MA, Huizinga JL, Kidd JM, de Rooij DG, Hu Y-C, Mueller JL. 2019. A Neofunctionalized X-Linked Ampliconic Gene Family Is Essential for Male Fertility and Equal Sex Ratio in Mice. *Current Biology* **29**:3699–3706. DOI: https://doi.org/10.1016/j.cub.2019.08.057, PMID: 31630956

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution* **35**:1547–1549. DOI: https://doi.org/10.1093/molbev/msy096, PMID: 29722887

Laetsch DR, Blaxter ML. 2017. BlobTools: Interrogation of genome assemblies Version 1; peer review: 2 approved with reservations. *F1000Research* **6**:1287. DOI: https://doi.org/10.12688/f1000research.12232.1

Lahn BT, Page DC. 2000. A human sex-chromosomal gene family expressed in male germ cells and encoding variably charged proteins. *Human Molecular Genetics* **9**:311–319. DOI: https://doi.org/10.1093/hmg/9.2.311, PMID: 10607842

Langmead B. 2010. Aligning short sequencing reads with Bowtie Current protocols in bioinformatics / editoral board, Andreas D Baxevanis. *Current Protocols in Bioinformatics* **Chapter 11**:Unit 11.7. DOI: https://doi.org/10.1002/0471250953.bi1107s32

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**:357–359. DOI: https://doi.org/10.1038/nmeth.1923, PMID: 22388286

Larracuente AM, Clark AG. 2013. Surprising differences in the variability of Y chromosomes in African and cosmopolitan populations of *Drosophila melanogaster*. *Genetics* **193**:201–214. DOI: https://doi.org/10.1534/genetics.112.146167, PMID: 23086221

Larracuente AM, Ferree PM. 2015. Simple method for fluorescence DNA in situ hybridization to squashed chromosomes. *Journal of Visualized Experiments* **95**:52288. DOI: https://doi.org/10.3791/52288, PMID: 25591075

Lawlor MA, Cao W, Ellison CE. 2021. A transposon expression burst accompanies the activation of Y-chromosome fertility genes during *Drosophila* spermatogenesis. *Nature Communications* **12**:6854. DOI: https://doi.org/10.1038/s41467-021-27136-4
Lee YCG, Karpen GH. 2017. Pervasive epigenetic effects of *Drosophila* euchromatic transposable elements impact their evolution. *eLife* **6**:e25762. DOI: https://doi.org/10.7554/eLife.25762, PMID: 28695823
Lemeunier F, Ashburner M. 1984. Relationships within the melanogaster species subgroup of the genus *Drosophila* (Sophophora). *Chromosoma* **89**:343–351. DOI: https://doi.org/10.1007/BF00331251
Lemos B, Branco AT, Hartl DL. 2010. Epigenetic effects of polymorphic Y chromosomes modulate chromatin components, immune response, and sexual conflict. *PNAS* **107**:15826–15831. DOI: https://doi.org/10.1073/pnas.1010383107, PMID: 20798037
Lenormand T, Fyon F, Sun E, Roze D. 2020. Sex Chromosome Degeneration by Regulatory Evolution. *Current Biology* **30**:3001–3006. DOI: https://doi.org/10.1016/j.cub.2020.05.052, PMID: 32559446
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–2079. DOI: https://doi.org/10.1093/bioinformatics/btp352, PMID: 19505943
Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**:589–595. DOI: https://doi.org/10.1093/bioinformatics/btp698, PMID: 20080505
Loh BJ, Cullen CF, Vogt N, Ohkura H. 2012. The conserved kinase SRPK regulates karyosome formation and spindle microtubule assembly in *Drosophila* oocytes. *Journal of Cell Science* **125**:4457–4462. DOI: https://doi.org/10.1242/jcs.107979, PMID: 22854045
Lohe AR, Brutlag DL. 1987a. Identical satellite DNA sequences in sibling species of *Drosophila*. *Journal of Molecular Biology* **194**:161–170. DOI: https://doi.org/10.1016/0022-2836(87)90365-2, PMID: 3112413
Lohe AR, Brutlag DL. 1987b. Adjacent satellite DNA segments in *Drosophila* structure of junctions. *Journal of Molecular Biology* **194**:171–179. DOI: https://doi.org/10.1016/0022-2836(87)90366-4, PMID: 3112414
Lohe AR, Roberts PA. 1990. An unusual Y chromosome of *Drosophila simulans* carrying amplified rDNA spacer without rRNA genes. *Genetics* **125**:399–406. DOI: https://doi.org/10.1093/genetics/125.2.399, PMID: 2379820
Lohe AR, Roberts PA. 2000. Evolution of DNA in heterochromatin: the *Drosophila melanogaster* sibling species subgroup as a resource. *Genetica* **109**:125–130. DOI: https://doi.org/10.1023/a:1026588217432, PMID: 11293787
Mahadevaraju S, Fear JM, Akeju M, Galletta BJ, Pinheiro MMLS, Avelino CC, Cabral-de-Mello DC, Conlon K, Dell'Orso S, Demere Z, Mansuria K, Mendonça CA, Palacios-Gimenez OM, Ross E, Savery M, Yu K, Smith HE, Sartorelli V, Yang H, Rusan NM, et al. 2021. Dynamic sex chromosome expression in *Drosophila* male germ cells. *Nature Communications* **12**:892. DOI: https://doi.org/10.1038/s41467-021-20897-y, PMID: 33563972
Mahajan S, Bachtrog D. 2017. Convergent evolution of Y chromosome gene content in flies. *Nature Communications* **8**:785. DOI: https://doi.org/10.1038/s41467-017-00653-x, PMID: 28978907
Mahajan S, Wei KHC, Nalley MJ, Gibilisco L, Bachtrog D. 2018. De novo assembly of a young *Drosophila* Y chromosome using single-molecule sequencing and chromatin conformation capture. *PLOS Biology* **16**:e2006348. DOI: https://doi.org/10.1371/journal.pbio.2006348, PMID: 30059545
Malone CD, Lehmann R, Teixeira FK. 2015. The cellular basis of hybrid dysgenesis and Stellate regulation in *Drosophila*. *Current Opinion in Genetics & Development* **34**:88–94. DOI: https://doi.org/10.1016/j.gde.2015.09.003, PMID: 26451497
McKee BD, Karpen GH. 1990. *Drosophila* ribosomal RNA genes function as an X-Y pairing site during male meiosis. *Cell* **61**:61–72. DOI: https://doi.org/10.1016/0092-8674(90)90215-z, PMID: 2156630
McKee BD, Hong CS, Das S. 2000. 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* **109**:77–93. DOI: https://doi.org/10.1023/a:1026536200594, PMID: 11293799
McVey M, Lee SE. 2008. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends in Genetics* **24**:529–538. DOI: https://doi.org/10.1016/j.tig.2008.08.007, PMID: 18809224
Meiklejohn CD, Landeen EL, Gordon KE, Rzatkiewicz T, Kingan SB, Geneva AJ, Vedanayagam JP, Muirhead CA, Garrigan D, Stern DL, Presgraves DC. 2018. Gene flow mediates the role of sex chromosome meiotic drive during complex speciation. *eLife* **7**:e35468. DOI: https://doi.org/10.7554/eLife.35468, PMID: 30543325
Melnikova L, Biessmann H, Georgiev P. 2005. The Ku protein complex is involved in length regulation of *Drosophila* telomeres. *Genetics* **170**:221–235. DOI: https://doi.org/10.1534/genetics.104.034538, PMID: 15781709
Meyer GF. 1963. Die Funktionsstrukturen des Y-Chromosoms in den Spermatocytenkernen von *Drosophila hydei*, *D. neohydei*, *D. repleta* und einigen anderen *Drosophila-Arten*. *Chromosoma* **14**:207–255. DOI: https://doi.org/10.1007/BF00326814
Meyer GF, Hess O, Beermann W. 2004. Phasenspezifische Funktionsstrukturen in Spermatocytenkernen von *Drosophila melanogaster* und Ihre Abhängigkeit vom Y-Chromosom. *Chromosoma* **12**:676–716. DOI: https://doi.org/10.1007/BF00328946, PMID: 14473096
Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. 2019. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research* **47**:D419–D426. DOI: https://doi.org/10.1093/nar/gky1038, PMID: 30407594
Montchamp-Moreau C, Ginhoux V, Atlan A. 2001. The Y chromosomes of *Drosophila simulans* are highly polymorphic for their ability to suppress sex-ratio drive. *Evolution; International Journal of Organic Evolution* **55**:728–737. DOI: https://doi.org/10.1554/0014-3820(2001)055[0728:tycods]2.0.co;2, PMID: 11392391

Morgan AP, Pardo-Manuel de Villena F. 2017. Sequence and Structural Diversity of Mouse Y Chromosomes. *Molecular Biology and Evolution* **34**:3186–3204. DOI: https://doi.org/10.1093/molbev/msx250, PMID: 29029271

Mose LE, Perou CM, Parker JS. 2019. Improved indel detection in DNA and RNA via realignment with ABRA2. *Bioinformatics* **35**:2966–2973. DOI: https://doi.org/10.1093/bioinformatics/btz033, PMID: 30649250

Mueller JL, Mahadevaiah SK, Park PJ, Warburton PE, Page DC, Turner JMA. 2008. The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression. *Nature Genetics* **40**:794–799. DOI: https://doi.org/10.1038/ng.126, PMID: 18454149

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**:268–274. DOI: https://doi.org/10.1093/molbev/msu300, PMID: 25371430

Nouhaud P. 2018. Long-Read Based Assembly and Annotation of a *Drosophila Simulans* Genome. [bioRxiv]. DOI: https://doi.org/10.1101/425710

Ohta T. 1984. Some models of gene conversion for treating the evolution of multigene families. *Genetics* **106**:517–528. DOI: https://doi.org/10.1093/genetics/106.3.517, PMID: 6706111

Palumbo G, Bonaccorsi S, Robbins LG, Pimpinelli S. 1994. Genetic analysis of Stellate elements of *Drosophila melanogaster*. *Genetics* **138**:1181–1197. DOI: https://doi.org/10.1093/genetics/138.4.1181, PMID: 7896100

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**:289–290. DOI: https://doi.org/10.1093/bioinformatics/btg412, PMID: 14734327

Paredes S, Branco AT, Hartl DL, Maggert KA, Lemos B. 2011. Ribosomal DNA deletions modulate genome-wide gene expression: "rDNA-sensitive" genes and natural variation. *PLOS Genetics* **7**:e1001376. DOI: https://doi.org/10.1371/journal.pgen.1001376, PMID: 21533076

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**:417–419. DOI: https://doi.org/10.1038/nmeth.4197, PMID: 28263959

Peichel CL, Naftaly JA, Urton AFS, Cech JN. 2019. Assembly of a Young Vertebrate Y Chromosome Reveals Convergent Signatures of Sex Chromosome Evolution. [bioRxiv]. DOI: https://doi.org/10.1101/2019.12.12.874701

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**:290–295. DOI: https://doi.org/10.1038/nbt.3122, PMID: 25690850

Piergentili R, Bonaccorsi S, Raffa GD, Pisano C, Hackstein JHP, Mencarelli C. 2004. Autosomal control of the Y-chromosome kl-3 loop of *Drosophila melanogaster*. *Chromosoma* **113**:188–196. DOI: https://doi.org/10.1007/s00412-004-0308-2, PMID: 15338233

Piergentili R. 2007. Evolutionary conservation of lampbrush-like loops in drosophilids. *BMC Cell Biology* **8**:35. DOI: https://doi.org/10.1186/1471-2121-8-35, PMID: 17697358

Piergentili R, Mencarelli C. 2008. *Drosophila melanogaster* kl-3 and kl-5 Y-loops harbor triple-stranded nucleic acids. *Journal of Cell Science* **121**:1605–1612. DOI: https://doi.org/10.1242/jcs.025320, PMID: 18430782

Pimpinelli S, Bonaccorsi S, Fanti L, Gatti M. 2011. Immunostaining of mitotic chromosomes from *Drosophila* larval brain. *Cold Spring Harbor Protocols* **2011**:pdb.prot065524. DOI: https://doi.org/10.1101/pdb.prot065524, PMID: 21880821

Pisano C, Bonaccorsi S, Gatti M. 1993. The kl-3 loop of the Y chromosome of *Drosophila melanogaster* binds a tektin-like protein. *Genetics* **133**:569–579. DOI: https://doi.org/10.1093/genetics/133.3.569, PMID: 8454204

Quénerch'du E, Anand A, Kai T. 2016. The piRNA pathway is developmentally regulated during spermatogenesis in *Drosophila*. *RNA* **22**:1044–1054. DOI: https://doi.org/10.1261/rna.055996.116, PMID: 27208314

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841–842. DOI: https://doi.org/10.1093/bioinformatics/btq033, PMID: 20110278

Redhouse JL, Mozziconacci J, White RAH. 2011. Co-transcriptional architecture in a Y loop in *Drosophila melanogaster*. *Chromosoma* **120**:399–407. DOI: https://doi.org/10.1007/s00412-011-0321-1, PMID: 21556802

Reijo R, Lee TY, Salo P, Alagappan R, Brown LG, Rosenberg M, Rozen S, Jaffe T, Straus D, Hovatta O. 1995. Diverse spermatogenic defects in humans caused by Y chromosome deletions encompassing a novel RNA-binding protein gene. *Nature Genetics* **10**:383–393. DOI: https://doi.org/10.1038/ng0895-383, PMID: 7670487

Repping S, Skaletsky H, Brown L, Daalen SK, Korver CM, Pyntikova T. 2003. Polymorphism for a 1.6-Mb deletion of the human Y chromosome persists through balance between recurrent mutation and haploid selection. *Nature Genetics* **35**:247–251. DOI: https://doi.org/10.1038/ng1250

Rice WR. 1987a. The accumulation of sexually antagonistic genes as a selective agent promoting the evolution of reduced recombination between primitive sex chromosomes. *Evolution; International Journal of Organic Evolution* **41**:911–914. DOI: https://doi.org/10.1111/j.1558-5646.1987.tb05864.x, PMID: 28564364

Rice WR. 1987b. Genetic hitchhiking and the evolution of reduced genetic activity of the Y sex chromosome. *Genetics* **116**:161–167. DOI: https://doi.org/10.1093/genetics/116.1.161, PMID: 3596229

Rogers MJ. 2021. Y chromosome copy number variation and its effects on fertility and other health factors: a review. *Translational Andrology and Urology* **10**:1373–1382. DOI: https://doi.org/10.21037/tau.2020.04.06, PMID: 33850773

Roy V, Monti-Dedieu L, Chaminade N, Siljak-Yakovlev S, Aulard S, Lemeunier F, Montchamp-Moreau C. 2005. Evolution of the chromosomal location of rDNA genes in two *Drosophila species* subgroups: ananassae and melanogaster. *Heredity* **94**:388–395. DOI: https://doi.org/10.1038/sj.hdy.6800612, PMID: 15726113

Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC. 2003. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423**:873–876. DOI: https://doi.org/10.1038/nature01723, PMID: 12815433

Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. *Nature Methods* **17**:155–158. DOI: https://doi.org/10.1038/s41592-019-0669-3, PMID: 31819265

Russell SR, Kaiser K. 1993. *Drosophila melanogaster* male germ line-specific transcripts with autosomal and Y-linked genes. *Genetics* **134**:293–308. DOI: https://doi.org/10.1093/genetics/134.1.293, PMID: 8514138

Sackton TB, Montenegro H, Hartl DL, Lemos B. 2011. Interspecific Y chromosome introgressions disrupt testis-specific gene expression and male reproductive phenotypes in *Drosophila*. *PNAS* **108**:17046–17051. DOI: https://doi.org/10.1073/pnas.1114690108, PMID: 21969588

Samper E, Goytisolo FA, Slijepcevic P, van Buul PP, Blasco MA. 2000. Mammalian Ku86 protein prevents telomeric fusions independently of the length of TTAGGG repeats and the G-strand overhang. *EMBO Reports* **1**:244–252. DOI: https://doi.org/10.1093/embo-reports/kvd051, PMID: 11256607

Signor SA, New FN, Nuzhdin S. 2018. A Large Panel of *Drosophila simulans* Reveals an Abundance of Common Variants. *Genome Biology and Evolution* **10**:189–206. DOI: https://doi.org/10.1093/gbe/evx262, PMID: 29228179

Singh ND, Koerich LB, Carvalho AB, Clark AG. 2014. Positive and purifying selection on the *Drosophila* Y chromosome. *Molecular Biology and Evolution* **31**:2612–2623. DOI: https://doi.org/10.1093/molbev/msu203, PMID: 24974375

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

Smith AV, Orr-Weaver TL. 1991. The regulation of the cell cycle during *Drosophila* embryogenesis: the transition to polyteny. *Development* **112**:997–1008. DOI: https://doi.org/10.1242/dev.112.4.997, PMID: 1935703

Soh YQS, Alföldi J, Pyntikova T, Brown LG, Graves T, Minx PJ, Fulton RS, Kremitzki C, Koutseva N, Mueller JL, Rozen S, Hughes JF, Owens E, Womack JE, Murphy WJ, Cao Q, de Jong P, Warren WC, Wilson RK, Skaletsky H, et al. 2014. Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. *Cell* **159**:800–813. DOI: https://doi.org/10.1016/j.cell.2014.09.052, PMID: 25417157

Stecher G, Tamura K, Kumar S. 2020. Molecular Evolutionary Genetics Analysis (MEGA) for macOS. *Molecular Biology and Evolution* **37**:1237–1239. DOI: https://doi.org/10.1093/molbev/msz312, PMID: 31904846

Sun C, Skaletsky H, Rozen S, Gromoll J, Nieschlag E, Oates R, Page DC. 2000. Deletion of azoospermia factor a (AZFa) region of human Y chromosome caused by recombination between HERV15 proviruses. *Human Molecular Genetics* **9**:2291–2296. DOI: https://doi.org/10.1093/oxfordjournals.hmg.a018920, PMID: 11001932

Tao Y, Hartl DL, Laurie CC. 2001. Sex-ratio segregation distortion associated with reproductive isolation in *Drosophila*. *PNAS* **98**:13183–13188. DOI: https://doi.org/10.1073/pnas.231478798, PMID: 11687638

Tao Y, Araripe L, Kingan SB, Ke Y, Xiao H, Hartl DL. 2007. A sex-ratio meiotic drive system in *Drosophila simulans* II: an X-linked distorter. *PLOS Biology* **5**:e293. DOI: https://doi.org/10.1371/journal.pbio.0050293, PMID: 17988173

Thornton K. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* **19**:2325–2327. DOI: https://doi.org/10.1093/bioinformatics/btg316, PMID: 14630667

Tobler R, Nolte V, Schlötterer C. 2017. High rate of translocation-based gene birth on the *Drosophila* Y chromosome. *PNAS* **114**:11721–11726. DOI: https://doi.org/10.1073/pnas.1706502114, PMID: 29078298

Vogt PH, Edelmann A, Kirsch S, Henegariu O, Hirschmann P, Kiesewetter F, Köhn FM, Schill WB, Farah S, Ramos C, Hartmann M, Hartschuh W, Meschede D, Behre HM, Castel A, Nieschlag E, Weidner W, Gröne HJ, Jung A, Engel W, et al. 1996. Human Y chromosome azoospermia factors (AZF) mapped to different subregions in Yq11. *Human Molecular Genetics* **5**:933–943. DOI: https://doi.org/10.1093/hmg/5.7.933, PMID: 8817327

Wakimoto BT, Hearn MG. 1990. The effects of chromosome rearrangements on the expression of heterochromatic genes in chromosome 2L of *Drosophila melanogaster*. *Genetics* **125**:141–154. DOI: https://doi.org/10.1093/genetics/125.1.141, PMID: 2111264

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE* **9**:e112963. DOI: https://doi.org/10.1371/journal.pone.0112963, PMID: 25409509

Wang M, Branco AT, Lemos B. 2018. The Y Chromosome Modulates Splicing and Sex-Biased Intron Retention Rates in Drosophila. *Genetics* **208**:1057–1067. DOI: https://doi.org/10.1534/genetics.117.300637

Wei KH-C, Lower SE, Caldas IV, Sless TJS, Barbash DA, Clark AG. 2018. Variable Rates of Simple Satellite Gains across the *Drosophila* Phylogeny. *Molecular Biology and Evolution* **35**:925–941. DOI: https://doi.org/10.1093/molbev/msy005, PMID: 29361128

Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, Wang X-J, Buck D, Au KF. 2017. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* **6**:100. DOI: https://doi.org/10.12688/f1000research.10571.2, PMID: 28868132

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* **13**:555–556. DOI: https://doi.org/10.1093/bioinformatics/13.5.555, PMID: 9367129

Zhou J, Sackton TB, Martinsen L, Lemos B, Eickbush TH, Hartl DL. 2012. Y chromosome mediates ribosomal DNA silencing and modulates the chromatin state in *Drosophila*. *PNAS* **109**:9941–9946. DOI: https://doi.org/10.1073/pnas.1207367109, PMID: 22665801

Appendix 1

Low Pacbio coverage in heterochromatic regions

We find that PacBio coverage is lower than expected on Y chromosomes and in heterochromatic regions generally (***Figure 1—figure supplement 2***). We found a similar bias in the *D. melanogaster* genome (***Chang and Larracuente, 2019***), where the PacBio data were independently generated by a different group (***Kim et al., 2014***). While a previous paper suggests that CsCl might contribute to this bias (***Krsticevic et al., 2015***), we used Qiagen's Blood and Cell culture DNA Midi Kit for DNA extraction. Heterochromatin is underreplicated in the endoreplicated cells that undergo multiple rounds of S phase but with no cell division such as those in larval salivary glands cells (***Smith and Orr-Weaver, 1991***). Previous studies demonstrated that endoreplicated cells in the adult flies might contribute to lower coverage in Illumina sequencing data (***Flynn et al., 2020***). Therefore, these endoreplicated cells might also contribute to the bias in Pacbio coverage.

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 (***Figure 6—figure supplement 2***; ***Supplementary file 21***).

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 ***Figure 6***). The reconstructed 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; ***Supplementary files 17 and 18***, ***Figure 6***). 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.