Title

Rapid Functional and Sequence Differentiation of a Tandemly Repeated Species-Specific Multigene Family in Drosophila

Journal

Molecular Biology and Evolution, 34(1)

ISSN

0737-4038

Authors

Clifton, Bryan D
Librado, Pablo
Yeh, Shu-Dan
et al.

Publication Date

2017

DOI

10.1093/molbev/msw212

Rapid Functional and Sequence Differentiation of a Tandemly Repeated Species-Specific Multigene Family in *Drosophila*

Bryan D. Clifton,[1] Pablo Librado,[2] Shu-Dan Yeh,[3] Edwin S. Solares,[1] Daphne A. Real,[1] Suvini U. Jayasekera,[1] Wanting Zhang,[4] Mijuan Shi,[4] Ronni V. Park,[1] Robert D. Magie,[1] Hsiu-Ching Ma,[1] Xiao-Qin Xia,[4] Antonio Marco,[5] Julio Rozas,[6] and José M. Ranz*[,1]

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

[2]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark

[3]Department of Life Sciences, National Central University, Taoyuan City, Zhongli District, Taiwan

[4]Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, Hubei Province, China

[5]School of Biological Sciences, University of Essex, Colchester, United Kingdom

[6]Departament de Genètica, Microbiologia i Estadistica, and Institut de Recerca de la Biodiversitat, Universitat de Barcelona, Barcelona, Spain

***Corresponding author:** E-mail: jranz@uci.edu

Associate editor: Ilya Ruvinsky

Abstract

Gene clusters of recently duplicated genes are hotbeds for evolutionary change. However, our understanding of how mutational mechanisms and evolutionary forces shape the structural and functional evolution of these clusters is hindered by the high sequence identity among the copies, which typically results in their inaccurate representation in genome assemblies. The presumed testis-specific, chimeric gene *Sdic* originated, and tandemly expanded in *Drosophila melanogaster*, contributing to increased male-male competition. Using various types of massively parallel sequencing data, we studied the organization, sequence evolution, and functional attributes of the different *Sdic* copies. By leveraging long-read sequencing data, we uncovered both copy number and order differences from the currently accepted annotation for the *Sdic* region. Despite evidence for pervasive gene conversion affecting the *Sdic* copies, we also detected signatures of two episodes of diversifying selection, which have contributed to the evolution of a variety of C-termini and miRNA binding site compositions. Expression analyses involving RNA-seq datasets from 59 different biological conditions revealed distinctive expression breadths among the copies, with three copies being transcribed in females, opening the possibility to a sexually antagonistic effect. Phenotypic assays using *Sdic* knock-out strains indicated that should this antagonistic effect exist, it does not compromise female fertility. Our results strongly suggest that the genome consolidation of the *Sdic* gene cluster is more the result of a quick exploration of different paths of molecular tinkering by different copies than a mere dosage increase, which could be a recurrent evolutionary outcome in the presence of persistent sexual selection.

Key words: **gene amplification, functional diversification, newly evolved gene, *Sdic*, *Drosophila*.**

Article

Introduction

Genes restricted to one or a few closely related species are ubiquitous across phyla (Tautz and Domazet-Loso 2011; Long et al. 2013). Despite their young age, these genes can exert noteworthy effects on organismal viability and fertility (Chen et al. 2010; Mayer et al. 2015), therefore their study is instrumental for determining how early mutational mechanisms and evolutionary forces refine the functional attributes of a gene and its organismal impact shortly after its formation (Hahn 2009; Chen et al. 2013). This is especially important in the case of recent expansions of tandemly duplicated genes, which are thought to play a primary role during species adaptation and differentiation (Brown et al. 1998; Newcomb et al. 2005; Perry et al. 2007; Jugulam et al. 2014).

Genome consolidation of recent duplicates can be achieved throughout different evolutionary paths in which natural selection and genetic drift contribute with different intensities (Innan and Kondrashov 2010; Katju and Bergthorsson 2013). In particular, the expansion dynamics of gene clusters is commonly thought to be associated with a beneficial effect via increased gene dosage (Ohno 1970; Kondrashov 2012). However, this process can be subsequently accompanied by some degree of functional diversification among the duplicates through a secondary functional attribute of the gene product (Bergthorsson et al. 2007). A relevant constraint on functional paralog divergence to consider is the homogenizing effect exerted by interlocus gene conversion, i.e., the non-reciprocal recombination process that results in the transfer of DNA stretches between similar

non-allelic sequences, which is particularly relevant in the case of young tandemly arranged duplicates (Chen et al. 2007; Osada and Innan 2008; Casola et al. 2010). Importantly, this homogenizing effect also impacts the retention probability of the duplicates and therefore their ability to contribute to species adaptation (Walsh 1987; Innan 2003; Katju 2012).

Critically, the analysis of the functional and evolutionary dynamics of recent tandem expansions of species-specific genes is hindered precisely by the repetitive nature and high sequence identity of the constituent copies. These features limit the resolution of microarray and quantitative PCR technologies as well as the information derived from short-read based sequencing technologies, which typically results in an inaccurate representation of these gene clusters in current genome assemblies in the form of sequence errors or copies being collapsed (Hemingway et al. 2004; Bariami et al. 2012; Krsticevic et al. 2015).

The *Sperm-specific dynein intermediate chain* (*Sdic*) multigene family originated in the *D. melanogaster* lineage less than 4.9 mya (Obbard et al. 2012). The *Sdic* ancestral copy started its formation with a local segmental duplication of two adjacent genes on the *X* chromosome, *AnxB10* and *sw*. This was followed by point mutations and indels of varying size that obliterated sections along the parental genes, resulting in a fusion event between their inner copies, with *AnxB10* not contributing to the transcribed region of *Sdic*, and a *de novo* exon acquisition from a previously non-coding sequence of *sw* (fig. 1*A*) (Nurminsky et al. 1998b). Subsequently, *Sdic*

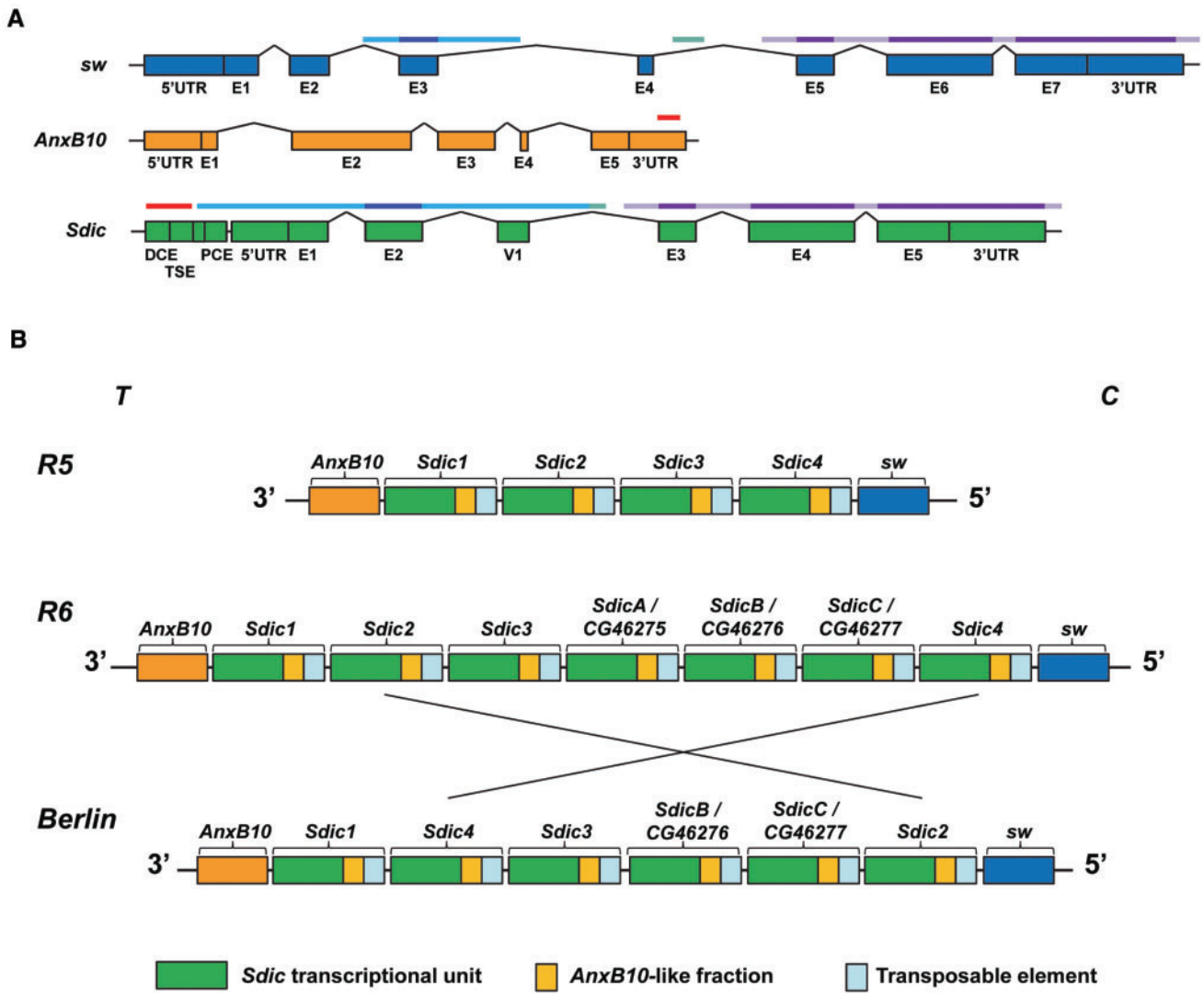

Fig. 1. Organizational features of the *Sdic* region of *D. melanogaster*. (*A*) Sequence stretches of the parental genes *sw* and *AnxB10* that contribute to the structure of the chimeric protein-coding gene *Sdic*. Top colored bars denote sequence stretches from parental genes that correspond to sequence stretches in *Sdic*. Dark and light tones, exonic and intronic sequence in *sw* respectively. (*B*) Different organization of the *Sdic* region in three assemblies of the *D. melanogaster* genome in the ISO_1 strain. The *Sdic* cluster is composed of tandem repeats, each consisting of three parts: *Sdic*, originated primarily from stretches of *sw*; another putative transcriptional unit originated from *AnxB10* named *AnxB10-like*; and a ~785 nt stretch from the transposable element Rt1c (Nurminsky et al. 1998b; Ponce and Hartl 2006). The relative location (black lines) and number of repeats vary between assemblies, which determine the size of the region: ~31 kb in Release 5 (*R5*); ~46 kb in the assembly GCA_000778455.1 (*Berlin*); and ~54 kb in Release 6 (*R6*). *T*, telomere; *C*, centromere. Distances and lengths of different features are not to a scale.

became repeatedly tandemly duplicated, representing one of the most noticeable gene family expansions in *D. melanogaster* (Hahn et al. 2007). One *Sdic* copy has been shown to be expressed only in males, with its encoded product present in the tail of mature spermatocytes, collectively pointing toward a role in male fertility. Based on functional features and comparative sequence analysis, the *Sdic* protein was classified as an axonemal, rather than cytoplasmic, dynein intermediate chain (Nurminsky et al. 1998b). Genome engineering experiments coupled with phenotypic tests ultimately uncovered that the *Sdic* region boosts sperm competitive ability (Yeh et al. 2012), in line with its presumed adaptive nature (Kulathinal et al. 2004), making *Sdic* one of the few examples of a recently formed gene cluster that is unambiguously linked to sexual selection.

Due to its short age, highly tandemly repeated nature, and role in adaptive evolution, the *Sdic* multigene family has the potential to reveal key insights about the mode and tempo of the functional evolution that accompanies the formation and consolidation of similar gene clusters in the genome. However, the most recent release of the *D. melanogaster* genome sequence (Release 6) includes the presence of additional copies compared with the previous release (Release 5) (dos Santos et al. 2015), whereas functionally validated information only exists for one of the *Sdic* copies (Nurminsky et al. 1998b). Therefore, the actual structure of the *Sdic* cluster, and the extent to which the different copies exhibit identical functional attributes at the protein and expression levels, remain uncertain. Thus, resolving these questions is essential to evaluating whether the gene cluster is evolving in a concerted manner or has started a diversification process in which some of the copies have entered into a pseudogenization process. Additionally, a genome-wide analysis of the architecture of sexual antagonism in *D. melanogaster* indicated that the variable expression of one of the *Sdic* copies was associated with opposed effects on male and female fitness (Innocenti and Morrow 2010). In summary, the key structural and functional aspects of the *Sdic* gene cluster continue to remain elusive, impeding a correct analysis of the region's patterns of change and a precise view of its contribution to fitness.

Here we have investigated the evolutionary history of the constituent members of the *Sdic* gene cluster. This study first seeks to precisely reconstruct and annotate one the most challenging regions of the euchromatic fraction of the *D. melanogaster* genome by leveraging the increased resolution associated with long-read sequencing technologies, which have been shown to be instrumental in comprehensive studies of complex genomic regions including tandemly arranged duplicates (Huddleston et al. 2014; Krsticevic et al. 2015); second, to evaluate how different molecular mechanisms and evolutionary forces have shaped the current levels and patterns of DNA variability among the copies, ultimately recreating the most plausible scenario underlying the expansion of the cluster; and third, to determine the degree of functional diversification among different *Sdic* copies by performing a copy-specific monitoring of their expression, paying special attention to sex differences and a potential impact on female fitness.

We present a much more complex organizational and functional portrait of the evolution of the *Sdic* multigene family than previously thought (Nurminsky et al. 1998b; Ponce and Hartl 2006). For this, we devised analytical approaches tailored to accommodate the sequence similarity among the copies in order to leverage multiple available assemblies and preassemblies generated by long-read sequencing technologies (Kim et al. 2014; McCoy et al. 2014; Berlin et al. 2015) and RNA-seq datasets from different developmental stages and body parts (Graveley et al. 2011; Brown et al. 2014). We uncover differences with the current annotation of the *Sdic* region, both in number of copies and internal positioning (dos Santos et al. 2015). Our proposed evolutionary scenario for the formation of the *Sdic* multigene family involves a minimum of four unequal-crossing over events, pervasive gene conversion, and two episodes of positive selection. Despite the young age of this multigene family, we find clear signs of expression diversification across biological conditions with a varying expression breadth among its members, including expression in females although without resulting in decreased fertility according to phenotypic tests. Additionally, our results suggest that the *Sdic* protein may not function only as a sperm-specific axonemal dynein intermediate chain. Collectively, the *Sdic* multigene family epitomizes how quickly a tandemly arranged multigene family can functionally diversify at both the coding and regulatory levels, even in the face of gene conversion, through the acquisition of uneven sexually dimorphic expression.

Results

Assessing the Assembly of the *Sdic* Region

The *Sdic* region is located at 19C1 on the *X* chromosome and is composed of tandem repeats absent in other *Drosophila* species (supplementary fig. S1, Supplementary Material online). Each repeat consists of three parts of which the transcriptional unit that encodes the *Sdic* protein is the most relevant (fig. 1*B*). Releases 5 and 6 of the genome assembly of the ISO_1 strain differ considerably at the *Sdic* region (Hoskins et al. 2007; dos Santos et al. 2015). Release 5 included four copies of the *Sdic* repeat whereas Release 6 added three new copies (*CG46275*, *CG46276*, and *CG46277*; hereafter *SdicA*, *SdicB*, and *SdicC*, respectively), in addition to substantial sequence changes for copies *Sdic3* and *Sdic4* (fig. 1*B*; supplementary table S1, Supplementary Material online). This copy number increase is in good agreement with previous estimates at the molecular and computational levels (Benevolenskaya et al. 1995; Yeh et al. 2012). The fewer number of repeats in Release 5 could be the result of collapsed Sanger sequencing reads of high sequence identity.

To verify the organization of the *Sdic* region in Release 6, we examined other assemblies for the strain ISO_1 based on long sequencing reads (table 1 and supplementary text, Supplementary Material online). Long reads are more likely to harbor sequence stretches distinctive of particular individual or adjacent repeats, informing about their internal

Table 1. Organization of the *Sdic* Region of *D. melanogaster* in Different Assemblies.

Assembly	Sequence Technology	Number of Scaffolds*	Number of *Sdic* Copies	Copy Order (T. . .*AnxB10* ←←. . .←← *sw*. . .C)	Region Size (kb)[¶]
BAC10C18[a]	Sanger	1	4	*AnxB10* – 1 – 2 – 3 – 4 – *sw*	30.742
R6[b]	Sanger	1	7	*AnxB10* – 1 – 2 – 3 – A – B – C – 4 – *sw*	53.701
Berlin[c]	SMRT	1	6	*AnxB10* – 1 – 4 – 3 – B – C – 2 – *sw*	45.959
PBcR[d]	SMRT	1	6	*AnxB10* – 1 – 4 – 3 – B – C – 2 – *sw*	46.387
FALCON[e]	SMRT	2	4 (*0012*)	*AnxB10* – 1 – 4 – 3 – B – *sw*	30.391
			3 (*0143*)	*sw* – 2 – C – 3. . ..	22.688
SLR[f]	Illumina TruSeq	6	ctg100000969823	. . .4 –? –? . . .	NA
			ctg100000969503	. . .? –? . . .	
			ctg100000969502	*sw* -? . . .	
			ctg100000964644 (RC)	*AnxB10* – 1 – 4. . .	
			ctg100000964565 (RC)	. . ..? –?	
			431	. . .? –? –? –? . . .	

NOTE.—SMRT, single-molecule real-time. A, *CG46275*; B, *CG46276*; and C, *CG46277*. T, telomere; C, centromere.
[a]Hoskins et al. (2007); Release 5; GenBank accession number AC011705.11. BLASTn analysis indicates that this BAC includes the region upstream of *sw* at one end and 47 nt of *AnxB10* that are absent in *AnxB10-like* at the other.
[b]dos Santos et al. (2015); Release 6; GenBank assembly accession number: GCA_000001215.4.
[c]Berlin et al. (2015); GenBank assembly accession number: GCA_000778455.1.
[d]Kim et al. (2014).
[e]S. Koren and C.S. Chin, unpublished data. Contig IDs are indicated in brackets.
[f]McCoy et al. (2014); GenBank assembly accession number: GCA_000705575.1.
*Upon BLASTn using the exonic sequences of *Sdic1* in Release 6.
[¶]From the first nucleotide at the 5′ of the TE part of the most upstream *Sdic* repeat through the last nucleotide at the 3′ UTR of the most downstream *Sdic* repeat.

positioning. We examined four assemblies: three assembled from the same set of single-molecule real-time (SMRT) sequencing reads, differing only in their assembly methods (Kim et al. 2014; Berlin et al. 2015; S. Koren and C.S. Chin, unpublished data; see Material and Methods), and one obtained with Illumina TruSeq Synthetic Long-Reads (SLRs) (McCoy et al. 2014). Two of the SMRT-based assemblies, *Berlin* and *PBcR* hereafter (table 1), produced an unfragmented *Sdic* region (Kim et al. 2014; Berlin et al. 2015). Using a set of diagnostic sequence motifs for each *Sdic* copy (supplementary table S2, Supplementary Material online), we located all *Sdic* repeats in the assemblies and proceeded with their precise annotation. For the two unfragmented reconstructions, we found the same number of copies, arranged in the same fashion, although displaying some sequence differences. Critically, both reconstructions differ from Release 6 in having one less copy of the two that are identical in sequence (*Sdic3* and *SdicA*), as well as in the relative order of the copies, with *Sdic2* and *Sdic4* switching places (fig. 1*B*). Collectively, these results strongly support that the *Berlin* and *PBcR* assemblies should be considered as an alternative to Release 6 for the *Sdic* region, especially the former given the improvements associated with locality-sensitive hashing-based assemblies (Berlin et al. 2015).

Despite providing a fragmented assembly, the extremely low error rate associated with Illumina TruSeq sequencing (McCoy et al. 2014) makes SLRs especially appropriate to validate the reconstruction of the *Sdic* region in the Release 6 and *Berlin* assemblies (Berlin et al. 2015). The rationale is that the absence of differences between a particular SLR and one of the assemblies likely reflects the actual sequence in the ISO_1 strain. Using BLASTn, we retrieved 319 SLRs encompassing exonic sequences from the *Sdic* copies. Next, we filtered out reads that were so long that they contained the same region from two copies as assessed by Blast2seq (Johnson et al. 2008), which could lead to misassembly (Krsticevic et al. 2015), or so short that they did not retain motifs distinctive of individual copies. The combination of these criteria led us to consider 122 4–7.6 kb long SLRs, which were mapped against the two assemblies using BLASR (Chaisson and Tesler 2012) (supplementary fig. S2, Supplementary Material online). Most SLRs showed higher sequence identity in their alignment with one of the two assemblies, with 43 SLRs differing in which *Sdic* copy they were mapping against, which followed different patterns (supplementary table S3 and fig. S3, Supplementary Material online). Importantly, thorough scrutiny of the alignments revealed that the selected SLRs aligned more optimally with the *Berlin* assembly than with the Release 6 (supplementary fig. S4 and text, Supplementary Material online).

To determine the support level for each *Sdic* copy in the two assemblies, we focused on 107 SLRs showing high quality alignments and found a more even coverage across *Sdic* copies in the *Berlin* assembly (supplementary fig. S5 and text, Supplementary Material online). We also screened some diagnostic sequence stretches indicative of a more accurate reconstruction of the region. Specifically, we determined whether any SLR supported distinctive junctions (*Sdic1-Sdic2*, *Sdic2-Sdic3*, and *SdicC-Sdic4* in Release 6; *Sdic1-Sdic4*, *Sdic4-Sdic3*, and *SdicC-Sdic2* in the *Berlin* assembly) and same-copy differences in the two assemblies (supplementary table S4, Supplementary Material online). For both features, we found SLRs solely supporting the *Berlin* assembly. On balance, our results indicate that the *Berlin* assembly most accurately recapitulates the *Sdic* region in the ISO_1 strain.

Sequence Diversity

The six annotated copies of *Sdic* in the *Berlin* assembly (Berlin et al. 2015) range in nucleotide sequence identity percentage from 93.9% to 99.1%, with a median value of 97.6% from the

start to stop codons (supplementary table S5, Supplementary Material online). This identity level decreases only moderately when the whole gene fraction is considered (93.4–98.9%, median = 97.45%). From the transcriptional start to stop site, most nucleotide differences and indels accumulate in exons 4 and 5, the intron residing between them, and the 3'UTR. Only considering differences that result in amino acid replacements, excluding those due to frameshift mutations and deletions (see below), all nine non-synonymous changes found reside in exons 4 and 5, none of them being present across all *Sdic* copies. For the same alignable regions, only two synonymous changes are detected.

At the amino acid level, the sequence identity among the different *Sdic* protein variants ranges from 86.1% to 100%, with Sdic3 and SdicB being identical (supplementary fig. S6, Supplementary Material online). In terms of domain composition, the *Sdic* protein variants harbor either six or four WD40 motifs as confirmed by protein domain search in INTERPRO (supplementary fig. S6, Supplementary Material online); all sw proteins possess six WD40 motifs (supplementary fig. S6, Supplementary Material online). Based on the number of carboxyl end WD40 motifs, we grouped the putative *Sdic* proteins in two sets. The four WD40 motif-containing set includes Sdic1-PC and Sdic4-PE and is characterized by the shortest protein variants as a result of shifts in splice sites. Sdic1-PA also belongs within this first set of variants, although it exhibits a conspicuous structure as a result of three deletions in exon 5 (supplementary fig. S7, Supplementary Material online). Further, the six WD40 motif-containing set is characterized by a carboxyl end either identical to that of sw (all Sdic2 isoforms) or affected by several amino acid deletions and replacements (SdicB-PA, SdicC-PA, and Sdic3-PE, Sdic3-PF, Sdic3-PG). Importantly, the nucleotide differences that alter the donor splice site at the 3' end of exon 4 in *Sdic4* and *SdicC* also mediate the automatic conversion of ancestrally intronic sequence from *sw* into the *Sdic* coding sequence. In fact, for *SdicC*, the whole intronic sequence is read through such that it connects exons 4 and 5 (supplementary fig. S7, Supplementary Material online).

In addition to the WD40 motifs, all the Sdic and sw protein variants harbor a *cytoplasmic dynein 1 intermediate chain 1/2* domain (supplementary fig. S6, Supplementary Material online). Further, sequence comparison of the newly evolved N-terminus of the Sdic protein variants against other known axonemal dynein intermediate chain proteins revealed a negligible level of sequence similarity, which was in good agreement with the lack of significant matches in sequence similarity searches with BLASTp (Altschul et al. 1997). Collectively, these results are suggestive of a cytoplasmic role for the Sdic protein variants, without ruling out their function in the axoneme, which would take place through a non-canonical axonemal domain.

Molecular Evolution of the *Sdic* Multigene Family

The evolution of tandemly arranged gene duplicates often involves an initial phase driven by gene conversion, followed by a second phase where genetic drift and/or selection limit further sequence homogenization, enabling functional divergence (Fawcett and Innan 2011). Taking advantage of the validated *Berlin* assembly, we evaluated the relative contributions of gene conversion and adaptive diversification to the evolution of the six *Sdic* copies.

The analysis of the 5′–3′ distribution of the between-copy-variation supported the distinction of two broad sections within *Sdic*. The 5′ section begins at the transcription start site and ends at the 12 nt long gap present in the stretch that codes for the fourth WD40 domain. The 3′ section proceeds from this gap to the transcription stop site (supplementary fig. S8 and S13, Supplementary Material online). GeneConv (Sawyer 1989) revealed 23 statistically significant gene conversion tracts $P_{adj} < 0.05$), suggesting a scenario where the inner copies (*Sdic2*, *Sdic3*, *Sdic4*, *SdicC*, and *SdicB*) exchange DNA segments with each other, as well as the 5′ regions with *Sdic1*, and the 3′ regions with *sw* (supplementary table S6, Supplementary Material online). This is in line with the physical positions of *Sdic1* and *sw* as the most outermost genes in the region that are involved in these putative gene conversion events. Five out of the 23 gene conversion tracts show lengths larger than the maximum documented genome-wide in *D. melanogaster* (Casola et al. 2010). This unusual length may be due to the high *Sdic* sequence identity, which precludes the accurate delineation of converted tracts, resulting in the artifactual joining of adjacent stretches of exchanged DNA. Further, the boundaries of these converted tracts show a clear co-localization with the five likely recombination breakpoints inferred by ACG (O'Fallon 2013), which split *Sdic* into six partitions with independent evolutionary histories (P1-P6; fig. 2A). P1-P4 would correspond to the 5′ section of the *Sdic* sequence whereas the 3′ section would span P5–P6.

Overall, our results suggest that gene conversion is a major contributor to the shaping of the *Sdic* multigene family's pattern of variability. Nevertheless, the inspection of the local gene genealogies (fig. 2A) revealed that the statistical significance supporting the putatively converted DNA segments is partly driven by the accumulation of singletons (i.e., mutations in a single *Sdic* copy; long branches in the local genealogies of P1, P3, P5, and P6; fig. 2A). Given that all mutations are confined to one copy, GeneConv systematically infers that the remaining copies must be homogenizing their DNA sequences by exchanging DNA, a pattern also compatible with other evolutionary scenarios, including a relaxation of purifying selection and the action of positive selection. Using models especially devoted to quantifying the impact of natural selection on coding and non-coding regions (see "Material and Methods"), we found that all *Sdic* copies are evolving under purifying selection, with ~90–95% of their nucleotide positions being invariable or having substitutions rates lower than the synonymous substitution rate. However, the intensity of purifying selection does vary across copies and particularly across partitions. For example, the exonization of the intronic region of *sw* in *Sdic* likely resulted in a stochastic accumulation of mutations in the *sw* intron but not the homologous *Sdic* exon, from which they were purged. This is reflected as a long branch in the local genealogy of partition P1, a pattern that could mimic the signal of positive selection (*sw-AnxB10* branch in the P1 genealogy, fig. 2A).

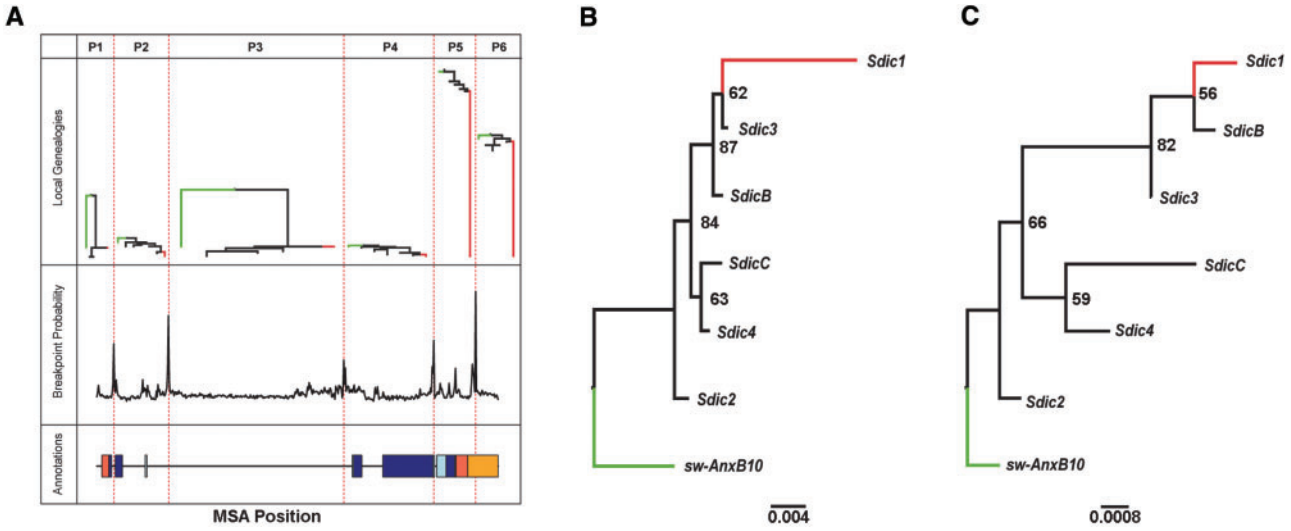

Fig. 2. Molecular evolution of the *Sdic* multigene family. (*A*) Top, local gene genealogies for each of the six DNA partitions (labeled by P1–P6) inferred with ACG. The DNA stretches from the different partitions are separated by recombination breakpoints depicted by a red dashed line. Using the exon–intron annotations of all copies except *Sdic4* as a reference, and after omitting stretches of sequence associated with deletions, partitions P5 harbors 11 non-synonymous and eight synonymous substitutions; partitions P1–P4 harbor 5 and 3, respectively. P6 does not include *Sdic4*, as this copy only contains missing data in this region. Middle panel, breakpoint posterior probability as estimated by ACG. Bottom panel, summarization of the exon–intron boundaries of *Sdic* following the color code in supplementary fig. S8, Supplementary Material online. MSA, multiple sequence alignment. (*B*) Maximum Likelihood phylogeny of the *Sdic* multigene family members, using a composite sequenced comprised of the homologous *sw* and *AnxB10* (*sw-AnxB10*) as an outgroup. The numbers in the internal nodes indicate the bootstrap support after 1,000 replicates. (*C*) Up-close view of the gene genealogy for the P4 partition. This partition has likely not exchanged information by gene conversion or been affected by other evolutionary forces that could potentially obscure the true duplication history of the *Sdic* gene copies. Local gene genealogies are represented with FigTree (http://tree.bio.ed.ac.uk/software/figtree/; last accessed December 1, 2015). Branches colored in red and green highlight *Sdic1* and *sw-AnxB10*, respectively. Scale bars indicate the number of nucleotide substitutions per site.

The test conducted is also especially robust at detecting positive selection in the face of potentially confounding factors, such as relaxed purifying selection or GC-biased gene conversion (see "Materials and Methods"). We identified two lineages showing statistical evidence for positive selection (supplementary table S7, Supplementary Material online). The first corresponds to the basal lineage leading to the ancestor of all *Sdic* copies in P1 and P3, and the second to the external lineage leading to *Sdic1* in P5. The first episode of positive selection occurred after the formation of the ancestral *Sdic* gene, probably driving mutations responsible for its expression to fixation, such as the acquisition of a translation start site. The second subsequent episode exclusively affected *Sdic1* in partition 5, which has accommodated multiple indels and other nucleotide differences that have led to multiple amino acid replacements (supplementary fig. S8, Supplementary Material online). Interestingly, partition P5 encompasses the constitutive fraction of the 3′UTR, which has undergone a profound remodeling of its miRNA binding site composition across copies, especially in the case of *Sdic1* (see below).

We tentatively reconstructed a scenario of duplications that leads to the contemporary organization of the *Sdic* region in the reference strain ISO_1 (fig. 3). For that, we took into consideration the phylogenetic relationship among the *Sdic* copies inferred from the gene tree topology exhibited by partition P4, as well as key shared diagnostic changes (e.g., in the promoter region –see below–). Unlike a gene topology based on the whole *Sdic* sequence, P4's topology has experienced limited gene conversion and does not exhibit singleton enrichment, and hence more faithfully recapitulates the evolutionary history of the duplication events and the correct gene tree topology of the family (Slightom et al. 1985; McGrath et al. 2009) (fig. 2*B*–*C*). The proposed scenario puts forward that upon formation of the ancestral *Sdic*, a duplication event took place giving rise to two copies. One of the two copies, the one adjacent to *sw*, would have evolved to what is known as *Sdic2*. In parallel, the other copy would have become duplicated again giving rise to two copies, the most downstream from *sw* being the ancestor of *Sdic1*, *Sdic3*, and *SdicB* (*Sdic1/3/B*), and the middle copy being the ancestor to *SdicC* and *Sdic4* (*SdicC/4*). Protocopies *Sdic1/3/B* and *SdicC/4* would have then duplicated jointly, increasing the number of copies from three to five, originating the precursors of *Sdic1* and *Sdic4* on the downstream side, and the ancestors of both *SdicC* and *Sdic3* and *SdicB* (*Sdic3/B*) near the middle of the cluster. An additional duplication of the protocopy *Sdic3/B* would have then occurred, giving rise to the precursors of *Sdic3* and *SdicB*. Only the temporal sequence of origination of *Sdic1*, *Sdic3*, and *SdicB* conflicts with their phylogenetic relationship, which suggests a different sequence of events: *Sdic1/3/B* → *Sdic3* and *Sdic1/B*, then *Sdic1/B* → *Sdic1* and *SdicB*. Nevertheless, the ancestral node joining *Sdic1*, *Sdic3*, and *SdicB* exhibits a low bootstrap value being this parsimonious scenario also supported by the occurrence of 0 amino acid replacements and 13 silent changes between *Sdic3* and *SdicB*.

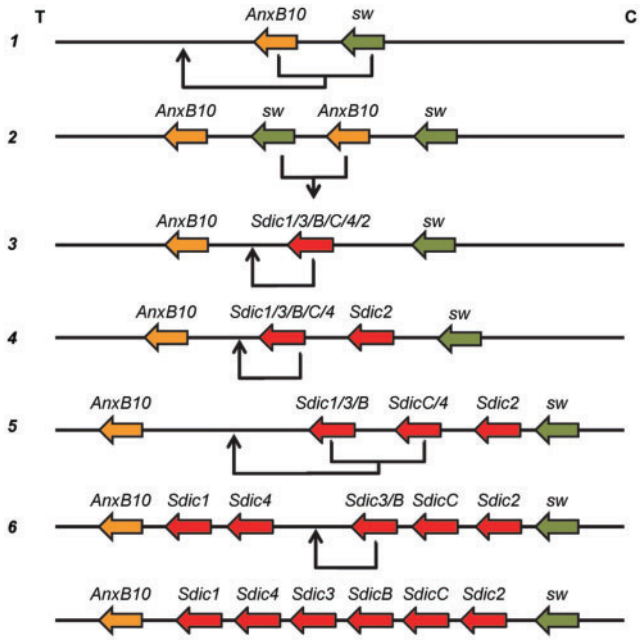

Fig. 3. Most parsimonious reconstruction of the formation of the *Sdic* region. An unequal crossing-over event between regions upstream of *sw* and downstream of *AnxB10* resulted in a segmental duplication of *sw* and *AnxB10*, although other more complex rearrangement scenarios cannot be ruled out (Bauters et al. 2008) (*1*). This was followed by the creation of the ancestral *Sdic* copy (*Sdic1/3/B/C/4/2*) through a series of mutations, which notably involved a large deletion event involving the middle copies of *sw* and *AnxB10* (*2*); a TE also became inserted upstream of the ancestral *Sdic* copy (data not shown). An unequal crossing-over event involving sequence stretches upstream and downstream of the ancestral *Sdic*, but in different homologous chromosomes, would have then resulted in a tandem duplication of the ancestral *Sdic* copy (*3*). Next, a similar unequal crossing-over event resulted in the tandem duplication of the *Sdic* copy closest to *AnxB10* (*4*). Subsequently, a third unequal crossing-over event occurred amid the region between *AnxB10* and its closest copy and the region between the two copies closest to *sw* resulting in a tandem duplication of the two copies closest to *AnxB10* (*5*). Finally, a fourth unequal crossing-over event resulted in a single-copy tandem duplication leading to the formation of the sixth *Sdic* copy (*6*). Several gene conversion events have likely occurred between *Sdic* copies. After step 3, it is uncertain where the unequal crossing-over events occurred due to the high similarity of the copies. This proposed scenario is in overall good agreement with the phylogenetic tree in fig. 2C, with the exception of the sequential generation of *Sdic1*, *Sdic3*, and *SdicB*. Nevertheless, this tree exhibits low bootstrap values. Black arrows, duplication events. T, telomere; C, centromere.

In the proposed scenario, the tandem duplication of the *Sdic* region would have come about via four unequal crossing-over events.

Expression Diversification among *Sdic* Copies

Previous characterization of *Sdic* expression was limited to *Sdic1* ((Nurminsky et al. 1998b; Mikhaylova and Nurminsky 2011). To evaluate potential expression differences among *Sdic* copies, we focused on two amplicons for which the design of specific primers was more feasible. One amplicon is associated exclusively with *Sdic1* whereas the other is shared between *Sdic4* and *SdicC* (hereafter *Sdic**). RT-PCR experiments with the OR-R strain uncovered that both *Sdic1* and *Sdic** are expressed in not just testes, but also ovaries, demonstrating that expression of these copies is not male specific (supplementary fig. S9, Supplementary Material online). *Sdic* female expression was also reproduced in the African strain ZW-109 (supplementary fig. S10, Supplementary Material online). Furthermore, we detected expression of both amplicons in both male and female heads (supplementary fig. S9, Supplementary Material online). In order to better quantify expression differences across tissues, sexes, and strains, we performed qRT-PCR experiments. The results confirmed high expression levels of *Sdic1* and *Sdic4* in testes from the two strains, as well as lower expression levels in ovaries and heads from both sexes (supplementary table S8 and fig. S11, Supplementary Material online). Interestingly, in ZW-109, *Sdic4*, but not *Sdic1*, was overexpressed in male relative to female heads, a pattern not observed for OR-R. These results support a much more complex spatial expression profile for *Sdic* than previously reported (Nurminsky et al. 1998b).

Even if no disruptive amino acid replacement or premature stop codon has altered the functionality of the different *Sdic* protein variants, the pseudogenization of some of the copies can arise from mutations within the promoter region. We observe two nucleotide differences in the promoter region of *Sdic3* and *SdicB* in relation to the remaining *Sdic* copies (supplementary fig. S12, Supplementary Material online). These two nucleotide differences were confirmed in *Sdic3* and *SdicB* by 3 and 4 SLRs, respectively. Importantly, one of these differences falls within a sequence stretch that is similar to a motif in the *βTub85D* gene promoter responsible for testis-expression specificity (Michiels et al. 1989).

In order to both determine the potential impact of the nucleotide differences within the promoter region and generate a more comprehensive expression profile of the *Sdic* copies, we searched for copy-specific motifs and scrutinized their presence—no mismatch allowed—across ~3.15 billion RNA-seq reads representing 59 biological samples from different anatomical parts and developmental timepoints (Graveley et al. 2011; Brown et al. 2014). This measure was necessary as many reads have the potential to map against several *Sdic* copies or *sw*. After corroborating their absence in *sw*, five motifs were delineated within the most 3′ third of *Sdic1*, *Sdic2*, *Sdic3*, *Sdic4*, and *SdicC* (supplementary table S10 and fig. S13, Supplementary Material online); no informative motif was found for *SdicB*.

Given the conservative nature of our approach, we pooled all reads from the libraries associated with the same biological condition. In this way, we maximized our capability to detect reads containing the diagnostic motifs, which was used as evidence of expression. The number of reads for which we detected perfect alignments, corrected by the sequencing depth of the biological condition in question, was adopted as proxy for expression level (supplementary table S9, Supplementary Material online). In spite of limitations derived from, for example, the fact that some motifs have the potential to survey more than one transcript for a particular

copy whereas others are specific to a single mRNA transcript variant, it was possible to uncover distinctive characteristics for the expression profile of the different *Sdic* copies (fig. 4*A*–*B*, supplementary fig. S14, Supplementary Material online).

We found evidence of expression for all five copies surveyed, which, combined with the absence of premature stop codons and evidence of purifying selection, reinforces the notion that none of the *Sdic* copies has entered into a pseudogenization process in the ISO_1 strain. From the developmental perspective, all copies showed sustained expression from third instar larvae throughout adulthood, although episodic expression of *Sdic3* was detected in earlier developmental stages. The expression level of the *Sdic* copies increases during the pupal stage, reaching maximum values in 5-day-old males, which correlates well with the testes expression evidence obtained via RT- and qRT-PCR experiments for particular *Sdic* copies. In fact, it is in samples unambiguously linked to males only (eight out of 59) that all *Sdic* copies show their highest expression levels. Considering the six samples (three developmental and three anatomical, roughly 10% of the total) in which each copy shows the highest expression levels, we find *Sdic1* and *Sdic4* displaying the most marked trend, with five out of the six samples being linked to males. Among the anatomical samples linked to males, *Sdic1* stands out by showing its highest expression levels in testes and accessory glands of 4-day-old males, whereas *Sdic3* showed its highest expression levels in head samples from males of different ages. Further, although the developmental samples do not show evidence of systematic expression of the *Sdic* copies in females, the anatomical samples clearly show evidence for the expression of *Sdic3* in eight out of 11 samples unambiguously linked to females. Interestingly, we detect profound variation among *Sdic* copies in their contribution to the expression profile of particular biological conditions

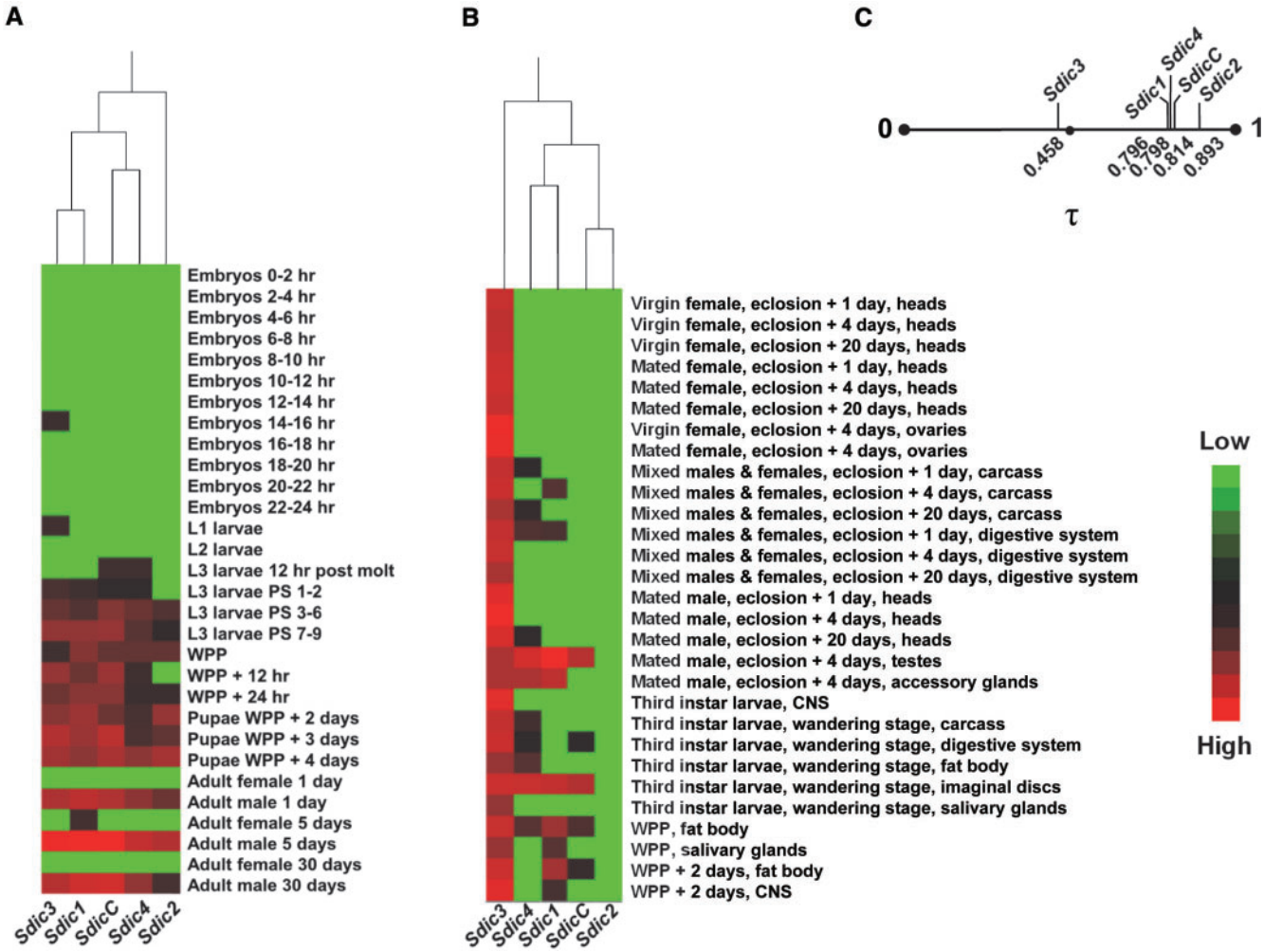

Fig. 4. Expression profile of five *Sdic* copies. Heatmap for developmental stages (*A*) and anatomical samples (*B*) showing evidence of expression diversification among the *Sdic* copies surveyed. Red, high expression; black, intermediate expression; green, lower expression. Fifty-nine biological conditions were examined. The data were obtained in two different large-scale expression surveys (Graveley et al. 2011; Brown et al. 2014), which might differ in their power to detect lowly expressed transcripts, even in similar, although not identical, conditions. (*C*) Expression specificity, τ, upon considering all conditions. τ values range from 0 to 1, with higher values corresponding to more restricted expression and lower values to broader expression across conditions (Yanai et al. 2005). Log10 normalized expression values were used in the analyses. Examples of the detected reads in relevant conditions are provided in supplementary fig. S14, Supplementary Material online. CNS, central nervous system; hr, hour; Lx, larval stage x; PS, puff stage; WPP, white prepupae.

not previously shown for this multigene family. For example, *Sdic3* contributes disproportionately more to the global expression of *Sdic* in the central nervous system of third instar larvae and 2-day-old white prepupae than any other copy. Likewise, we find marked differences in expression specificity values (τ) among copies (fig. 4C). In fact, Monte Carlo simulations showed that *Sdic3* possesses a significantly wider expression breadth (i.e., lower τ value) than the rest of the assayed copies ($P < 0.001$).

Variation in expression attributes among the *Sdic* copies can arise through both the pre- and post-transcriptional regulation. The currently annotated promoter sequences are virtually identical barring two nucleotides substitutions. These sequence changes differentiate *Sdic3* and *SdicB* from the rest of the copies, which could result in differential competing ability to recruit transcriptional machinery in the particular biological conditions in which the constituents of this machinery are in limited concentrations. In fact, *Sdic3* exhibits a clearly different expression breadth compared with the rest of the surveryed copies. Alternatively, differences in expression attributes could result from the recruitment of a slightly different set of downstream regulators. This might have happened through the severe 3′UTR remodeling across *Sdic* copies, resulting in differential post-transcriptional regulation via microRNAs. To explore this, we scanned the 3′ UTRs of all *Sdic* and *sw* transcripts for canonical miRNA target sites. We identified target sites for up to 54 distinct mature microRNAs (supplementary table S11, Supplementary Material online). By considering the gain/loss profile of orthologous miRNA target sites, we observed that only four target sites were conserved across all *Sdic* and *sw* transcripts. In fact, *sw* and *Sdic2* had a very similar targeting profile (supplementary fig. S15A, Supplementary Material online), suggesting a profound remodeling process of the 3′UTRs occurred after the divergence between *Sdic2* and the rest of *Sdic* copies (supplementary fig. S15B, Supplementary Material online). *Sdic1*, the copy characterized by the most male-biased profile, also exhibits the most markedly different miRNA binding site profile. *Sdic1* has the largest number of specific, novel target sites (14), harboring sites in exclusive for 10 miRNAs. Overall, we observed regulated *Sdic* expression throughout development and across body parts, the absence of expression silencing, and incipient differences among copies. How the interplay between promoter differences and remodeled 3′UTR miRNA binding site compositions contribute to the observed expression differences is not apparent at this time.

The *Sdic* Region and Female Fertility

All *Sdic* copies are expressed in males whereas 3–4 copies (*Sdic1*, *Sdic3*, and either *Sdic4*, *SdicC*, or both) show expression in females. Further, microarray experiments coupled with hemiclonal analysis pointed to *Sdic3*, now several copies based on our improved annotation, as a locus that displays sexual antagonism with regard to variable gene expression (Innocenti and Morrow 2010); *sw* did not show this pattern. As the *Sdic* region enhances sperm competitive ability (Yeh et al. 2012), this opens the possibility that the *Sdic* region as a whole can have an opposed effect on the fitness of the sexes. We examined the effects of deleting the *Sdic* region in females under the hypothesis that there would be a fitness boost if *Sdic* expression impairs female fertility.

We generated synthetic genotypes for the *Sdic* region using previously engineered deletions of the entire *Sdic* region via non-homologous recombination (Yeh et al. 2012) (supplementary fig. S16A, Supplementary Material online). This was done upon reassuring that the changes introduced to the annotation of the *Sdic* region were compatible with no *Sdic* copy remaining in *X*(19C1), which could compromise the interpretation of any phenotypic test (supplementary fig. S17, Supplementary Material online). We assayed three relevant parameters for female fertility: female productivity, i.e., the progeny number; number of eggs laid; and egg hatching rate. Homozygous females for the deletion of the *Sdic* region (A^{-d} and E^{-d}) were compared against wild-type females for the region (B^{+} and I^{+}) by monitoring differences in female productivity over a 33-day-period (Methods and supplementary fig. S16B, Supplementary Material online). The knock-out strains did not exhibit increased productivity relative to their wild-type counterparts and w^{1118}, another control strain (supplementary table S12, Supplementary Material online). We found statistically significant differences in each timepoint examined, but they mostly resulted from a consistently low productivity of the wild-type control I^{+} (supplementary table S12, Supplementary Material online). In relation to the other two wild-type strains B^{+} and w^{1118}, the knock-out strains E^{-d} and A^{-d} did not show any consistent pattern, with at least one of them displaying no significant differences in productivity for most of the timepoints assayed.

No difference in productivity among females with and without the *Sdic* region could result from counteracting factors, e.g., a higher number of eggs laid being offset by a lower hatching rate. We tested for differences in these two parameters over a 6-day period and found no evidence that the absence of the *Sdic* region correlates with a higher number of eggs laid or a higher hatching rate (supplementary table S13-S14 and fig. S16C, Supplementary Material online). Failure to find statistically significant differences could result from a lack of power due to limited sample size, particularly in the case of hatching rate. However, the global trend seems to be robust, with two of the wild-type strains (B^{+} and w^{1118}) showing very similar values to those of the knockout strains. Overall, these results indicate that *Sdic* expression in females does not impair the fertility of this sex, which does not exclude that it can impact negatively other fitness traits.

Discussion

Our analysis of the *Sdic* region in *D. melanogaster* represents a step forward in the generation of accurate portraits of the organizational, sequence, and functional evolution of recently originated, tandemly arranged multigene families. This is needed as our current knowledge is primarily based on tandemly arrange families of ancient origin such as the globins or rRNA genes (Brown et al. 1972; Zimmer et al. 1980), cases involving young tandem duplicates with a limited number of members (Osada and Innan 2008), or cases in which the

functional data is limited or lacking (Moore and Purugganan 2003). Genomic regions harboring recently expanded gene clusters are hotspots for structural and functional change, having the potential to foster adaptive evolution (Brown et al. 1998; Newcomb et al. 2005; Perry et al. 2007; Jugulam et al. 2014). By coupling long-read sequencing technologies (Eid et al. 2009) with RNA-seq data from multiple biological conditions, and tailored analytical approaches that accommodate the particularities of members of these type of multigene families, we can now perform unparalleled multilevel characterizations of these complex genomic regions.

At the organization level, the combined use of different long-sequencing read technologies has prompted us to propose a different organization for the *Sdic* multigene family in the ISO_1 strain from the one currently accepted (dos Santos et al. 2015). This alternative organization differs in both number and internal arrangement of the copies. To account for the six copies in this alternative organization, we propose a duplication scenario involving a minimum of four unequal crossing-over events. Further, the inter-copy variability patterns are compatible with a scenario of rampant inter-locus gene conversion, especially involving the outermost members of the cluster. Despite the homogenizing effects of gene conversion, we found a preferential accumulation of mutations towards the 3′ end of the *Sdic* copies, affecting both coding and non-coding sequence, which would have been driven partially by positive selection. Examples of positive selection overcoming the effects of gene conversion have also been documented for other recently originated tandem duplicates (Innan 2003; Osada and Innan 2008). Importantly, the role of positive selection in shaping the patterns of nucleotide polymorphism and divergence in the *Sdic* region has been controversial (Brookfield 2001; Kulathinal et al. 2004). We found evidence that copy differentiation at the sequence level is compatible with at least two episodes of positive selection, one shortly after the origin of the ancestral copy, and a more recent episode exclusively affecting the 3′ end of one copy (*Sdic1*). These signatures of positive selection and the lack of evidence for pseudogenization of the *Sdic* copies scrutinized provide strong support to the adaptive role of *Sdic*.

The six copies documented encode a variety of *Sdic* proteins which differ primarily at their C-terminus, where the protein sw presumably interacts with the dynein heavy chain, as inferred from its ortholog in *Dictyostelium* (*dicA*; Ma et al. 1999). Importantly, all Sdic and sw variants possess a common *cytoplasmic dynein 1 intermediate chain 1/2* domain, suggesting Sdic could function similarly to sw. However, the lack of coiled-coil and serine-rich domains at the N-terminus of Sdic would presumably prevent the Sdic variants from interacting with the dynactin protein complex, which mediates the interaction of the dynein protein complex with a variety of subcellular structures (Nurminsky et al. 1998a; Ma et al. 1999). Overall, Sdic and sw might share a limited set of common interactions with other protein complex subunits and subcellular structures. In fact, these structural differences, and the expression profile exhibited by some *Sdic* copies, are suggestive of a Sdic protein that interacts with non-axonemal dynein complexes present in tissues possessing both ciliated (e.g., sperm) and non-ciliated cells (e.g., salivary glands and imaginal discs). Whether or not Sdic interacts with axonemal dynein complexes cannot be inferred from our results, but the fact that the silencing of the whole multigene family results in a significant reduction in sperm competitive ability does not allow us to discard this possibility (Yeh et al. 2012).

The *Sdic* multigene family shows a pattern of expression consistent with quick regulatory diversification among copies. As is the case for other recently originated genes, *Sdic* was likely expressed in testes at a very early stage (Kaessmann 2010; Zhao et al. 2014). This is the only expression attribute in adults shared across all copies, whereas expression in females was displayed by 3–4 copies, varying across adult samples, including some (*Sdic1* and *Sdic3*) that were inferred to be among the most recently generated in the gene family. *Sdic*'s testis expression could have resulted from a rather simple promoter motif with incipient testis-biased expression (Nurminsky et al. 1998b; FitzGerald et al. 2006), a benign molecular environment (Schmidt and Schibler 1995; Sassone-Corsi 2002), or both. Subsequently, selective pressures such as post-mating male–male competition (Kleene 2005; Singh and Kulathinal 2005) would have mediated the retention and expansion of *Sdic*, as supported by phenotypic assays (Yeh et al. 2012). Exactly when the broadening of expression took place relative to the origination of some the copies is unclear at this time, as is how the differences in promoter sequence and 3′UTR miRNA binding site composition led to the observed expression differences. Nevertheless, these unclarified aspects point to some interesting directions. First, whereas functional broadening over evolutionary time is a hallmark of many old duplicates (Assis and Bachtrog 2013; Kaessmann 2010), including expression in both sexes, *Sdic3* highlights how quickly this broadening trend can occur. Second, functional diversification of tandemly arranged duplicates might proceed through post-transcriptional regulatory changes driven by the evolution of a unique composition of miRNA binding sites (Wang and Adams 2015), as could be the case for *Sdic1*, revealing an important path for the diversification of DNA-mediated duplicates.

The functional complexity of the *Sdic* copies, revealed here through their protein domain compositions and expression profiles, questions whether the phenotypic impact of the *Sdic* region is confined to post-mating male–male competition. It is possible that *Sdic* expression in females can result in a sexually antagonistic effect as circumstantial evidence suggests (Innocenti and Morrow 2010), fitting into the notion that the *X* chromosome, where *Sdic* resides, is a key genomic reservoir of sexually antagonistic genetic variation (Rice 1984; Gibson et al. 2002). Our results for three parameters of female fertility suggest that should this antagonistic effect exist, it impacts either a more subtle fertility component or a completely different type of trait from those tested here.

Regardless of the organismic impact of the *Sdic* region, our results show that the amplification of *Sdic* has not consisted merely in a gene dosage increase. Nevertheless, it remains a challenge to fully understand the evolutionary implications of the *Sdic* amplification. We hypothesize that the Sdic protein

could have facilitated the emergence of a secondary, unrefined function of sw (Hughes 1994) or novel interactions between the dynein complex and other protein complexes or cellular components via the novel N-terminus. Additionally, sw has been shown to interact with the p150-Glued subunit of dynactin in a dosage-dependent manner, suggesting that Sdic, which is essentially identical to sw but cannot bind the p150-Glued subunit, could act as a competitive inhibitor of the interaction between the dynein and dynactin complexes (Boylan et al. 2000). Whether it is because of an enhanced secondary or an entirely novel function, the benefit of *Sdic* could have become more apparent upon its overexpression via copy number increase (Bergthorsson et al. 2007), with some of the copies subsequently undertaking different paths of evolutionary tinkering. This pattern is compatible with the variation in domain composition and expression profiles seen for the *Sdic* copies in the ISO_1 strain. Equivalent multilevel characterization of the *Sdic* gene cluster in other *D. melanogaster* strains as performed here will help gauge some key aspects. The first is whether *Sdic*'s functional refinement is still ongoing, with some of the copies possibly undergoing pseudogenization, or alternatively whether the existing copies are part of a diversification process associated with balancing selection, both scenarios driven by the permanent action of sexual selection. The second aspect is whether there is an optimal range of copies refractory to the extreme outcomes of unequal crossing-over, i.e., the complete loss of *Sdic* or an unbearably high copy number which would both be detrimental.

Materials and Methods

Assembly and Annotation Analysis

All assemblies used are associated with sequencing experiments that made use of the ISO_1 isogenic strain *y; cn bw sp* (Adams et al. 2000). These include: the complete sequence of BAC10C18 (GenBank accession number AC011705.11); Release 6 plus ISO1 MT (GCA_000001215.4; dos Santos et al. 2015); assembly ASM77845v1, which is based on SMRT sequencing reads ASM77845v1 (GCA_000778455.1; Berlin et al. 2015); and an assembly based on Illumina TruSeq SLRs (GCA_000705575.1; McCoy et al. 2014). The assembly ASM77845v1 was generated using the Celera assembler (v8.2) and MHAP as overlapper. Using the same reads as assembly ASM77845v1, two additional preassemblies just differing in computational pipeline aspects, were included. The preassembly reported in Kim et al. (2014) uses the overlapper implemented in the HGAP (hierarchical genome assembly process) pipeline and can be retrieved from http://cbcb.umd.edu/software/pbcr/dmel_cons_asm.tar.gz (last accessed December 1, 2015). The other SMRT based preassembly was generated using the FALCON v0.1 assembler, which can be retrieved from https://s3.amazonaws.com/datasets.pacb.com/2014/Drosophila/reads/dmel_FALCON_diploid_assembly.tgz (last accessed December 1, 2014). Contigs containing *Sdic* copies that are part of different assemblies were identified using *Bowtie2* v2.2.3 (Langmead and Salzberg 2012) under parameter settings *–fast-local* and *–no-unal*, whereas using the sequences of the annotated exons of the *Sdic* copies in Release 6 as a query. The annotation of the *Sdic* region in the assembly GCA_000778455.1 was done taking the gene structure of each *Sdic* copy in Release 6 as a reference.

In the case of the scrutiny of SLRs to test the validity of particular assemblies, FASTQ files (Dm4-1 to Dm4-3, and Dm5-1 to Dm5-3) were downloaded from the Illumina BaseSpace site and tested for significant similarity with *Sdic* exonic sequences using BLASTn v2.2.30 (Altschul et al. 1990). The mapping of SLRs against particular assemblies was done using BLASR v1.3.1 (Chaisson and Tesler 2012) under the default minimum percent identity and setting *-bestn 1* in order to prevent multiple alignments. Prior to this, the *Sdic* region in each assembly under comparison was indexed using the program *sawriter*, which is part of the SMRT Analysis toolkit available at the Pacific Biosciences Developer's Community Network Website (DevNet: http://www.smrtcommunity.com/DevNet; last accessed December 1, 2015). TABLET v1.14.10.20 (Milne et al. 2013) was used for alignment visualization and confirmation of key motifs.

Molecular Evolution Mode

A multiple sequence alignment (MSA) composed of the six *Sdic* copies, from the start of the promoter to the end of the 3′UTR, was assembled including as well an artificial composite sequence comprised of the homologous *sw* and *AnxB10* regions (*sw-AnxB10*) as an outgroup. Using MEGA v6.06 (Tamura et al. 2013), sequence alignments were performed with MUSCLE and refined by visual inspection. Levels of divergence along the sequence alignment, plus the number of synonymous and non-synonymous substitutions, were calculated with DnaSP v5 (Librado and Rozas 2009). The maximum likelihood (ML) phylogenetic tree was reconstructed using RAxML v8.12 (Stamatakis 2014) with 1,000 bootstrap replicates.

Gene conversion tracts were inferred using the GeneConv program (Sawyer 1989) under the assumption that no nucleotide mismatches occurred among the tracts, reflecting the negligible probability of these events happening during the very early evolutionary stages of a multigene family like *Sdic*. We applied the Bonferroni correction to obtain the adjusted probability with which a particular tract experienced gene conversion. As GeneConv tracts might modify the local gene genealogy, we further examined whether *Sdic* exhibits incongruent gene genealogies along its sequence by estimating the recombination breakpoints with the ACG program (O'Fallon 2013), which implements explicit models that fully capture the coalescent process with recombination. The ACG Markov chain was run for 20,000,000 iterations, with a burn-in period of 5,000,000.

The HyPhy batch script, written by Oliver Fredigo (https://github.com/ofedrigo/TestForPositiveSelection/blob/master/nonCodingSelection.bf; last accessed January 15, 2016), was used to test for positive selection acting on specific *Sdic* copies (Haygood et al. 2007). This script evaluates whether the substitution rate in a focal class of sites, which can be comprised of any kind of functional category, is higher than in a neutral class of sites (here represented by the synonymous sites). The

statistical significance of this test is assessed by comparing two nested models by means of a Likelihood Ratio Test (LRT). The null model assumes three classes of sites, including positions that are (i) selectively neutral, (ii) evolving under purifying selection, or (iii) purged in background lineages, but neutrally evolving in the foreground branch. The alternative model replaces class (iii) with two extra classes that assume a fraction of the sites are evolving under positive selection in the foreground lineages, but under either (iv) neutral or (v) purifying selection in the background lineages. Thus, this test enables distinguish between positive and relaxed purifying selection, as the latter is already accounted for in the null model. To accommodate for the different gene tree topologies found for each partition along the MSA, this test was separately conducted for each of the *Sdic* sequence partitions identified by the ACG recombination breakpoints. Exclusively for this analysis, we included a second artificial composite sequence comprised of the orthologous stretches to *sw* and *AnxB10* in *D. simulans*, which was used as a more external outgroup. This enabled to clearly distinguish, within each partition, whether basal episodes of positive selection occurred in the lineage leading to the ancestor to all *Sdic* copies or in that leading to the *D. melanogaster* composite *sw-AnxB10*.

Strains and Fly Husbandry

D. melanogaster strains used are listed in supplementary table S15, Supplementary Material online. Flies were reared on dextrose-cornmeal-yeast medium in a 25C chamber under constant lighting conditions. Adult virgins were collected within 6 h of eclosion, sorted by sex, and then cultured separately in groups of $\leq$10 individuals. At 4–6 days post-eclosion, entire adult whole bodies and other dissected biological samples (male and female heads, testes, and ovaries) were homogenized and stored in TRIzol (Life Technologies) at -80 °C. Dissections were done separately for each type of biological sample in ice-cold 1× PBS solution. All sorting, scoring, collecting, counting, and manipulation of flies was performed under CO_2 anesthesia.

Total RNA Extraction and cDNA Synthesis

For the strains Oregon-R and Zimbabwe-109, total RNA was extracted from three biological replicates corresponding to each strain by sex by tissue combination. Following manufacturer's instructions, total RNA was extracted from tissues previously homogenized in TRIzol. DNA traces were removed by treating 10 µg of each sample with Turbo DNA-free DNase (Ambion). RNA integrity and purity were confirmed using gel electrophoresis and a NanoDrop spectrophotometer respectively. cDNAs for each sample were generated using 1 µg of DNase-treated total RNA, oligo(dT) primers, and SuperScript III reverse transcriptase (Invitrogen) in the presence of RNaseOUT recombinant RNase inhibitor (Invitrogen). All female samples were tested for male contamination by RT-PCR of the *Y*-linked gene *CG41561*. cDNA quality was confirmed by RT-PCR of *Gapdh2*.

PCR-Based Expression Profiling

RT-PCRs were performed using TaKaRa Ex Taq polymerase (Clontech), 2 µL cDNA template, and appropriate primers. The correct identity of each amplicon was confirmed by gel electrophoresis, Sanger sequencing, and subsequent BLASTn analysis. qRT-PCR experiments were performed essentially as described (Yeh et al. 2014). Possible reference genes were selected based on their expression stability as shown by modENCODE RNA-seq data in FlyBase (dos Santos et al. 2015), as well as the expression profile between the sexes as reported in the Sex Bias Gene Expression Database (Gnad and Parsch 2006). Subsequent verification of expression stability, as indicated by the GeNorm program (Statminer, TIBCO Spotfire suite v6.5.3 -Perkin Elmer-), led us to use two reference genes: *clot* and *CG14903*. Estimates for expression differences were obtained using the $-2^{\Delta\Delta Cq}$ method (Livak and Schmittgen 2001). *P*-values were calculated using the Limma moderate *t*-test (Smyth 2004) within the Statminer package and the Benjamini-Hochberg multiple test correction (Benjamini and Hochberg 1995). Each normalized Ct value, x_i, was transformed according to:

$$(-1 \times \log_b y_i) + 1$$

where $y_i = (x_i + |a| + 1)$, a is the minimum value in the range of initial normalized Ct values ($x_1, \ldots, x_n$), and b is the maximum of the initially adjusted values ($x_i + |a| + 1, \ldots, x_n + |a| + 1$). Accordingly, the highest normalized Ct value is scaled to 0 and the lowest to 1. Primers used are listed in supplementary table S16, Supplementary Material online.

RNA-seq Analysis

Ninety-six SRA files corresponding to 59 types of biological samples were retrieved from NCBI using the SRA Toolkit (Graveley et al. 2011; Brown et al. 2014). Reads with remaining adapters, with a percentage of N sites >10%, or with >=50% nucleotides with a quality value $Q \leq 5$ were discarded. One diagnostic motif, a sequence unique to a specific *Sdic* copy, for each of the *Sdic* copies (excluding *SdicB*, for which none could be found) was extended both upstream and downstream up to a total length of 130 nt. All reads from all libraries were then examined for a perfect alignment involving $\geq$76 nt with each of the extended diagnostic motifs using TopHat 2.0.12 (Kim et al. 2013), making sure that the core diagnostic motif was always included. Raw counts per library were obtained using a custom shell script. The level of expression was estimated as the number of reads per kilobase per million reads (RPKM; Mortazavi et al. 2008), although in this case the variable length has no effect since all the motifs are 130 nt long. Within-biological-sample normalized expression values were subsequently log10 transformed. Heatmaps were generated by hierarchical clustering on principal components using FactoMineR (Lê et al. 2008; Diaz-Castillo et al. 2012). Expression specificity, τ, was quantified as described (Yanai et al. 2005). For the Monte Carlo simulation analysis, log10 transformed normalized expression values were shuffled 10,000 times and τ was recalculated each time for each

copy. The resulting dataset allowed for calculating the probability of obtaining by chance alone a τ larger or equal to the one observed.

MicroRNA Binding Site Composition

3′UTR sequences were extracted for all *Sdic* transcripts according to our annotation, and for all *sw* transcripts according to FlyBase (dos Santos et al. 2015). The presence of canonical microRNA sites (7mer-A1, 7mer-m8, and 8mer) as previously described (Bartel 2009), was examined using an in-home Perl script and the current microRNA annotation of *D. melanogaster* in miRBase v.21 (Kozomara and Griffiths-Jones 2014). Gains/losses of microRNA target sites were mapped to the *Sdic* phylogeny using the Dollo v3.695 parsimony method implemented in PHYLIP (Felsenstein 2005).

Phenotypic Assays

For the productivity assay, virgin females either possessing (A^{+}, I^{+}) or devoid (B^{-d}, E^{-d}) of the *Sdic* region of the *X* chromosome were crossed with naïve wild-type males of the Oregon-R strain. Females from the strain w^{1118} were also used as a control for productivity levels of the source genetic background used to create the engineered strains used here (Yeh et al. 2012). Three naïve Oregon-R males were aged to 5 days old then mated to three 1-day-old virgin females from each of the experimental and control strains. Twenty-five replicates of each mating pair were assembled and the adult individuals were transferred to a fresh vial every other day. To compensate for decreasing male fecundity with age, males were removed on day 15 and replaced with another four males, which were in turn removed on day 29. The total progeny emerged from each vial associated with days 1, 3, 11, 13, 21, 31, and 33 was recorded. The progeny number produced was normalized by the number of females still alive at the moment of transferring from the vial associated with that particular day.

In the case of the egg-laying and egg-hatching assays, 10 five-day-old Oregon-R naïve males were mated separately to 10 virgin females of the same age from each of the five strains under comparison for 24 h. Three replicates of each of these crosses were set up. Petri dishes with grape-juice agar were used for easy egg detection against a dark background. To induce egg-laying, yeast was added to the agar (Waskar et al. 2005). Additionally, several scratches were made on the surface of the agar to increase surface area (Atkinson 1983). The adults of each replicate were transferred to a new plate every 24 h for 5 consecutive days and discarded on day 6. The egg number on each plate was recorded immediately after the adults were removed. After incubating for an additional 24 h, the plates were reexamined for unhatched eggs, the number of which was also recorded. These data was used to calculate the hatching rate and the number of eggs laid per female. JMP 12.1 (SAS Institute) was used for statistical analyses.

In Situ Hybridization

A ~4.23 kb *Sdic* genomic fragment present in all *Sdic* copies was generated by PCR and Sanger sequenced for verification. Probe labeling and hybridization on polytene chromosome squashes was performed as described (Ranz et al. 1997). Cytological analysis of the hybridizations was done using the photomap of *D. melanogaster* (Lefevre 1976) with a Nikon Eclipse 90i-automated microscope under phase contrast.

Supplementary Material

Supplementary figures S1–S17 and tables S1–S16 and text are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Esther Betran, Brandon Gaut, Manyuan Long, John Parsch, and Therese Markow for comments; Carolus Chan, Kania Gandasetiawan, Felix Mesak, Andrei Tatarenkov, Hayden Tran, and EunBi Yang for technical help; and Rahul Warrior for providing us one stock. We are also in debt to the HPC Biocluster at UCI for grating us access to their resources. This work was supported by the National Science Foundation (MCB-1157876 to J.M.R.) and by the Ministerio de Economía y Competitividad of Spain (CGL2013-45211 to J.R.). E.S. and D.R. are grateful to the Bridges to the Baccalaureate Program supported by a National Institutes of Health grant (R25-GM056647). With exception of one author, E.S., the rest declare to have no competing interests. E.S. has stock ownership in Pacific Biosciences.

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* 287:2185–2195.

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

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.

Assis R, Bachtrog D. 2013. Neofunctionalization of young duplicate genes in *Drosophila*. *Proc Natl Acad Sci U S A.* 110:17409–17414.

Atkinson WD. 1983. Gregarious oviposition in *Drosophila melanogaster* is explained by surface texture. *Aust J Zool.* 31:925–929.

Bariami V, Jones CM, Poupardin R, Vontas J, Ranson H. 2012. Gene amplification, ABC transporters and cytochrome P450s: unraveling the molecular basis of pyrethroid resistance in the dengue vector, *Aedes aegypti*. *PLoS Negl Trop Dis.* 6:e1692.

Bartel DP. 2009. MicroRNAs: target recognition and regulatory functions. *Cell* 136:215–233.

Bauters M, Van Esch H, Friez MJ, Boespflug-Tanguy O, Zenker M, Vianna-Morgante AM, Rosenberg C, Ignatius J, Raynaud M, Hollanders K, et al. 2008. Nonrecurrent MECP2 duplications mediated by genomic architecture-driven DNA breaks and break-induced replication repair. *Genome Res.* 18:847–858.

Benevolenskaya EV, Nurminsky DI, Gvozdev VA. 1995. Structure of the *Drosophila melanogaster* annexin X gene. *DNA Cell Biol.* 14:349–357.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol.* 57:289–300.

Bergthorsson U, Andersson DI, Roth JR. 2007. Ohno's dilemma: evolution of new genes under continuous selection. *Proc Natl Acad Sci USA.* 104:17004–17009.

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

Boylan K, Serr M, Hays T. 2000. A molecular genetic analysis of the interaction between the cytoplasmic dynein intermediate chain and the glued (dynactin) complex. *Mol Biol Cell.* 11:3791–3803.
Brookfield JF. 2001. Population genetics: the signature of selection. *Curr Biol.* 11:R388–3R390.
Brown CJ, Todd KM, Rosenzweig RF. 1998. Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Mol Biol Evol.* 15:931–942.
Brown DD, Wensink PC, Jordan E. 1972. A comparison of the ribosomal DNA's of *Xenopus laevis* and *Xenopus mulleri*: the evolution of tandem genes. *J Mol Biol* 63:57–73.
Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S, Suzuki AM, et al. 2014. Diversity and dynamics of the Drosophila transcriptome. *Nature* 512:393–399.
Casola C, Ganote CL, Hahn MW. 2010. Nonallelic gene conversion in the genus Drosophila. *Genetics* 185:95–103.
Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* 13:238.
Chen JM, Cooper DN, Chuzhanova N, Ferec C, Patrinos GP. 2007. Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet.* 8:762–775.
Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. *Nat Rev Genet.* 14:645–660.
Chen S, Zhang YE, Long M. 2010. New genes in Drosophila quickly become essential. *Science* 330:1682–1685.
Diaz-Castillo C, Xia XQ, Ranz JM. 2012. Evaluation of the role of functional constraints on the integrity of an ultraconserved region in the genus Drosophila. *PLoS Genet.* 8:e1002475.
dos Santos G, Schroeder AJ, Goodman JL, Strelets VB, Crosby MA, Thurmond J, Emmert DB, Gelbart WM, FlyBase C. 2015. FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res.* 43:D690–D697.
Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323:133–138.
Fawcett JA, Innan H. 2011. Neutral and non-neutral evolution of duplicated genes with gene conversion. *Genes (Basel)* 2:191–209.
Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
FitzGerald PC, Sturgill D, Shyakhtenko A, Oliver B, Vinson C. 2006. Comparative genomics of Drosophila and human core promoters. *Genome Biol.* 7:R53.
Gibson JR, Chippindale AK, Rice WR. 2002. The X chromosome is a hot spot for sexually antagonistic fitness variation. *Proc Biol Sci.* 269:499–505.
Gnad F, Parsch J. 2006. Sebida: a database for the functional and evolutionary analysis of genes with sex-biased expression. *Bioinformatics* 22:2577–2579.
Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, et al. 2011. The developmental transcriptome of *Drosophila melanogaster*. *Nature* 471:473–479.
Hahn MW. 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. *J Hered.* 100:605–617.
Hahn MW, Han MV, Han SG. 2007. Gene family evolution across 12 Drosophila genomes. *PLoS Genet.* 3:e197.
Haygood R, Fedrigo O, Hanson B, Yokoyama KD, Wray GA. 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat Genet.* 39:1140–1144.
Hemingway J, Hawkes NJ, McCarroll L, Ranson H. 2004. The molecular basis of insecticide resistance in mosquitoes. *Insect Biochem Mol Biol.* 34:653–665.
Hoskins RA, Carlson JW, Kennedy C, Acevedo D, Evans-Holm M, Frise E, Wan KH, Park S, Mendez-Lago M, Rossi F, et al. 2007. Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin. *Science* 316:1625–1628.
Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, Hon L, Sudmant PH, Graves TA, Alkan C, Dennis MY, et al. 2014. Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res.* 24:688–696.
Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci* 256:119–124.
Innan H. 2003. A two-locus gene conversion model with selection and its application to the human RHCE and RHD genes. *Proc Natl Acad Sci USA.* 100:8793–8798.
Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet.* 11:97–108.
Innocenti P, Morrow EH. 2010. The sexually antagonistic genes of Drosophila melanogaster. *PLoS Biol.* 8:e1000335.
Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. 2008. NCBI BLAST: a better web interface. *Nucleic Acids Res.* 36:W5–W9.
Jugulam M, Niehues K, Godar AS, Koo DH, Danilova T, Friebe B, Sehgal S, Varanasi VK, Wiersma A, Westra P, et al. 2014. Tandem amplification of a chromosomal segment harboring 5-enolpyruvylshikimate-3-phosphate synthase locus confers glyphosate resistance in *Kochia scoparia*. *Plant Physiol.* 166:1200–1207.
Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* 20:1313–1326.
Katju V. 2012. In with the old, in with the new: the promiscuity of the duplication process engenders diverse pathways for novel gene creation. *Int J Evol Biol.* 2012:341932.
Katju V, Bergthorsson U. 2013. Copy-number changes in evolution: rates, fitness effects and adaptive significance. *Front Genet.* 4:273.
Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36.
Kim KE, Peluso P, Babayan P, Yeadon PJ, Yu C, Fisher WW, Chin CS, Rapicavoli NA, Rank DR, Li J, et al. 2014. Long-read, whole-genome shotgun sequence data for five model organisms. *Sci Data.* 1:140045.
Kleene KC. 2005. Sexual selection, genetic conflict, selfish genes, and the atypical patterns of gene expression in spermatogenic cells. *Dev Biol.* 277:16–26.
Kondrashov FA. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc Biol Sci.* 279:5048–5057.
Kozomara A, Griffiths-Jones S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* 42:D68–D73.
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 (Bethesda)* 5:1145–1150.
Kulathinal RJ, Sawyer SA, Bustamante CD, Nurminsky D, Ponce R, Ranz JM, Hartl DL. 2004. Selective sweep in the evolution of a new sperm-specific gene in *Drosophila*. In: Nurminsky D, editor. Selective Sweep. Austin, Texas: Kluwer Academic/Plenum Publishers. p. 1–12.
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 9:357–359.
Lê S, Josse J, Husson F. 2008. FactoMineR: an R package for multivariate analysis. *J Stat Softw.* 25:1–18.
Lefevre G. 1976. A photographic representation and interpretation of the polytene chromosomes of *Drosophila melanogaster* salivary glands. In: Ashburner MA, Novitski E, editors. The Genetics and Biology of Drosophila. London: Academic Press. p. 31–66.
Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25:1451–1452.
Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 25:402–408.
Long M, VanKuren NW, Chen S, Vibranovski MD. 2013. New gene evolution: little did we know. *Annu Rev Genet.* 47:307–333.
Ma S, Trivinos-Lagos L, Graf R, Chisholm RL. 1999. Dynein intermediate chain mediated dynein-dynactin interaction is required for interphase microtubule organization and centrosome replication and separation in Dictyostelium. *J Cell Biol.* 147:1261–1274.

Mayer MG, Rodelsperger C, Witte H, Riebesell M, Sommer RJ. 2015. The orphan gene dauerless regulates Dauer development and intraspecific competition in nematodes by copy number variation. *PLoS Genet.* 11:e1005146.

McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, Petrov DA, Fiston-Lavier AS. 2014. Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. *PLoS One* 9:e106689.

McGrath CL, Casola C, Hahn MW. 2009. Minimal effect of ectopic gene conversion among recent duplicates in four mammalian genomes. *Genetics* 182:615–622.

Michiels F, Gasch A, Kaltschmidt B, Renkawitz-Pohl R. 1989. A 14 bp promoter element directs the testis specificity of the Drosophila beta 2 tubulin gene. *EMBO J.* 8:1559–1565.

Mikhaylova LM, Nurminsky DI. 2011. Lack of global meiotic sex chromosome inactivation, and paucity of tissue-specific gene expression on the Drosophila X chromosome. *BMC Biol.* 9:29.

Milne I, Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, Shaw PD, Marshall D. 2013. Using Tablet for visual exploration of second-generation sequencing data. *Brief Bioinform.* 14:193–202.

Moore RC, Purugganan MD. 2003. The early stages of duplicate gene evolution. *Proc Natl Acad Sci USA.* 100:15682–15687.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods.* 5:621–628.

Newcomb RD, Gleeson DM, Yong CG, Russell RJ, Oakeshott JG. 2005. Multiple mutations and gene duplications conferring organophosphorus insecticide resistance have been selected at the Rop-1 locus of the sheep blowfly, *Lucilia cuprina*. *J Mol Evol.* 60:207–220.

Nurminsky DI, Nurminskaya MV, Benevolenskaya EV, Shevelyov YY, Hartl DL, Gvozdev VA. 1998a. Cytoplasmic dynein intermediate-chain isoforms with different targeting properties created by tissue-specific alternative splicing. *Mol Cell Biol.* 18:6816–6825.

Nurminsky DI, Nurminskaya MV, De Aguiar D, Hartl DL. 1998b. Selective sweep of a newly evolved sperm-specific gene in Drosophila. *Nature* 396:572–575.

O'Fallon BD 2013. ACG: rapid inference of population history from recombining nucleotide sequences. *BMC Bioinformatics* 14:40.

Obbard DJ, Maclennan J, Kim KW, Rambaut A, O'Grady PM, Jiggins FM. 2012. Estimating divergence dates and substitution rates in the Drosophila phylogeny. *Mol Biol Evol.* 29:3459–3473.

Ohno S. 1970. Evolution by Gene Duplication. New York: Springer-Verlag.

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

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet.* 39:1256–1260.

Ponce R, Hartl DL. 2006. The evolution of the novel Sdic gene cluster in *Drosophila melanogaster*. *Gene* 376:174–183.

Ranz JM, Segarra C, Ruiz A. 1997. Chromosomal homology and molecular organization of Muller's elements D and E in the *Drosophila repleta* species group. *Genetics* 145:281–295.

Rice WR. 1984. Sex chromosomes and the evolution of sexual dimorphism. *Evolution* 38:735–742.

Sassone-Corsi P. 2002. Unique chromatin remodeling and transcriptional regulation in spermatogenesis. *Science* 296:2176–2178.

Sawyer S. 1989. Statistical tests for detecting gene conversion. *Mol Biol Evol.* 6:526–538.

Schmidt EE, Schibler U. 1995. High accumulation of components of the RNA polymerase II transcription machinery in rodent spermatids. *Development* 121:2373–2383.

Singh RS, Kulathinal RJ. 2005. Male sex drive and the masculinization of the genome. *Bioessays* 27:518–525.

Slightom JL, Chang LY, Koop BF, Goodman M. 1985. Chimpanzee fetal G gamma and A gamma globin gene nucleotide sequences provide further evidence of gene conversions in hominine evolution. *Mol Biol Evol.* 2:370–389.

Smyth GK. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol.* 3:Article3.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol.* 30:2725–2729.

Tautz D, Domazet-Loso T. 2011. The evolutionary origin of orphan genes. *Nat Rev Genet.* 12:692–702.

Walsh JB. 1987. Sequence-dependent gene conversion: can duplicated genes diverge fast enough to escape conversion? *Genetics* 117:543–557.

Wang S, Adams KL. 2015. Duplicate gene divergence by changes in microRNA binding sites in Arabidopsis and Brassica. *Genome Biol Evol.* 7:646–655.

Waskar M, Li Y, Tower J. 2005. Stem cell aging in the Drosophila ovary. *Age (Dordr)* 27:201–212.

Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21:650–659.

Yeh SD, Do T, Chan C, Cordova A, Carranza F, Yamamoto EA, Abbassi M, Gandasetiawan KA, Librado P, Damia E, et al. 2012. Functional evidence that a recently evolved Drosophila sperm-specific gene boosts sperm competition. *Proc Natl Acad Sci USA.* 109:2043–2048.

Yeh SD, von Grotthuss M, Gandasetiawan KA, Jayasekera S, Xia XQ, Chan C, Jayaswal V, Ranz JM. 2014. Functional divergence of the miRNA transcriptome at the onset of Drosophila metamorphosis. *Mol Biol Evol.* 31:2557–2572.

Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science* 343:769–772.

Zimmer EA, Martin SL, Beverley SM, Kan YW, Wilson AC. 1980. Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. *Proc Natl Acad Sci USA.* 77:2158–2162.