Germline novelty through recurrent copy-number, protein, and regulatory evolution of the synaptonemal complex

Kevin H-C Wei[1,2,3,*], Ching-ho Chang[4], Kamalakar Chatla[3], Ananya Krishnapura[3], Samuel P Appiah[3], Jacki Zhang[3], Robert L Unckless[5], Justin P Blumenstiel[6], and Doris Bachtrog[3]

[1] Department of Zoology, University of British Columbia, Vancouver, BC, Canada

[2] Life Sciences Institute, University of British Columbia, Vancouver BC, Canada

[3] Department of Integrative Biology, University of California Berkeley, Berkeley, CA, USA

[4] Basic Sciences Division, Fred Hutch Cancer Center, Seattle, WA, USA

[5] Department of Molecular Biosciences, University of Kansas, Lawrence, KS, USA

[6] Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, KS, USA

* Corresponding author (wei.kevin@ubc.ca)

ABSTRACT

The synaptonemal complex (SC) is a protein-rich structure necessary to tether homologous chromosomes for meiotic recombination and faithful segregation. Despite being found in most major eukaryotic taxa implying a deep evolutionary origin, components of the complex can exhibit unusually high rates of sequence evolution, particularly in *Drosophila* where orthologs of several components could not be identified outside of the genus. To understand the cause of this paradoxical lack of conservation, we examine the evolutionary history of the SC in Drosophila, taking a comparative phylogenomic approach with high species density to circumvent obscured homology due to rapid sequence evolution. We find that in addition to elevated rates of coding evolution due to recurrent and widespread positive selection, components of the SC, in particular the central element *cona* and transverse filament *c(3)G* have diversified through tandem and retro-duplications, repeatedly generating paralogs with novel germline functions. Strikingly, independent *c(3)G* duplicates under positive selection in separate lineages both evolved to have high testes expression and similar structural changes to the proteins, suggesting molecular convergence of novel function. In other instances of germline novelty, two *cona* derived paralogs were independently incorporated into testes-expressed lncRNA. Surprisingly, the expression of SC genes in the germline is exceedingly prone to change suggesting recurrent regulatory evolution which, in many species, resulted in high testes expression even though *Drosophila* males are achiasmic. Overall, our comprehensive study recapitulates the adaptive sequence evolution of several components of the SC, and further uncovers that the lack of conservation not only extends to other modalities including copy number, genomic locale, and germline regulation, it may also underlie repeated germline novelties especially in the testes. Given the unexpected and frequently elevated testes expression in a large number of species and the ancestor, we speculate that the function of SC genes in the male germline, while still poorly understood, may be a prime target of constant evolutionary pressures driving repeated adaptations and innovations.

INTRODUCTION

Meiotic recombination, the exchange of non-sister, homologous chromosomes through physical crossovers, is an essential genetic mechanism universal to sexually reproducing eukaryotes. It allows for the shuffling of homologous alleles generating novel allelic combinations. This is necessary for maintaining nucleotide diversity and efficacy of selection; without it, chromosomes (like on the non-recombining, degenerate Y or W chromosomes) will irreversibly accumulate deleterious mutations ultimately leading populations to go extinct. At the cellular level, meiotic pairing, synapsis, and resolution of double strand breaks into crossovers are critical for stabilizing meiotic bivalents as failure is typically associated with skyrocketing aneuploidy rates. Therefore, recombination is a crucial genetic process that is necessary for reproductive fitness and species survival.

Despite the critical functionality of recombination and the deep conservation across eukaryotes, aspects of this fundamental genetic mechanism are surprisingly prone to change. Recombination rate has been repeatedly shown to vary drastically between closely related species (1). Adaptive explanations typically invoke changing environmental (e.g. temperature (2)) or genomic conditions (e.g. repeat content (3)) requiring commensurate shifts in recombination rate to maintain fitness optima (4, 5). Others have suggested intragenomic conflicts with selfish elements (6) or sexual conflict creating unstable equilibria for optimal fitness (7, 8). However, some have argued that changes in recombination rate have little impact on fitness and rate changes are the byproduct of selection on other aspects of the meiotic processes (9). Several key findings supporting the adaptive interpretation come from *Drosophila* as multiple genes in the pathways necessary for recombination show signatures of rapid evolution due to positive selection (10–13). Moreover, because recombination is absent in *Drosophila* males and the SC does not assemble during spermatogenesis (14), sexual antagonism due to sex-specific optima of crossover rates is unlikely the underlying driver of adaptive recombination evolution, at least in species with sex-specific achiasmy. Why recombination, an essential genetic mechanism, is prone to change and whether such changes are adaptive remain central questions in evolutionary genetics (15–17).

The paradox of poor conservation but crucial function is exemplified by the synaptonemal complex (SC), a crucial machinery necessary for meiotic recombination in plants, animals, and major lineages of fungi (18). It is a protein complex that acts as zippers to tether homologs together along the chromosome axes during meiotic prophase I and forms train track-like structures which have been visualized under electron microscopy across eukaryotic taxa (19, 20). The SC is mirrored along a central axis composed of central element proteins that are tethered by the transverse filaments to lateral elements on two sides anchoring into chromatin (Figure 1A) (21). This highly stereotypical configuration is found in baker's yeast, mice, and flies, indicative of an evolutionary ancient structure. Yet, despite the deep evolutionary origin and functional necessity across wide eukaryotic domains, there are many examples of unexpected exceptions. At the extreme are recombining species such as the fission yeast that entirely forego the SC (22). In another instance, the SC of *Caenorhabditis* has been reconfigured such that the transverse filament – typically a single gene in most SCs – is composed of at least four genes (23). Therefore, parts of the SC appear to be curiously flexible in composition whereby different analogous but perhaps non-homologous pieces can be recruited and replaced (24).

Consistent with this flexibility, sequences of SC components are often poorly conserved at shorter evolutionary time scales (12, 25). In *Drosophila*, positive selection appears to repeatedly drive the sequence evolution of the SC, which is composed of the central elements *corona* (*cona*) (26) and *corolla* (27), the transverse filament *c(3)G* (28), and the lateral elements *orientation disruptor* (*ord*) (29) and *c(2)M (30).* Orthologs of the central region components, *corolla*, *cona,* and *c(3)G,* could not be found outside of the *Drosophila* genus (12) either reflecting divergence so extensive that orthology is no longer recognizable, or novel acquisitions of SC components. Flexibility in SC composition may explain how these molecular transitions are possible without major fitness impacts, but cannot account for why SC genes appear to be evolving under recurrent adaptation. The recent explosion of high quality *Drosophila* species genome assemblies (31–37) offer a unique opportunity to understand the genetic and evolutionary mechanisms driving the strikingly rapid divergence of SC genes. Here, we systematically revisit the evolution of the SC in *Drosophila* by examining the genomes and

transcriptomes of 48 species scattered across the entire *Drosophila* phylogeny, with dense representation from three key species groups (*melanogaster*, *obscura*, and *immigrans*). In our exhaustive analyses, we uncovered frequent duplications of several SC components generating paralogs with novel functions, in addition to repeated sequence evolution due to positive selection. Further, we revealed unexpectedly high rates of expression divergence and regulatory turnover in not just the ovary but also the male germline, where SC genes are thought to have no function. In fact, testes-biased expression of SC genes appears to be the norm, and likely the ancestral state, suggesting SC components have crucial function in male germline, despite the absence of male recombination. Altogether our study revealed a highly dynamic evolutionary history with repeated bouts of copy number, sequence, and regulatory evolution that contribute to the overall poor conservation of SC genes. Further, the surprising transcriptional activity of SC genes in the male germline raises new possibilities for SC functions unrelated to recombination under repeated directional selection in addition to their roles in chiasmate meiosis in the female germline.

RESULTS

Poor sequence conservation and frequent duplications of components of the SC

To identify *Drosophila* SC homologs we elected to focus on only species with high quality genome assemblies with either available annotations and/or RNA-seq data (Supplementary table 1). In addition, we strategically generated highly contiguous assemblies of two additional species (*D. hypocausta* and *D. niveifrons*, belonging to the immigrans group; Supplementary table 2), and testes and ovaries RNA-seq of eight species (*D. subobscura*, *D. arawakana*, *D. dunni*, *D. innubila*, *D. funebris*, *D. immigrans*, *D. hypocausta*, and *D. niveifrons*) to either annotate previously unannotated genomes or to refine previous annotations (Supplementary table 1). Altogether, we compiled a total of 47 species spanning the two major arms of the *Drosophila* genus (the *Sophophora* and *Drosophila* subgenera), with three species groups particularly well-represented (*melanogaster*, *obscura*, and *immigrans* groups) (Figure1B) and the outgroup species *Scaptopdrosophila lebanonensis*.

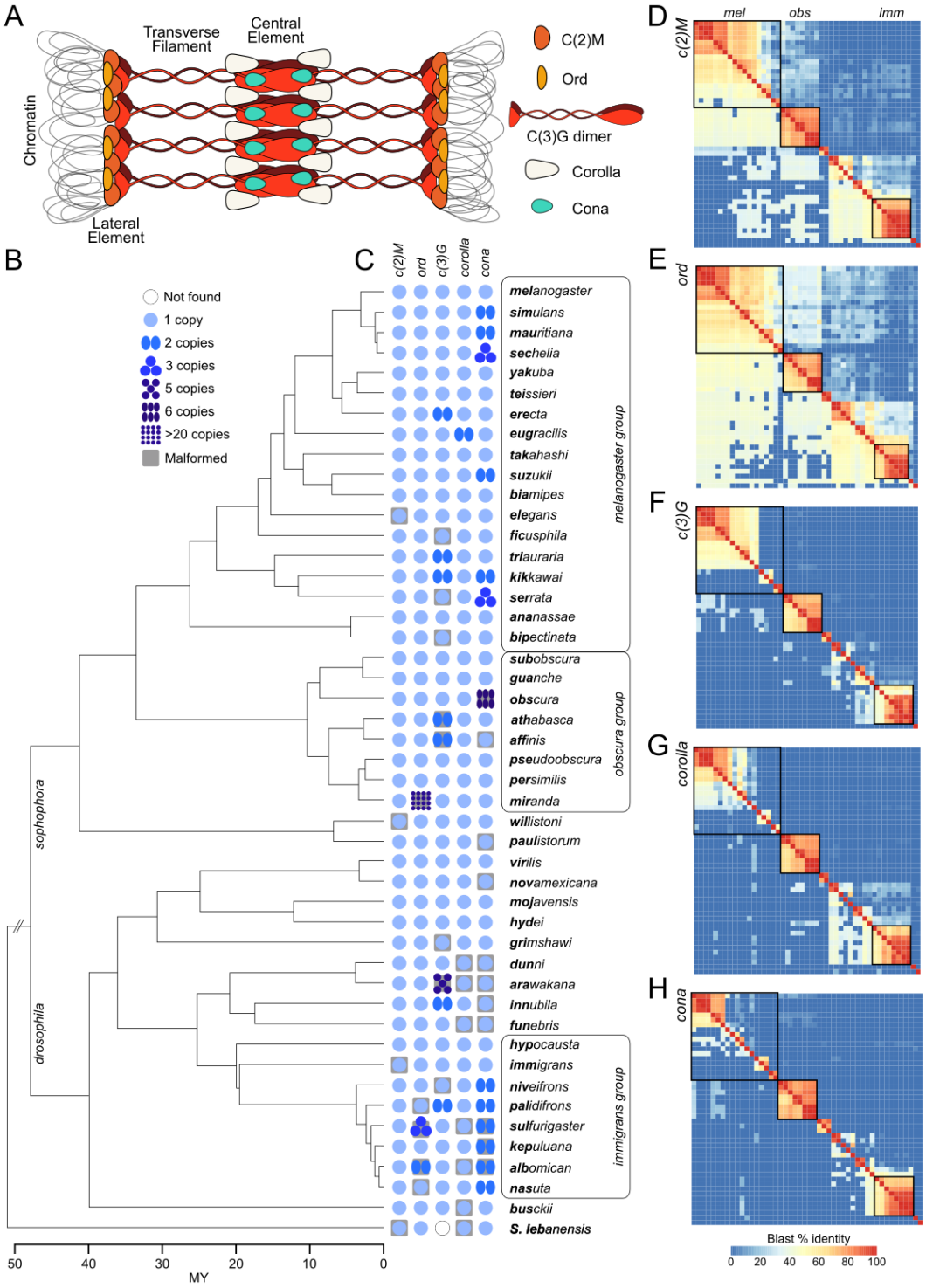

Figure 1. Sequence conservation, or the lack thereof, of synaptonemal complex components across the *Drosophila* genus. A. Cartoon diagram of the *Drosophila* SC and its primary constituents. B. Phylogenetic relationships of the 48 species used in this study. Bold characters in the species name denote species shorthands. Species dense groups are labeled and boxed. C. Presence, copy number, and absence of SC components across the phylogeny. The number of blue circles indicated the copy number. D-H. Pairwise blast sequence alignments between orthologs from representative species across the genus. Alignments above the diagonal are from nucleotide blasts of the transcript or CDS sequences using blastn. Alignments below the diagonal are from protein blasts of the amino acid sequence using blastp. % blast identity is the summation of length of blast alignments across the genes multiplied by the % sequence identity.

Using a multi-step reciprocal best blast hit approach, we sought to identify orthologs and paralogs across the species (Figure 1B and C; Material and methods). However, gene structure of SC components are frequently malformed regardless of the source of the annotation (publicly available, or our own). They are often mis-annotated as truncated or chimeric gene products or entirely missing in the annotation (Supplementary table 3; for examples see Supplementary figure 1), likely due to the combination of exacerbating factors such as poor sequence conservation, frequent presence of tandem duplicates, low RNA-seq reads, and in some cases assembly errors. To ensure proper sequence alignments, we therefore curated all SC genes and manually re-annotated all erroneous ones ensuring at a minimum, well-formed CDSs and intact ORFs (Supplementary figure 1; see Materials and Methods). Note, because *cona* is a short gene with few exons, we hand-annotated its orthologs in 8 additional species (see below).

For the lateral elements *c(2)M* and *ord*, sequence homology is decently preserved (Figure 1D and 1E). However, for the central region genes (*c3G*, *cona*, *corolla*), DNA sequence homology quickly becomes unrecognizable outside of species groups, while weak protein homology is only occasionally recognizable. Previously, Hemmer and Blumentiel 2018 identified SC orthologs in a subset of fly species (12). Increased species and better annotations enabled us to identify orthologs previously missed (*cona* in *D. willistoni* and *corolla* in the outgroup) and resolved discrepant homology relationships (*cona* in the drosophila subgenus, see below). Moreover, we revealed a surprising number of duplications, with *c(2)M* being the only SC gene that remained single-copy. All SC paralogs were previously unaccounted; the only exception being *ord* duplicates in *D. miranda*, which was identified to have rampantly amplified creating over 20 copies (Supplementary figure 2) on the specie's unique neo-sex chromosomes (38). Of the poorly conserved components, *c(3)G* and *cona* in particular have recurrent copy number changes, having more than two copies in 8 and 13 species, respectively. Such propensity to duplicate is strikingly epitomized by the five *c(3)G* and six *cona* copies in *D. arawakana* and *D. obscura,* respectively.

Young tandem SC paralogs and evidence for repeated historic retroduplication and pseudogenization events

Based on protein trees of the SC components, we find that the current copy number distributions reflect at least 3, 7, 1, and 5 independent duplications of *ord*, *c(3)G*, *corolla*, and *cona*, respectively. The majority of paralogs are recent species-specific duplications resulting in short branch lengths (Figure 2A and 3A, and supplementary figures 3). The genomic locations of the copies further reveal that tandem duplications account for the majority of the observed copies. For the transverse filament *c(3)G*, four of the seven duplication events are tandems (Figure 2A and B), three of which (including the five copies in *D. arawakana)* are recent and species-specific. *D. athabasca* and *affinis* share an older tandem duplicate that carries an additional gene (Figure 2C); one of the copies which we designated *c(3)G2* is shorter, while showing poorer conservation and longer branch lengths between the orthologs, suggestive of adoption of a new function. Similarly, for *cona*, three of the five duplication events are tandems, including the 6 copies in *D. obscura* (Figure 3A and B).

Both *c(3)G* and *cona* experienced several instances of retroduplications. *cona* offers two clear examples of old events (*serrata* and *nasuta* subgroups) leading to non-syntenic paralogs shared across many species in the *serrata* and *nasuta* subgroups (Figure 3A). *c(3)G*'s duplication history appears more convoluted but offers unique insight into its dynamic evolution. In *D. kikkawai*, *triauraria*, and *innubila*, *c(3)G* paralogs are found in different regions or chromosomes. The latter two are old duplication events indicated by the long branches separating the paralogs. For *D. triauraria*, the duplication creating *c(3)G2* predated the split in the serrata species subgroup, but is no longer found in the derived lineages, indicating subsequent loss. For *D. innubila,* the phylogeny indicates that the duplication occurred after the split from *D. funebris* (Figure 2A). However, synteny information suggests this is not the true relationship as *c(3)G1* is found to be in the same syntenic block shared across most of the subgenus while the duplicate *c(3)G2* is found in a different synteny block shared with D. *funebris*, *arawakana*, and *dunni* (Supplemental Figure 4). This synteny pattern is therefore more parsimonious with an old duplication in the last common ancestor of the four species with the original copy remaining

only in *D. innubila*. In this scenario, non-allelic gene conversion likely homogenized the duplicates in *D. innubila*, obscuring the true phylogenetic relationship.

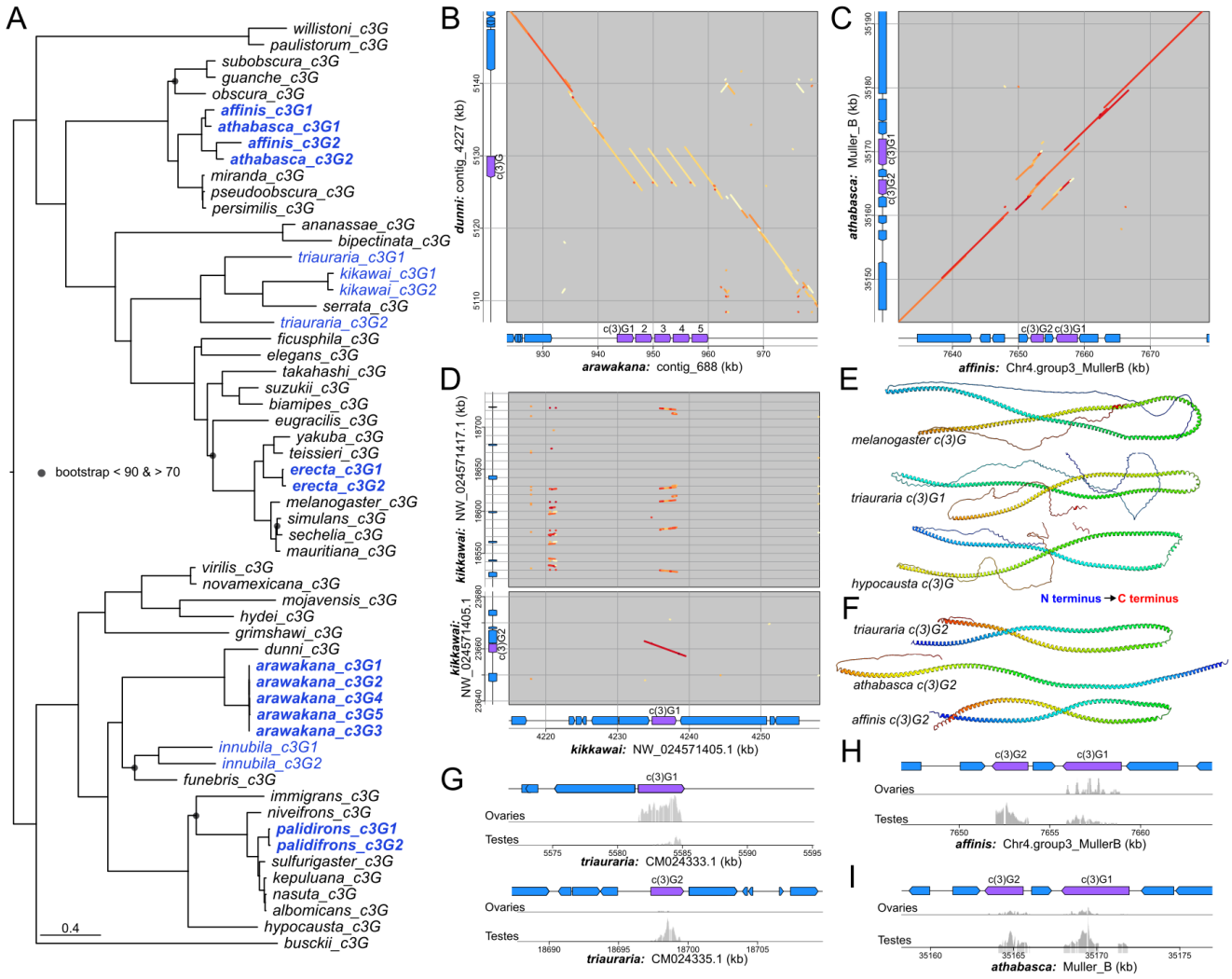

Figure 2. Complex evolution history of the transverse filament *c(3)G*. A. Gene tree of *c(3)G* orthologs and paralogs; nodes with poor bootstrap support (< 90) are demarcated by gray circles. Duplicates are labeled in blue and tandem duplicates are bolded. B, C. Dotplots showing synteny of genomic regions surrounding *c(3)G* between sister species and/or paralogs. The color of the dots represent the % sequence identity front blastn alignments with darker red equating to higher identity. The gene annotations of the compared regions are displayed with *c(3)G* labeled in purple and other neighboring genes in blue. D. Dot plots of regions surrounding *c(3)G1* and its retroduplicate c(3)G2 (bottom), and additional genomic regions with truncated c(3)G alignments in *D. kikkawai*. E, F. Alphafold predictions of the protein structure of *c(3)G* homologs. For coiled-coil domain prediction, see Supplementary Figure 7. Note, because *c(3)G* self-dimerizes, the folding of the monomers displayed in E are unlikely to be the true conformation *in situ*. G-I. Germline expression of c(3)G duplicates in *D. triauraria* (G), affinis (H), and *athabasca* (I).

In addition to the one *c(3)G* retroduplicate in *D. kikkawai*, we curiously identified numerous loci across the genome to be 5' truncated homologs, none of which were annotated or have RNA-seq reads mapping (Figure 2D). These truncated and nonfunctional duplicates, along with the two loss events mentioned, raise the possibility that *c(3)G* experienced not only repeated duplications, but also repeated pseudogenization events. A similar pattern of nonfunctional duplicates is also observed with *corolla* in *D. arawakana*; despite only one full length *corolla*, there are four adjacent tandem copies that lack the 5' exon and therefore likely non-functioning (Supplementary figures 5). Further examining the syntenic relationships of the SC homologs, we find that while the lateral elements have maintained the same local microsynteny consistent with lack of movement, the central region genes have repeatedly relocated to different chromosomes, or different locations on the same chromosome, with the X being the likely ancestral home for all three (Supplementary figure 4A). Such recurrent movements through transpositions are unusual for flies as chromosomal gene content and microsyteny are largely stable while broad chromosome-scale synteny is scrambled by large scale inversions (39). Curiously, one recent relocation occurred in the common ancestor of the *pseudoobscura* species, moving *c(3)G* from Muller B to an euchromatic repeat block on Muller E (Supplementary figure 6), intimating that such movements may be mediated by the instability of repetitive sequences. Since most of these movements no longer have extant paralogs, corresponding pseudogenization events were likely common. Therefore, while most observable paralogs are young tandem duplicates, retroduplications and pseudogenization events have frequently occurred for *c(3)G, cona,* and even *corolla* which has few remnants of duplicates, thus accounting for the existence of many recent and species-specific paralogs but fewer old, shared duplicates.

***c(3)G* and *cona* paralogs with novel germline functions**

Much like other transverse filaments, *C(3)G* has an extensive coiled-coil domain flanked by globular domains at the N- and C- termini that connect to the central and lateral elements, respectively (18, 40). Despite the poor sequence conservation, we find that this canonical structure appears to be conserved across the genus based on protein folding (Figure 2E) (41, 42) and coiled-coil predictions

(Supplementary figure 7) (43). This unique evolutionary property of structural but not sequence conservation is also observed by Kursel et al. in *Caenorhabditis*, whereby central element genes have conserved coiled-coil domains and near invariant protein lengths, but neutrally evolving sequences *(25)*. In *D. athabasca*, *affinis*, and *triauraria*, while *c(3)G1*s produce longer proteins (690, 692, 830 AAs respectively) predicted to have the canonical structure (Figure 2E), the paralogs *c(3)G2s* all produce

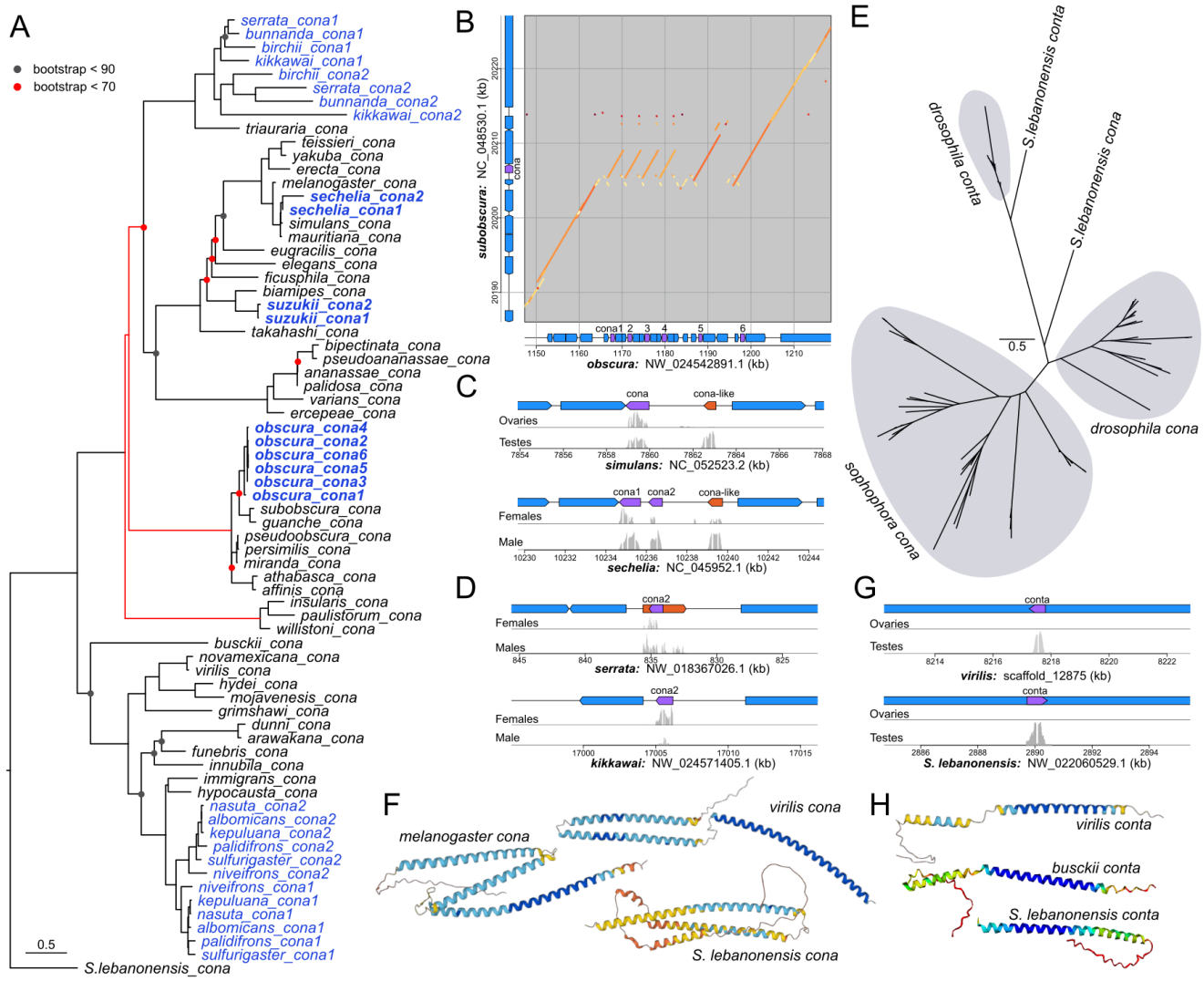

Figure 3. Repeated duplication of the central element *cona*. A. Gene tree of cona homologs; nodes with poor bootstrap support < 90 and < 70 are demarcated by gray and red circles, respectively. Given the poor concordance between alignment methods (Supplementary figure 3), we manually adjusted some branches (red) to better align with known species relationships. Duplicates are labeled in blue and tandem duplicates are bolded. B. Dotplot of syntenic regions surrounding *cona* duplicates in *D. obscura* and *subobscura*. C. Gene structure and expression of *cona* paralogs in *D. simulans* and *sechelia*. Annotated lncRNA of gene with homology to cona is in orange in the gene tracks. D. Same as C but for *D.serrata* and *kikkawai*. F. Alphafold prediction of cona ortholog proteins in *Drosophila* and outgroup species. Darker blue indicates higher confidence in the structure prediction. F. Unrooted protein tree of *cona* and its old duplicate *conta;* major lineages are labeled. G. Annotation and germline expression of *conta* (purple) in *D. virilis* and the outgroup *S. lebanonensis*. H. Alphafold prediction of *conta* in representative species.

notably shorter proteins (361, 395, and 319 AAs, respectively) with the flanking globular domains truncated, if not entirely absent, suggesting they no longer function as transverse filaments that can tether the SC (Figure 2F). Further, these paralogs are highly expressed in the testes but lowly expressed in the ovaries (Figure 2G-I), incongruent with the expectation of female meiotic function. Despite being independent duplications in lineages separated by over 25 million years, *c(3)G2* in *D. triauraria* and *D. athabasca/affinis* display remarkably similar structural and regulatory evolution, revealing molecular convergence for likely male germline function.

Several *cona* duplicates similarly show properties that deviate from its characterized function in SC formation during female meiosis. At least two recent duplication events occurred within the simulans clade generating two upstream paralogs, one ancestral to the three *simulans* species while the other found only in *D. sechelia* (Figure 3C). The *sechelia*-specific duplicate generates a complete but short ORF and likely protein coding, but the shared paralog only shows homology at the 3', lacks a complete ORF, and is annotated as a long non-coding RNA (Figure 3C). To evaluate whether this paralog, which we named cona-like, is transcriptionally active, versus a pseudogenized paralog, we examined RNA-seq data, and found high expression in the testes and males but low-to-no expression in ovaries or females across all *simulans* species (Figure 3C), strongly suggesting testes function as a lncRNA. Adding to the intrigue, this is not the only instance of a *cona* paralog generating lncRNA. In the *serrata* group, the retroduplicate, *cona2*, is shared across the species (Figure 3A and D), but only in *D. serrata* does it generate a lncRNA. Unlike cona-like in the simulans clade, this paralog has a well-formed ORF and is expressed in both sexes, suggesting high protein coding potential. However, the lncRNA is anti-sense as confirmed with strand-specific RNA-seq (Supplementary figure 7) and includes additional flanking sequences that only show expression in males. *cona2* likely generates a functional protein in females and ovaries but was incorporated in the anti-sense direction into lncRNA production in the testes of *D. serrata*. Altogether, these results suggest that both *c(3)G* and *cona* paralogs have repeatedly adopted germline functions unrelated to SC formation.

coronetta (_conta_) is an ancient testes-expressed paralog of _cona_

Previously, Hemmer and Blumenstiel identified *cona* homologs in the *drosophila* subgenus that were highly diverged from *sophophora cona (12)*. One of the proteins they identified through reciprocal best protein blast hit was *GJ20698* in *D. virilis,* a gene producing a short peptide of 109 AA. However, all other orthologs of *GJ20698*, which we were able to find across the *drosophila* subgenus as well as the outgroup, were not reciprocal best hits to *sophophora cona*, often having no identifiable *sophophara* homologs at all. Instead, increased density of species enabled us to identify *D. virilis's GJ16397* (the 2nd best hit to *sophophora cona*) which has orthologs across the *drosophila* subgenus and *S. lebanonensis* that are reciprocals best hits with *sophophora cona* (Figure 1F, 3A). The gene tree affirms that *sophophora cona* is more closely related to *GJ16397,* which we conclude to be the true *cona* ortholog (Figure 3E). Similar to *c(3)G*, *cona* has maintained the same conserved structure of three coils (Figure 3F), despite poor protein homology. *GJ20698*, which we named *coronetta* (*conta*), appears to be a distant paralog, and, given its presence in the outgroup, likely emerged prior to the last common ancestor of *Drosophila* and *Scaptodrososphila*. Unlike *cona*, *conta* sequence is conserved (Figure 3E), found in the same syntenic region (in the intron of the gene *teiresias;* Supplementary figure 8), and expressed in the testes but not ovaries (Figure 3G). Structural prediction of *CONTA* reveals distinctly shorter coiled structures (Figure 3H). Interestingly, we were unable to find *conta*'s ortholog in *sophophora* – not even in the same syntenic location (Supplementary figure 8). This strongly suggests that *conta* has been lost, further underscoring the propensity for SC paralogs to precipitate novel germline function that may be evolutionarily fleeting.

Elevated rates of SC protein evolution are consistent with recurrent positive selection and further accelerated by repeated duplications

Poor sequence homology can result from relaxed constraint due to reduced negative selection or adaptive protein evolution due to positive selection. Previously, Kursel et al. reported that the elevated rate of SC protein evolution in *Caenohabditis* reflect relaxed sequence constraint while the coiled-coil domains and protein lengths are both highly conserved (25). However, Hemmer and Blumenstiel identified both elevated rates of protein evolution and signatures of positive selection for Drosophila SC

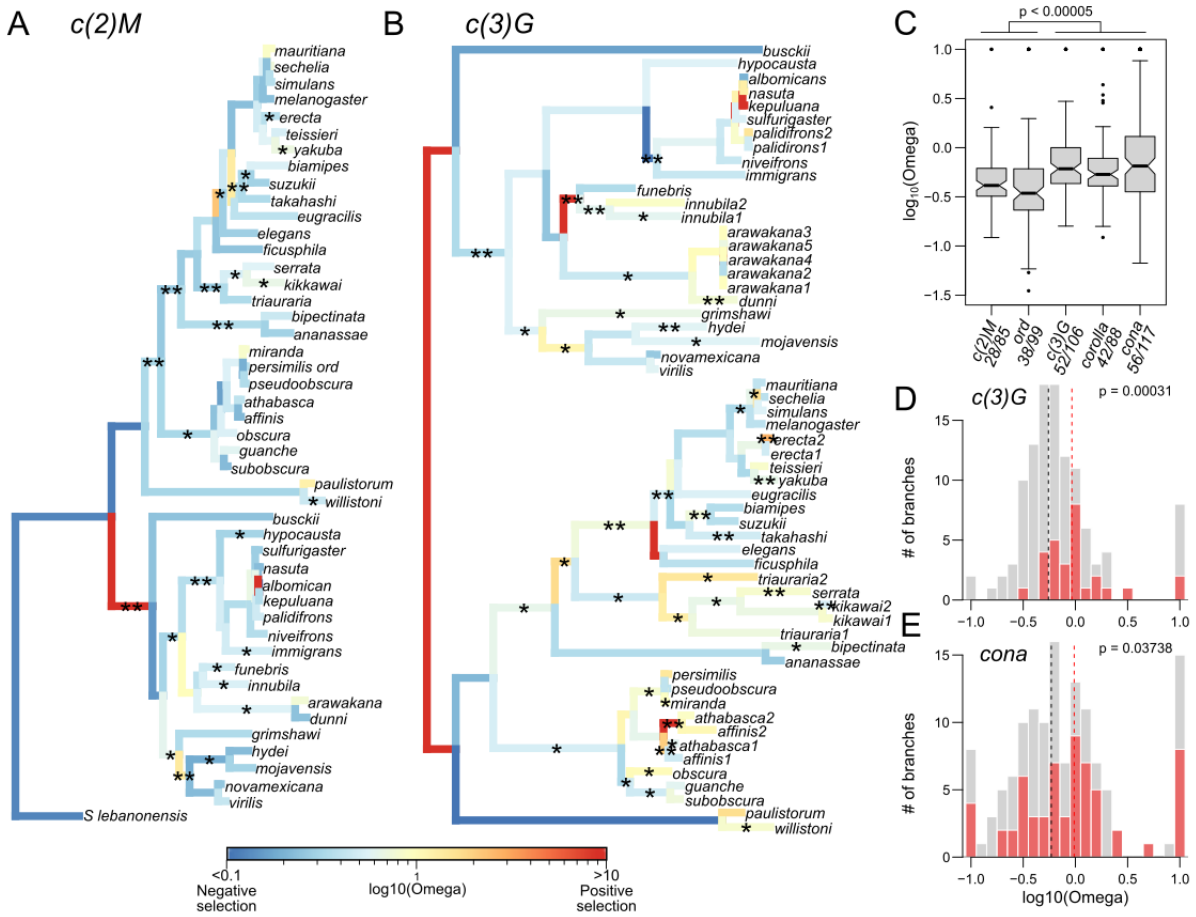

Figure 4. Rate of protein evolution and signatures of positive selection of SC components. A-B. Along the gene trees, branches are colored by their branch-specific rates of protein evolution (omega) with warmer colors representing higher omega. Branches inferred to have significant positive selection in part of the protein are labeled with asterisks (* = p < 0.05 anc ** = p < 0.001). See supplementary figure 10 for the remaining SC genes. C. Distribution of branch-specific omega values for the different SC components trees. P-values are from pairwise Wilcoxon's rank sum tests comparing between the lateral and central region genes. Ratio below each gene indicates the number of significant branches or branches with omega > 1 over total number of branches. D, E. Comparison of the distribution of omega values on branches following duplications (red) versus unduplicated branches (gray). Vertical dotted lines mark the median omega values for duplicated (red) and unduplicated branches (black). P-values were inferred from one-tailed Wilcoxon's rank sum test.

genes (12). We reassessed the rates of protein evolution by estimating the branch-specific ratio of nonsynonymous and synonymous rate of protein evolution, represented by omega. Values approaching 0 indicate negative selection while values close to or greater than 1 indicate relaxed constraint and positive selection, respectively (44). Notedly, gene-wide omega values, which are predominantly negative, are typically the composite of several modes of evolution as different residues and/or domains of the protein can be under different forms and levels of selection (45, 46). Our curated, species dense SC orthologs and paralogs enabled not only branch-specific, gene-wide estimates, but

also detection of significant positive selection occurring only at portions of the protein coding sequence with the Hyphy package (47).

For the more conserved, lateral elements, despite the better preserved sequence homology (Figure 1D and E) and predominantly negative gene-wide omega (Figure 4A and Supplementary figure B), multiple branches still show signatures of positive selection; 28 out of 85 and 38 out of 99 branches display either positive gene-wide omega or significant site-specific selection for *c(2)M* and *ord*, respectively. For the poorly conserved central elements (Figure 4B & supplementary figure 10B-C), not only do they have significantly higher omega than the lateral elements ($p < 0.00005$, pairwise Wilcoxon's rank sum tests; Figure 5C) indicative of higher rate of protein evolution, over 40% of the branches show either positive omega or significant signatures of positive selection. Further, we used PAML to infer the rate of protein evolution within the three well represented species groups, and we, again, consistently find evidence of positive selection. Altogether these results demonstrate that all components of the SC have a history of recurrent adaptive evolution with the central region genes under strikingly frequent and repeated positive selection.

Copy number expansions can allow genes to diversify leading to new functions or subdivision of function among the paralogs both of which can be associated with elevated omega. To test whether the recurrent duplications of SC components lead to elevated rates following functional diversification, we examined branches after duplications and found that for *c(3)G* and *cona,* gene-wide omega values on such branches have median values of 0.925 and 0.914, both significantly higher than single copy branches (Figure 4B and C; $p = 0.000031$ and 0.03788, respectively, one-tailed Wilcoxon's rank sum test). In particular, the *c(3)G2* paralogs with novel testes function in *D. triauraria*, *athabasca*, and *affinis* all show clear signatures of adaptive evolution post duplication (Figure 4B). We note that the terminal branches of the latter two species appear neutrally evolving, but the parent branch displays highly elevated and significant omega, indicating that the paralog evolved under strong positive selection in the common ancestor. Further, other c(3)G duplicates, even the recent ones, show signatures of adaptive evolution including those in *D. erecta*, *kikkawai, and innubila*. Similarly for *cona* duplicates, nearly all the branches of *cona2* in the serrata group show signatures of positive selection, which

accounts for the overall longer branch lengths than those of *cona1* within the same group (Supplementary figure 10C). Thus, repeated duplications further accelerated the already rapid protein evolution of components of the SC.

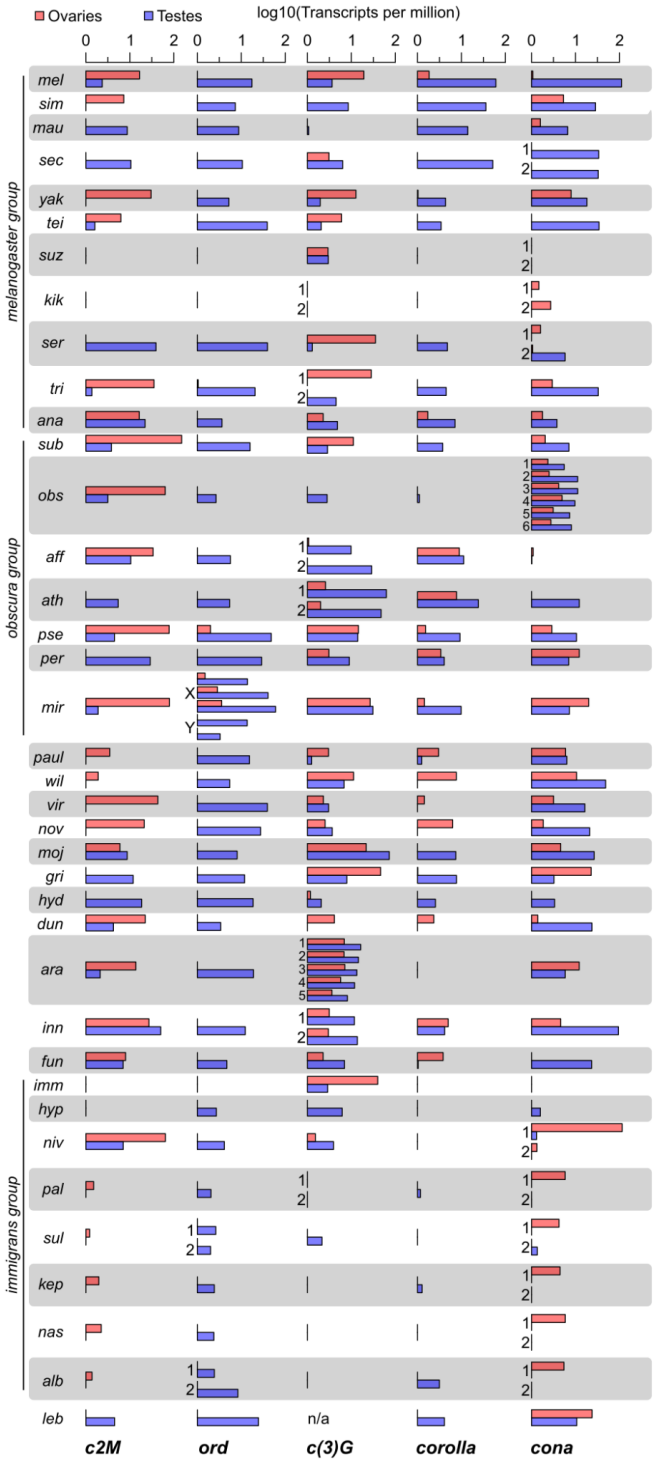

Figure 5. Recurrent germline regulatory turnover of SC components. Ovaries (red) and testes (blue) expression (transcript per million in log scale) of SC genes across 38 species. Duplicates are labeled. Gray and white horizontal bars differentiate between species.

Poor regulatory conservation of SC genes in both female and male germlines

In *Drosophila*, the SC proteins are thought to primarily function in females because males are achiasmic and mutants do not show obvious testes meiotic defects (48–51), other than *ord* which also maintains sister chromatid cohesion in the male germline (52, 53). However, a recent elegant study of the male germline by Rubin et al. found that the progression of pre-meiotic chromosome pairing is slower in *cona* and *c(3)G* mutants, revealing putative male germline function despite the absence of SC assembly (54). Inspired by these results, we examined ovaries and testes expression in a subset of 38 species (Figure 6A) with either available RNA-seq datasets, or ones we generated (Supplementary table 3). Other than *ord* which is testes-biased across all species, we surprisingly find that expression of SC genes – even the single copy ones – is highly unstable and the naive expectation of high ovary and low-to-no testes expression is incorrect in the majority of the species. Other than *c(2)M*, testes-biased expression appears to be the norm rather than the exception as SC components are more highly expressed in the testes in over 70% of the species. We further examined available testes single cell RNA-seq data of *D. miranda* (55), and find expression concentrated in the pre-meiotic cell types such as germline stem cells and spermatogonia (Supplementary Figure 11), reminiscent of their reported pre-meiotic activity in *D. melanogaster* males. Thus, even though they are primarily known for their role in female meiosis, SC genes can also be highly active in testes, arguing for critical male germline function. This is further supported by heavily testes-biased expression of *corolla*, *c(2)M*, and *ord*, in the outgroup *S. lebanonensis*, suggestive of critical testes function in the ancestor.

The striking lability of germline SC expression is particularly evident from several closely related species pairs whereby expression rapidly switches between testes- and ovaries-bias. For instance, *c(2)M* is ovaries-biased in *D. simulans* but testes-biased in the sister species *D. sechelia and mauritania*. Similar rapid expression change of *c(2)M* is also observed in *D. pseudoobscura*/*D. persimilis* and *D. athatbasca*/*D. affinis* sister pairs. Other curious patterns include elevated testes but low-to-no ovary expression such as *c(3)G* in *D. obscura* and *D. hypocausta*. Most puzzling, there are also multiple lineages where components of the SC show little-to-no expression in both gonads, which contributed to the difficulty of annotating them. For example, *D. suzukii* has low gonad

expression of all components except *c(3)G*. The most extensive regulatory conservation is in the *nasuta* subgroup where the expression of SC components are consistently low across the species, especially for *c(3)G* and *corolla*. These drastic differences between species cannot be simply driven by differential tissue contribution in the dissections, as the expression of the SC components are poorly correlated, with many instances of high expressions of one component and low expression of others within the same sample. While it is tempting to infer function outside of the germline based on the absence of expression, even *D. melanogaster* ovaries show low expression of some of the SC components like *ord* and *cona* arguing that normal SC function likely does not require large transcript counts. Nevertheless, these expression patterns reveal that despite the essential roles in regulating crossovers, germline transcriptional regulation of many SC genes is highly labile and constantly evolving, echoing their protein and copy number evolution.

DISCUSSION AND CONCLUSION

Our exhaustive survey for SC orthologs resulted in the surprising identification of many duplicates across *Drosophila*. Duplication is an important mechanism to diversify gene function as the resulting paralogs can either evolve novelty or compartmentalize existing functions, with reduced selective constraint on the gene as a single copy. Indeed, we find elevated rates of protein evolution following duplications indicative of both relaxed constraint and adaptive evolution. Many of the duplicates are young and species-specific showing no evidence of expression differences, and therefore unlikely to have diverged in function. For some of the older *c(3)G* and *cona* paralogs, we observed repeated and independent acquisition of distinct activities surprisingly in the male germline, such as testes-specific expressions and incorporation into lncRNA production. For reasons still unclear, the testes appear to be a unique regulatory environment, producing the largest repertoire of lncRNAs (56, 57) and *de novo* genes (58–61), both of which can have critical roles in spermatogenesis. In light of our unexpected finding of frequent testes expression, we speculate that testes activity of SC genes – even single copy ones – may provide the opportunities to generate paralogs that can diversify. This has occurred multiple times, independently, such as with *c(3)G2* in *D. triauraria* and the common ancestor

of *D. athabasca* and *affinis*, and *cona2* in the *serrata* subgroup. This path to functional novelty or diversification in the testes may be further facilitated by the elevated retrotransposon activity in spermatocytes (62) which can conceivably provide the necessary machineries for both retro- and tandem duplications.

However, despite many paralogs some of which have adopted putatively novel functions, pseudogenization of duplicates also appears common. We have identified multiple instances where extant duplicates in one species have been lost in neighboring lineages (e.g. *c(3)G2* in *D. triauraria*, and *D. innubila*, and *conta*). We also found several examples of remnants of SC duplicates, including truncated copies of *c(3)G* and *corolla*. Such dynamic copy number changes raise a perplexing conundrum: why are *cona* and *c(3)G* prone to produce duplicates under positive selection only for the duplicates to end up as pseudogenes. One clue may come from the exceptional duplication of *ord* in *D. miranda*, where neo-X- and neo-Y-linked gametologs, along with other meiosis-related genes, have massively amplified in tandem (38). The amplification is hypothesized to be the result of dosage dependent sex-ratio meiotic drivers precipitating an arms race for gene copy numbers on sex chromosomes that were recently autosomal (38, 55). Similar dynamics of repeated copy number evolution is also observed for sex ratio drivers and suppressors that manipulate DNA packaging in X- and Y- bearing sperms (63–65). In such models of meiotic conflicts, temporary/young duplications may act to increase the gene dosage to either induce selfish transmission (such as biased sex ratio) or to act as suppressors of drive that restore fitness reduction associated with non-mendelian transmission. Once the conflict is resolved, drivers will pseudogenize and degenerate as it imparts no other fitness benefits.

Since the intricate orchestration of SC assembly across meiotic prophase is necessary for recombination and faithful disjunction, we find the regulatory variability of SC genes in the ovaries to be highly unexpected. Jarringly, multiple SC genes have little-to-no ovary expression in several instances. It is tempting to conclude repeated loss of meiotic function in the ovaries, but we find other possibilities more likely. If SC proteins have long half-lives, minimal transcript production may be sufficient to support robust SC assembly. However, this possibility cannot address why species evolved to have

drastically different expressions. Alternatively, instead of reflecting regulatory evolution, expression lability of SC genes may be responses to environmental fluctuations which is consistent with recombination rate being sensitive to environmental conditions such as temperature (2) and stresses (66). This possibility implies that species likely evolved physiological responses to regulate SC assembly. Such mechanisms can be beneficial in ensuring optimal recombination rates to modulate the amount of genotype diversity in the offspring (66) or proper progression of meiosis in suboptimal cellular conditions like extreme temperatures (9).

Our analyses of the protein coding evolution demonstrate that SC genes have an extensive history of recurrent adaptation with the central region genes being frequent targets of positive selection. While the elevated rates of evolution are partly driven by paralogs diversifying in germline function, orthologs without duplicates also show signatures of positive selection across the gene trees. This contrasts from the SC genes in *Caenorhabditis,* the protein sequence of which are evolving neutrally (25). Further, while *C(3)G* appears structurally conserved, the lengths of the proteins are far more variable with a coefficient of length variation 5 times higher than that of worms (0.17 vs. ~0.03). While this could reflect repeated adaptation in *Drosophila* female meiosis and meiotic recombination, our findings that SC genes frequently function in testes where it is also ancestrally highly expressed compel us to consider additional avenues under recurrent positive selection, especially since spermatogenesis is fruitful grounds for meiotic conflicts, sexual selection, and molecular innovations. Pleiotropy tends to increase molecular constraint (67). However, given the sequence tolerance of the SC, dual function of *Drosophila* SC genes in both oogenesis and spermatogensis may instead predicate a unique scenario where positive selection in the latter has little pleiotropic impact on the former. Dissecting the function of SC genes in the *Drosophila* male germline, which is ironically achiasmate, will therefore be critical to understanding the diversity and evolution of meiotic recombination.

MATERIALS AND METHODS

High molecular weight DNA extraction and genome assembly

To assemble the genomes of *D. hypocausta* and *D. neivefrons*, we followed the Nanopore long read sequencing pipeline from (33, 68). In short, high molecular weight DNA was extracted using the Qiagen Blood & Cell Culture DNA Midi Kit from ~100 males of *D. hypocausta* strain 15115-1871.04 from the National Drosophila Species Stock Center and ~50 females of *D. niveifrons* strain LAE-276 from the Kyorin Drosophila Species Stock Center. DNA strands were hand spooled after precipitation, followed by gentle washing with supplied buffers.

RNA-seq preparation and analyses

5 pairs of ovaries and testes were dissected from adult females and males and stored in Trizol at -80 degrees, followed by standard RNA extraction. RNA-seq libraries were generated using either the NEBNext RNA Library Prep Kit for Illumina with the Stranded and mRNA isolation Modules or the Illumina Truseq Stranded mRNA Library Prep kit. After quality check with the Fragment Analyzer QB3-Berkeley, the libraries were sequenced by Novogene. We aligned the reads (both ones we generated and downloaded from SRA) using hisat2 (v2.2.1) (69) on either pair-end or single-end mode to their respective genomes with the –dta flag to allow for downstream transcriptome assembly. After sorting the aligned reads with samtools (v1.5) (70), we used the featureCount (v2.0.3) in the Subread package (71) for read-counting over genes, allowing for non-uniquely mapped reads (-M flag). Read count tables were processed and analyzed in R (v4.2.2) and Rstudio (v2022.12.0). For gene expression analyses, we normalized the read counts across samples by converting them to transcript per million (TPM) (72). For species where we needed to do de novo gene annotation, we used stringtie v2.1.6 (73) on default for genome-guided transcript assembly.

Gene annotation and manual curation of gene structures

For species that required gene annotation, we ran three rounds of maker (74). For evidence-based ab initio gene prediction in the first round, we supplied the transcript assembly from stringtie, de novo repeat index from RepeatModeler2 (75), transcript sequence from closely related-species and protein sequence data from D. melanogaster and D. virilis downloaded from FlyBase. The maker results from round one were used to train the species specific gene model using SNAP (76). The resulting

snap.hmm file was fed back into maker for round 2. We iterated this process again, refining the gene models for a 3rd round of maker.

For malformed or missing annotations, we first visualized the gene structures and RNA-seq reads mapping around them using IGV (v2.16.0) (77). Additionally, we manually defined the region of the genome showing gene homology by blastn-ing the well-formed ortholog from a closely related species to the genome. The combination of RNA-seq reads mapping the blast-hit boundaries provided evidence to correct erroneous exon-intron injunctions, truncated annotations, chimeric gene structures, and absent annotations. To update the annotations file (.gff file), we used the genome browser GenomeView (v2250) (78) to manually edit or add the gene structures including mRNAs, exons, CDSs. All edited genes have at least full open reading frames, although 5' and 3' UTRs are missing. All manually annotated features were marked by the flag "hand" in the gffs. The updated .gff is then exported and sorted using GFF3sort (79), and transcript sequences are retrieved using gffread (0.9.12) (80). In several instances, we noticed assembly errors leading to malformed genes. One was *ord* in D. nasuta which had a stretch of N's within the gene body indicating scaffolding points. The other was in D. *neivifrons* where c(3)G was annotated as two fragments. This was due to a deletion of a single nucleotide in the genome causing a shifted reading frame which led to malformed annotations. The deletion was revealed by RNA-seq read mapping, whereby all reads showed a one basepair insertion. We fused the fragmented annotations into one, and corrected the transcript sequence to rectify the erroneous deletion. Lastly, we initially could not identify *corolla* in the primary NCBI genome assembly of *D. funebris* strongly suggesting gene is loss; however, we were subsequently able to identify it in an unplaced repeat-rich contig in a separate assembly.

Homolog search with reciprocal best blast hits of transcripts and/or coding sequences

We first used blastn to identify homologous transcripts between species pairs, and in cases where no clear best hit was identified, we then used tblastn to identify homologous transcript with protein sequences. The absence of protein hits with tblastn is then followed by blasting to the genome, to ensure the absence of an ortholog in the transcript sequences is not merely the result of absent annotation.

To identify orthologs and paralogs using a reciprocal best blast hit strategy, we reciprocally blastn-ed transcript sequences from species pairs using the commands:

```
blastn -task blastn -query species1.transcripts -db species2.transcripts -outfmt "6 qseqid sseqid pident length qlen slen  mismatch gapopen qstart qend sstart send evalue bitscore" -evalue 1
blastn -task blastn -query species2.transcripts  -db species1.transcripts -outfmt "6 qseqid sseqid pident length qlen slen  mismatch gapopen qstart qend sstart send evalue bitscore" -evalue 1.
```

For publicly available genomes with annotation files, we generated the transcript sequences using gffread, otherwise we used transcript sequences generated by maker. We then used grep to identify the blast hits and checked whether they are reciprocal best hits of each other. For the sophophora and drosophila sub-genera, we used *D. melanogaster* and *D. virilis* sequences downloaded from flybase as the focal species and blasted them first to their close relatives. When one species yields no blast hit for a gene, we then use other closely related species where the orthologs was successfully identified. If no hits can be identified for a species or a clade, we then repeat the same procedure using tblastn to identify translated protein sequences, as amino acid can be more conserved than nucleotide sequence. If tblastn fails to identify a homologous transcript, we then blastn-ed to the genome sequence. True absences/loss of a gene will yield poor or no blast hits, while missing annotation will result in clear noncontiguous hits with gaps corresponding to introns.

Microsynteny surrounding homolgs and chromosome placement

We extracted the sequences of the homologs including 50kb up and downstream using bedtools slop and bedtools getfasta. We then pairwise blastn-ed the sequences to each other and filtered out alignments with E-values of < 0.01 or shorter than 100 bp. To infer the extent of homology in the flanking sequences, we calculated the proportion of sequence aligned, excluding the positions of the homolog. Genes are deemed to be in non-syntenic regions if they share < 5% flanking homology. For species without Muller element designation of chromosomes, we assigned Muller elements by blastn-ing to the genome of a closely related species where the Muller elements have been determined..

Phylogeny construction

We retrieved the CDS for all genes, removed the stop codon, and converted them first to protein sequences using EMBOSS Transeq (81). For we then aligned the protein sequences using three aligners with the commands: prank (v.170427) -protein -showtree (82), mafft (v7.505) --localpair --maxiterate 1000 (83), and muscle (v5.1) (84). The resulting multi-sequence alignment fasta file were used as the input for iqtree (v1.6.12) (85) with the flags -AA and -bb 2000 for 2000 iterations of ultrafast boot-strapping (86). These trees were manually rooted with *S. lebanensis* as the outgroup species in FigTree (v1.4.4) (87), and then Node labels were added to the trees to facilitate downstream rate of evolution analyses with Hyphy with phytools (v1.5.1) (88) in R. We then selected the resulting trees with the best bootstrap support and concordance with species tree. *cona* trees were highly inconsistent across alignment methods with many poorly supported nodes.

Branch-specific rate of protein evolution and positive selection

We used TranslatorX (89) to align the CDS sequence based on the protein alignments. Providing the CDS alignments and the protein trees, we used the ABSREL module in HyPhy (v2.5.51) (47, 90) to infer the branch-specific rate of protein evolution (Omega) and significant signatures of positive selection. We wrote a custom script (github.com/weikevinhc/phyloparse) to parse the HyPhy .json output in R where the trees were reoriented with phytools and visualized with colors representing gene-wide omega values. The nominal p-values were used for significance. In addition, we used PAML (91) for the species-group specific test of recurrent positive selection. For comparisons with the rate of regulatory evolution, we reran HyPhy after removing sequences from species with no RNA-seq data.

Protein structure prediction with AlphaFold

Structures of proteins previously annotated in NCBI were retrieved from the AlphaFold Protein Structure Database (41). For genes we annotated, we used ColabFold (v1.5.2), an implementation of AlphaFold on the Google Colab platform (42) and selected num_recycles 24, producing structure predictions that were visualized in UCSF ChimeraX (92).

References

1. J. R. True, J. M. Mercer, C. C. Laurie, Differences in crossover frequency and distribution among three sibling species of Drosophila. *Genetics* **142**, 507–523 (1996).

2. C. Stern, An Effect of Temperature and Age on Crossing-Over in the First Chromosome of Drosophila Melanogaster. *Proc. Natl. Acad. Sci. U. S. A.* **12**, 530–532 (1926).

3. T. V. Kent, J. Uzunović, S. I. Wright, Coevolution between transposable elements and recombination. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372** (2017).

4. M. W. Feldman, S. P. Otto, F. B. Christiansen, Population genetic perspectives on the evolution of recombination. *Annu. Rev. Genet.* **30**, 261–295 (1996).

5. K. Bomblies, J. D. Higgins, L. Yant, Meiosis evolves: adaptation to external and internal environments. *New Phytol.* **208**, 306–323 (2015).

6. Y. Brandvain, G. Coop, Scrambling eggs: meiotic drive and the evolution of female recombination rates. *Genetics* **190**, 709–723 (2012).

7. T. Lenormand, The evolution of sex dimorphism in recombination. *Genetics* **163**, 811–822 (2003).

8. J. M. Sardell, M. Kirkpatrick, Sex Differences in the Recombination Landscape. *Am. Nat.* **195**, 361–379 (2020).

9. C. H. Morgan, H. Zhang, K. Bomblies, Are the effects of elevated temperature on meiotic recombination and thermotolerance linked via the axis and synaptonemal complex? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372** (2017).

10. J. A. Anderson, W. D. Gilliland, C. H. Langley, Molecular population genetics and evolution of Drosophila meiosis genes. *Genetics* **181**, 177–185 (2009).

11. C. L. Brand, M. V. Cattani, S. B. Kingan, E. L. Landeen, D. C. Presgraves, Molecular Evolution at a Meiosis Gene Mediates Species Differences in the Rate and Patterning of Recombination. *Curr. Biol.* **28**, 1289–1295.e4 (2018).

12. L. W. Hemmer, J. P. Blumenstiel, Holding it together: rapid evolution and positive selection in the synaptonemal complex of Drosophila. *BMC Evol. Biol.* **16**, 91 (2016).

13. K. Samuk, B. Manzano-Winkler, K. R. Ritz, M. A. F. Noor, Natural Selection Shapes Variation in Genome-wide Recombination Rate in Drosophila pseudoobscura. *Curr. Biol.* **30**, 1517–1528.e6 (2020).

14. J. G. Ault, H. P. Lin, K. Church, Meiosis in Drosophila melanogaster. IV. The conjunctive mechanism of the XY bivalent. *Chromosoma* **86**, 309–317 (1982).

15. K. R. Ritz, M. A. F. Noor, N. D. Singh, Variation in Recombination Rate: Adaptive or Not? *Trends Genet.* **33**, 364–374 (2017).

16. M. Zelkowski, M. A. Olson, M. Wang, W. Pawlowski, Diversity and Determinants of Meiotic Recombination Landscapes. *Trends Genet.* **35**, 359–370 (2019).

17. I. R. Henderson, K. Bomblies, Evolution and Plasticity of Genome-Wide Meiotic Recombination Rates. *Annu. Rev. Genet.* **55**, 23–43 (2021).

18. S. L. Page, R. S. Hawley, The genetics and molecular biology of the synaptonemal complex. *Annu. Rev. Cell Dev. Biol.* **20**, 525–558 (2004).

19. K. Schmekel, U. Skoglund, B. Daneholt, The three-dimensional structure of the central region in a synaptonemal complex: a comparison between rat and two insect species, Drosophila melanogaster and Blaps cribrosa. *Chromosoma* **102**, 682–692 (1993).

20. A. T. Carpenter, Electron microscopy of meiosis in Drosophila melanogaster females. I. Structure, arrangement, and temporal change of the synaptonemal complex in wild-type. *Chromosoma* **51**, 157–182 (1975).

21. S. E. Hughes, D. E. Miller, A. L. Miller, R. S. Hawley, Female Meiosis: Synapsis, Recombination, and Segregation in Drosophila melanogaster. *Genetics* **208**, 875–908 (2018).

22. J. Loidl, S. pombe linear elements: the modest cousins of synaptonemal complexes. *Chromosoma* **115**, 260–271 (2006).

23. K. Schild-Prüfert, *et al.*, Organization of the synaptonemal complex during meiosis in Caenorhabditis elegans. *Genetics* **189**, 411–421 (2011).

24. R. S. Hawley, Solving a meiotic LEGO puzzle: transverse filaments and the assembly of the synaptonemal complex in Caenorhabditis elegans. *Genetics* **189**, 405–409 (2011).

25. L. E. Kursel, H. D. Cope, O. Rog, Unconventional conservation reveals structure-function relationships in the synaptonemal complex. *Elife* **10** (2021).

26. S. L. Page, *et al.*, Corona is required for higher-order assembly of transverse filaments into full-length synaptonemal complex in Drosophila oocytes. *PLoS Genet.* **4**, e1000194 (2008).

27. K. A. Collins, *et al.*, Corolla is a novel protein that contributes to the architecture of the synaptonemal complex of Drosophila. *Genetics* **198**, 219–228 (2014).

28. S. L. Page, R. S. Hawley, c(3)G encodes a Drosophila synaptonemal complex protein. *Genes Dev.* **15**, 3130–3143 (2001).

29. H. A. Webber, L. Howard, S. E. Bickel, The cohesion protein ORD is required for homologue bias during meiotic recombination. *J. Cell Biol.* **164**, 819–829 (2004).

30. E. A. Manheim, K. S. McKim, The Synaptonemal complex component C(2)M regulates meiotic crossing over in Drosophila. *Curr. Biol.* **13**, 276–285 (2003).

31. S. Mahajan, K. H.-C. Wei, M. J. Nalley, L. Gibilisco, D. Bachtrog, De novo assembly of a young Drosophila Y chromosome using single-molecule sequencing and chromatin conformation capture. *PLoS Biol.* **16**, e2006348 (2018).

32. D. E. Miller, C. Staber, J. Zeitlinger, R. S. Hawley, Highly Contiguous Genome Assemblies of 15 Drosophila Species Generated Using Nanopore Sequencing. *G3* **8**, 3131–3141 (2018).

33. R. Bracewell, K. Chatla, M. J. Nalley, D. Bachtrog, Dynamic turnover of centromeres drives karyotype evolution in Drosophila. *Elife* **8** (2019).

34. T. Hill, H.-L. Rosales-Stephens, R. L. Unckless, Rapid divergence of the male reproductive proteins in the Drosophila dunni group and implications for postmating incompatibilities between species. *G3* **11** (2021).

35. B. Y. Kim, *et al.*, Highly contiguous assemblies of 101 drosophilid genomes. *Elife* **10** (2021).

36. M. Chakraborty, *et al.*, Evolution of genome structure in the Drosophila simulans species complex. *Genome Res.* **31**, 380–396 (2021).

37. K. H.-C. Wei, D. Mai, K. Chatla, D. Bachtrog, Dynamics and Impacts of Transposable Element Proliferation in the Drosophila nasuta Species Group Radiation. *Mol. Biol. Evol.* **39** (2022).

38. D. Bachtrog, S. Mahajan, R. Bracewell, Massive gene amplification on a recently formed Drosophila Y chromosome. *Nat Ecol Evol* **3**, 1587–1597 (2019).

39. S. W. Schaeffer, *et al.*, Polytene chromosomal maps of 11 Drosophila species: the order of genomic scaffolds inferred from genetic and physical maps. *Genetics* **179**, 1601–1655 (2008).

40. L. K. Anderson, *et al.*, Juxtaposition of C(2)M and the transverse filament protein C(3)G within the central region of Drosophila synaptonemal complex. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 4482–4487 (2005).

41. M. Varadi, *et al.*, AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).

42. M. Mirdita, *et al.*, ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).

43. C. Combet, C. Blanchet, C. Geourjon, G. Deléage, NPS@: network protein sequence analysis. *Trends Biochem. Sci.* **25**, 147–150 (2000).

44. M. Kimura, Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**, 275–276 (1977).

45. Z. Yang, J. P. Bielawski, Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15**, 496–503 (2000).

46. Z. Yang, R. Nielsen, Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43 (2000).

47. S. L. Kosakovsky Pond, *et al.*, HyPhy 2.5-A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Mol. Biol. Evol.* **37**, 295–299 (2020).

48. L. Sandler, D. L. Lindsley, B. Nicoletti, G. Trippa, Mutants affecting meiosis in natural populations of Drosophila melanogaster. *Genetics* **60**, 525–558 (1968).

49. B. S. Baker, A. T. Carpenter, Genetic analysis of sex chromosomal meiotic mutants in Drosophilia melanogaster. *Genetics* **71**, 255–286 (1972).

50. B. S. Baker, J. C. Hall, “Meiotic mutants: genic control of meiotic recombination and chromosome segregation” in *The Genetics and Biology of Drosophila*, N. E. Ashburner M, Ed. (Academic Press, 1976), pp. 351–434.

51. D. L. Lindsley, L. Sandler, S. J. Counce, A. C. Chandley, K. R. Lewis, The Genetic Analysis of Meiosis in Female Drosophila melanogaster [and Discussion]. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **277**, 295–312 (1977).

52. J. M. Mason, Orientation disruptor (ord): a recombination-defective and disjunction-defective meiotic mutant in Drosophila melanogaster. *Genetics* **84**, 545–572 (1976).

53. L. S. Goldstein, Mechanisms of chromosome orientation revealed by two meiotic mutants in

Drosophila melanogaster. *Chromosoma* **78**, 79–111 (1980).

54. T. Rubin, *et al.*, Premeiotic pairing of homologous chromosomes during *Drosophila* male meiosis. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2207660119 (2022).

55. K. H.-C. Wei, K. Chatla, D. Bachtrog, Single cell RNA-seq in Drosophila testis reveals evolutionary trajectory of sex chromosome regulation. *bioRxiv*, 2022.12.07.519494 (2022).

56. J. B. Brown, *et al.*, Diversity and dynamics of the Drosophila transcriptome. *Nature* **512**, 393–399 (2014).

57. K. Wen, *et al.*, Critical roles of long noncoding RNAs in Drosophila spermatogenesis. *Genome Res.* **26**, 1233–1244 (2016).

58. D. J. Begun, H. A. Lindfors, A. D. Kern, C. D. Jones, Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. *Genetics* **176**, 1131–1137 (2007).

59. M. T. Levine, C. D. Jones, A. D. Kern, H. A. Lindfors, D. J. Begun, Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 9935–9939 (2006).

60. E. Witt, S. Benjamin, N. Svetec, L. Zhao, Testis single-cell RNA-seq reveals the dynamics of de novo gene transcription and germline mutational bias in Drosophila. *Elife* **8** (2019).

61. S. B. Van Oss, A.-R. Carvunis, De novo gene birth. *PLoS Genet.* **15**, e1008160 (2019).

62. M. A. Lawlor, W. Cao, C. E. Ellison, A transposon expression burst accompanies the activation of Y-chromosome fertility genes during Drosophila spermatogenesis. *Nat. Commun.* **12**, 6854 (2021).

63. C. A. Muirhead, D. C. Presgraves, Satellite DNA-mediated diversification of a sex-ratio meiotic drive gene family in Drosophila. *Nat Ecol Evol* **5**, 1604–1612 (2021).

64. J. Vedanayagam, C.-J. Lin, E. C. Lai, Rapid evolutionary dynamics of an expanding family of meiotic drive factors and their hpRNA suppressors. *Nat Ecol Evol* **5**, 1613–1623 (2021).

65. C.-H. Chang, I. Mejia Natividad, H. S. Malik, Expansion and loss of sperm nuclear basic protein genes in Drosophila correspond with genetic conflicts between sex chromosomes. *Elife* **12** (2023).

66. N. D. Singh, *et al.*, Fruit flies diversify their offspring in response to parasite infection. *Science* **349**, 747–750 (2015).

67. M. Pavličev, J. M. Cheverud, Constraints Evolve: Context Dependency of Gene Effects Allows Evolution of Pleiotropy. *Annu. Rev. Ecol. Evol. Syst.* **46**, 413–434 (2015).

68. R. Bracewell, A. Tran, K. Chatla, D. Bachtrog, Chromosome-Level Assembly of Drosophila bifasciata Reveals Important Karyotypic Transition of the X Chromosome. *G3* **10**, 891–897 (2020).

69. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

70. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

71. Y. Liao, G. K. Smyth, W. Shi, The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* **41**, e108 (2013).

72. L. Pachter, Models for transcript quantification from RNA-Seq. *arXiv [q-bio.GN]* (2011).

73. M. Pertea, *et al.*, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).

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

75. J. M. Flynn, *et al.*, RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451–9457 (2020).

76. I. Korf, Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).

77. J. T. Robinson, *et al.*, Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

78. T. Abeel, T. Van Parys, Y. Saeys, J. Galagan, Y. Van de Peer, GenomeView: a next-generation genome browser. *Nucleic Acids Res.* **40**, e12 (2012).

79. T. Zhu, C. Liang, Z. Meng, S. Guo, R. Zhang, GFF3sort: a novel tool to sort GFF3 files for tabix indexing. *BMC Bioinformatics* **18**, 482 (2017).

80. G. Pertea, M. Pertea, GFF Utilities: GffRead and GffCompare. *F1000Res.* **9**, 304 (2020).

81. F. Madeira, *et al.*, Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res.* **50**, W276–W279 (2022).

82. A. Löytynoja, Phylogeny-aware alignment with PRANK. *Methods Mol. Biol.* **1079**, 155–170 (2014).

83. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

84. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

85. B. Q. Minh, *et al.*, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

86. D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

87. ,FigTree (July 13, 2023).

88. L. J. Revell, phytools: an R package for phylogenetic comparative biology (and other things): phytools: R package. *Methods Ecol. Evol.* **3**, 217–223 (2012).

89. F. Abascal, R. Zardoya, M. J. Telford, TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* **38**, W7–13 (2010).

90. M. D. Smith, *et al.*, Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* **32**, 1342–1353 (2015).

91. Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

92. E. F. Pettersen, *et al.*, UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).

93. J. Felsenstein, Phylogenies and the Comparative Method. *Am. Nat.* **125**, 1–15 (1985).