



Rapid evolutionary dynamics of an expanding family of meiotic drive factors and their hpRNA suppressors

Jeffrey Vedanayagam^{1,3}, Ching-Jung Lin^{1,2}, Eric C. Lai^{1,2,3}

¹Developmental Biology Program, Sloan Kettering Institute, 1275 York Ave, Box 252, New York, NY 10065, USA

²Weill Graduate School of Medical Sciences, Weill Cornell Medical College, New York, NY, 10065 USA

Abstract

Meiotic drivers are a class of selfish genetic elements whose existence is frequently hidden due to concomitant suppressor systems. Accordingly, we know little of their evolutionary breadth and molecular mechanisms. Here, we trace the evolution of the *Dox* meiotic drive system in *Drosophila simulans*, which affects male-female balance (*sex-ratio*). *Dox* emerged via stepwise mobilization and acquisition of multiple *D. melanogaster* gene segments including from protamine, which mediates compaction of sperm chromatin. Moreover, we reveal novel *Dox* homologs and massive amplification of *Dox* superfamily genes on X chromosomes of its closest sisters *D. mauritiana* and *D. sechellia*. Emergence of *Dox* loci is tightly associated with 359-class satellite repeats that flank *de novo* genomic copies. In concert, we find coordinated diversification of autosomal hairpin RNA-class siRNA loci that target subsets of *Dox* superfamily genes. Overall, we reveal fierce genetic arms races between meiotic drive factors and siRNA suppressors associated with recent speciation.

Introduction

While meiotic drive is widespread in plants, animals, and fungi^{1,2}, we know little about their origins, molecular functions, and short and long-term persistence in wild populations³. A particular type of meiotic drive is sex chromosome drive, where transmission of sex chromosomes (XY or ZW) deviates from Mendelian segregation. This frequently manifests in the heterogametic sex, yielding biased progeny *sex-ratio* of affected fathers that preferentially sire females⁴. SR drive systems occur broadly across eukaryotes, but are apparently lacking in some popular model systems. For example, strong SR drivers have not

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use: <https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms>

³Correspondence: JV: jeffreypat@gmail.com, ECL: laie@mskcc.org, 212-639-5578.

Author Contributions

JV and ECL conceived and designed the study. JV performed all the computational analyses. JV and C-JL collected testis tissue samples, and C-JL generated libraries for RNA-seq and small RNA sequencing. JV analyzed and interpreted the data with inputs from ECL. JV and ECL wrote the manuscript.

Competing Interests

The authors declare no competing interests.

been identified in *D. melanogaster* (*Dmel*), but its sister species *Drosophila simulans* (*Dsim*) harbors three different SR drive systems, termed Paris⁵, Durham⁶ and Winters^{7,8}. This highlights that SR drive and suppression systems can evolve with extraordinary dynamics.

Dsim and its immediate sister species *D. sechellia* (*Dsech*) and *D. mauritiana* (*Dmau*) comprise the *simulans*-clade, which diverged from a *Dmel*-like ancestor only ~250,000 years ago⁹. These closely-related species are amenable to introgression genetics^{10,11}, yielding both SR drive and hybrid sterility factors that preferentially disrupt spermatogenesis¹². The Durham drive system was uncovered during introgressions between *Dsim* and *Dsech*, where a minimal ~80Kb autosomal region was inferred to harbor a dominant SR suppressor (*Too much yin*, *Tmy*). In turn, *Tmy* was hypothesized to silence a still-unknown driver, whose deleterious functions are suppressed and thus cryptic in contemporary *Dsim*⁶. Subsequently, the Winters SR system was defined by a distinct suppressor termed *Not much yang* (*Nmy*), whose loss depletes male progeny⁸. The target of *Nmy* hpRNA was identified as *Distorter on the X* (*Dox*). Naturally occurring deletion mutations of *Dox* bypasses the need for wild-type *Nmy*, since *dox*; *nmy* double mutants restore equal *sex-ratio* and normal spermatogenesis^{7,13}.

Nmy encodes retroposed *Dox* sequence forming an inverted repeat⁹, and *Dox* has a paralog on the X chromosome termed *Mother of Dox* (*MDox*). These *Dsim* loci are all absent from the *Dmel* genome, suggesting emergence in *Dsim* or the *simulans*-clade ancestor. However, further insights into the evolution of these meiotic drive loci were hindered by inadequate genome assemblies. Recently reported PacBio genomes from the *simulans*-clade became available¹⁴ now facilitate such efforts. For example, our small RNA analyses identified another long inverted repeat within the minimal *Tmy* interval defined by introgression genetics; this region was uniquely assembled in PacBio but not short-read genomes^{6,15}. Remarkably, the *Tmy* and *Nmy* hpRNAs are related, suggesting evolutionarily relatedness of Winters and Durham systems.

Here, we utilize PacBio assemblies to delineate evolution of *Dox*-related systems. In particular, we (1) trace *Dox* origins from its constituent genes in *Dmel*, including from protamine, (2) uncover rampant proliferation of *Dox* superfamily loci on X chromosomes of *simulans*-clade species, (3) link flanking satellite repeats to expansion of *Dox* superfamily loci, and (4) show co-evolution of *Dox* superfamily meiotic drive loci with complementary hpRNA suppressor loci. These findings testify to ongoing genetic arms races in the *simulans*-clade and the involvement of RNAi in silencing meiotic drive.

Results

The chimeric *Dox* locus includes homology to protamine

Yun Tao reported the *Dsim* meiotic drive locus *Dox* arose from an insertion into a genomic region syntenic with *Dmel*, and that *Dox* bore homology to *Mother of Dox* (*MDox*)⁷. However, as *Dox/MDox* seemed to contain only short open reading frames, their coding status and molecular origins were unknown⁷.

With our interest in *Dox* function and its suppression by hairpin RNA (hpRNA) substrates of the endogenous RNAi pathway^{15,16}, we began to reconstruct the evolutionary origins of *Dox* sequences. As defined by RNA-seq from *Dsim* testis¹⁵, *Dox* encodes a 4.1kb spliced transcript (Fig. 1A). Similarity searches at nucleotide or coding levels in *Dsim* and *Dmel* revealed complex, chimeric origins of *Dox*. In the following sections, we document homologies to *Dmel* loci (*CG###* or common gene names) and/or *Dsim* loci (*GD###*). However, to be clear, for nearly all these loci, *Dmel* exhibits the ancestral state and lacks sequence insertions present in several *Dsim* homologs.

The *Dox* transcription unit is flanked by 359 satellite repeats, with another 359 fragment within the transcribed region (Fig. 1A). 359 belongs to the complex 1.688 satellite repeat family in *Drosophila*. In *D. melanogaster*, a large block of 359 satellites resides in pericentromeric heterochromatin on the X chromosome¹⁷, but *simulans*-clade species harbor expanded 359 satellites, including a large block within euchromatic X¹⁸.

The 5' end of *Dox* bears similarity to C-terminal-encoding and 3' UTR regions of *CG8664/GD15682* (designated "1"). Embedded within this is a fragment of *DNAREP1*, which belongs to the *Helitron* family of transposable elements. Downstream of this are sections with homology to *Protamine/GD21981*. We designate homology to *Protamine/GD21981* 5' UTR as *Dox* region "2", and the *Dox* region with coding homology as "3". Protamines are involved in chromatin compaction in post-meiotic spermatids¹⁹.

A small portion (63 bp) of the putative *Dox* ORF is homologous to another *Dmel* gene (*CG15306*), which is absent from *simulans*-clade species (*Dox* segment "4"). Following the internal 359 repeat, the terminal *Dox* transcript exhibits homology to *Cubilin* on the X chromosome (termed segment "5"). Thus, the extant *Dox* locus fuses regions of 4 different ancestral protein-coding genes, in addition to various repeat sequences.

Several of these *Dox* fragments have similarity to other genomic regions (Extended Data Fig. 1). We located nine matches to segment 1, six of which are located on the X: *Dox*, *MDox*, *GD27797-a* and *GD27797-b*, and two other hits at *CG5004/GD17329* and *CG15306*. As *GD27797-a/b* share similar segmental structures with *Dox/MDox*, we name these paralogs as "*ParaDox*" genes (hereafter, *PDox1* and *PDox2*). The three autosomal matches correspond to one or both arms of the *Nmy* and *Tmy* hpRNAs on 3R (Fig. 1B). Thus, acquisition of *CG8664/GD15682* sequence was an early step during *Dox* family evolution.

Segment 2, corresponding to the non-coding portion of the autosomal *Prot/GD21981* gene, hits many of the same loci as segment 1 (Fig. 1C). We classify these *Protamine* hits distinctly, as *CG8664/GD15682* and *CG5004/GD17329* contain only noncoding matches to the *Protamine* locus (including intronic portions), while the four X-linked *Dox* family loci also match its coding region (segment 3). *Protamine* homology can also be detected at *Nmy/Tmy* hpRNAs. Notably, the four *Dox* family loci retain clear coding potential that includes the protamine-like HMG box domain that binds DNA²⁰. While not recognized earlier⁷, Conserved Domain Database (CDDv3.19) retrieved significant hits ($e < 0.001$) that include signature residues of the general HMG box domain (Fig. 1A'). *Dox* factors even

exhibit homology to human HMG box domains (Fig. 1A'), emphasizing their likely function as chromatin factors.

The C-terminal 63 bp of the predicted Dox ORF, termed segment 4, corresponds to sequence from *CG15306*. No orthologous sequence can be found, but homology of *CG15306* to extant Dox family loci suggests that insertion of an ancestral Dox family gene at this location disrupted this gene in the *simulans*-clade ancestor. The *Dmel CG15306* fragment hits *PDox1*, *PDox2*, *MDox*, and *Dox* on the X, and *Nmy* and *Tmy* hpRNAs on 3R (Fig. 1D). Finally, segment 5 bears ~1.1kb homology to the C-terminal-encoding region of *Cubilin*. *Cubilin* matches to both *MDox* and *Dox*, but not other Dox family genes, indicating this was the most recent fusion during *MDox/Dox* evolution (Fig. 1E).

Beyond the HMG box domain, we examined possible evidence for other translated regions of Dox members. As *Cubilin* homologies at *MDox/Dox* are actually located on their antisense strands, any coding potential there would seem to be fortuitous. The *CG8664*-derived segment 1 overlaps the C-terminus of the parental gene, but mostly corresponds to the *CG8664*-3' UTR. Nevertheless, we find a potential open reading frame (termed ORF13) in this region (Extended Data Fig. 2). In addition, copies of a potential open reading frame encoded by protamine-derived segment 2 (ORF5) are aligned in Extended Data Fig. 2. Although ORF5 is formally from sequence upstream of the protamine transcription unit and 5' UTR, there are more in-frame and frame-preserving indels than frame-shifting changes across these loci. While there are no clues as to the significance of these other candidate ORFs, they provide additional support to the fusion events that generated Dox family genes (Extended Data Fig. 3).

In summary, *Dox* and *MDox* are members of a larger family of newly-emerged X-linked genes in *Dsim*, which were assembled from pieces of four protein-coding genes that are extant and syntenic in *Dmel*: *CG8664/GD15682*, *Prot/GD21981*, *CG15306*, and *Cubilin*, in addition to 359 satellite repeats (Figure 1A). Moreover, the largest ORF encoded by Dox family genes are similar to the DNA binding domain of protamine, a key sperm chromatin packaging factor.

Multistep origin of *Dox* genes from dispersed genomic loci

Given the complex and hybrid structure of *Dox* transcription units, we sought a parsimonious path for their assembly. Analyses of *Dmel* and *Dsim* syntenies suggest the following model. The key initial event regards how segment 1 from *CG8664/GD15682* might have joined with segments 2 and 3 from *Prot/GD21981* (Fig. 1A). Intriguingly, all extant similarities to Dox sequence on the X contain adjoining arrangements of segments 1, 2 and 3 (i.e., 1–2-HMG). *Prot/GD21981* are on the 2L arm, while *CG8664/GD15682* are on the X chromosome, and the fusion event likely happened in the *simulans*-clade ancestor. Our observations support a model where during the divergence of *simulans*-clade from the *Dmel* ancestor, a fusion of these genes from different chromosomes led to the emergence of a chimera. With evidence that protamine gene copies are already in flux²¹ (Fig. 2A), a likely protamine copy mobilized within a *simulans*-clade ancestor and inserted within the 3' UTR of *CG8664*, located on the X chromosome (Fig. 2B). However, while the contemporary *Dsim* copy of *CG8664/GD15682* contains segments 1 and 2 in its 3' UTR, it lacks the

HMG box-bearing segment 3 (Fig. 2C). Thus, we infer that the present-day *Dsim* genome no longer contains the full copy of the original insertion that generated ancestral *Dox* with its motley gang of motifs. We refer to this inferred gene model in the *simulans*-clade ancestor as the “original-*Dox*, *ODox*” (Fig. 2B, dotted box). The sublineage of *ODox*-related copies that lack the HMG box includes at least one other *Dsim*-specific locus, *GD17329* (the ortholog of *CG5004*) (Fig. 2C, D).

Further evidence of the lability of the inferred *ODox* locus is the fact that additional copies appear to have mobilized to other genomic locations and splintered further into derivatives that are recognizable by the juxtaposition of segments 1–2-HMG. Their relationships are again clouded by the fact that certain evolutionary intermediates are lacking in present-day genomes. For example, we infer that segment 4 was acquired by insertion of *ODox* into *CG15306*; however, the current *Dsim* locus does not encode an HMG box locus (Fig. 2E). Nevertheless, we can assign this as an evolutionary link in the *Dox* superfamily lineage, because the syntenic regions of *Dsech* and *Dmau* actually contain genes bearing domains 1–2-HMG-4 (Fig. 2F). Moreover, we can now observe that a *de novo* insertion of the gene *GD27797a* bearing segments 1–2-HMG-4 now exists with the intron of *Dsim GD24701* (Fig. 2G). We note that the ancestral allele, represented by its ortholog *Dmel CG43730*, contains a 359 satellite repeat at the equivalent intronic location. As mentioned, we named this gene “*ParaDox*”, and it has duplicated and exists as two nearly-identical copies in *D. simulans* (Fig. 2H). This appears to be the first association of a *Dox* superfamily gene with satellite sequences.

ParaDox appears to be the parent of *MDox* (Fig. 2I), which in turn is the parent of *Dox* (Fig. 2J). We deduce this order, based on the fact that all these loci share the full complement of 359–1-2-HMG-4–359 segments, but only *MDox* and *Dox* share segment 5, which is related to *Cubilin*. In fact, *MDox* is inserted at *Dsim GD16058*, the ortholog of *Cubilin*, establishing it as the “mother” of *Dox*⁷ (Extended Data Fig. 4). Subsequently, it mobilized between *Ptpmeg2/GD16051* and *CG42797/GD16956* to create *Dox*, which carries a *Cubilin* segment derived from *MDox* and gained a downstream 359 satellite (Fig. 2I and J).

Overall, we establish complex mobilization and insertional gymnastics for *Dsim* *Dox* loci (Fig. 2K), a foundation to interpret broader evolutionary dynamics of *Dox* superfamily genes.

Massive expansion of *Dox* loci across *simulans*-clade species

We next analyzed copy number and synteny of *Dox* superfamily loci from the *simulans*-clade sister species *D. mauritiana* (*Dmau*) and *D. sechellia* (*Dsech*), taking advantage of recent highly contiguous assemblies of all three *simulans*-clade genomes¹⁴.

Dsim MDox is flanked by *CG15317/GD16960* and *Cubilin/GD16058*, an arrangement preserved in *Dmau*, but not *Dsech* (Fig. 3A). By contrast, while *Dsim Dox* is flanked by *CG42797/GD16956* and *Ptpmeg2/GD16051*, the equivalent genomic regions of *Dmau* and *Dsech* resemble *Dmel*, and lack an intervening *Dox* gene. Thus, *Dsim Dox* may represent a derived insertion (Fig. 3A and Extended Data Fig. 5). We also observe both conservation and flux for *PDox* genes. *Dsim PDox1* is in the intron of *CG43740/GD24701*, with similar

locations of *PDox1* in *Dmau* and *Dsech* (Fig. 3A). In contrast, *Dsim PDox2* is flanked by *Hk/GD24648* and *CG12643/GD24647*, but comparable regions of *Dmau* and *Dsech* share the ancestral state with *Dmel* (Fig. 3A). *Dsim PDox* copies have notably higher homology to Tmy, while *Dox* and *MDox* have higher homology to Nmy (Extended Data Fig. 6), suggesting preferential targeting.

Intriguingly, we identify massive amplification of Dox superfamily genes in *Dmau* and *Dsech* (Fig. 3A). We segregated these into families based on sequence similarity (Fig. 3B) and relationships to hpRNAs. In *Dmau*, there are five members of the Dox family, of which only *MDox* is syntenic. These copies have higher homology to hpRNA *Nmy*. In addition, there are six other copies, which have higher homology to an apparent Tmy-like locus (see also later analysis of hpRNA evolution in the *simulans* clade). Their distinctive sequences suggest they form a distinct subfamily, which we term the UnorthoDox (UDox) genes. Although UDox and PDox families share higher homology to Tmy than Nmy, distinct sequences cluster these duplications separately (Fig. 3C). *Dsech* harbors four and seven duplicates of the PDox and UDox families, respectively. Using newly-generated testis RNA-seq data, we detect expression of these novel *Dmau* and *Dsech* Dox paralogs (Extended Data Fig. 7).

Surprisingly, out of nine instances of syntenic Dox superfamily genes in at least two *simulans*-clade species (Fig. 3A), only three cases appear to be clear orthologs (Fig. 3B). For the six other syntenic locations, the copies appear to be members of different Dox subfamilies (Fig. 3B and Supplementary Table 2). This non-intuitive situation suggests that gene conversions or independent insertions have occurred at these syntenic loci (Supplementary Table 1). This view is supported by inspection of sequence alignments, which emphasize that many loci are more similar to dispersed copies within the same species (Extended Data Fig. 8) as opposed to syntenic copies from another species (Extended Data Fig. 9). For example, while syntenic copies of *MDox* from *Dsim* and *Dmau* cluster together, many other syntenic copies show species-level clustering (Fig. 3C).

In addition to expansion of Dox family members, we observed expansions of other gene families in the same general region. From our testis RNA-seq data, we observe expression of the tyrosine phosphatase *Ptpmeg2* in all three *simulans*-clade species, and this gene is syntenic in *Dmel*. In *Dsim*, *Dox* is inserted adjacent to *Ptpmeg2* (Fig. 3A, Extended Data Fig. 5). In addition to this syntenic copy, there are 3 full length duplicates of *Ptpmeg2* in *Dsim*, and two and three additional full-length copies in *Dmau* and *Dsech*, respectively; additional partial copies in different species exist (Fig. 3A). Another gene with associated expansions is *Mkg-r*, a recently-emerged gene on the X chromosome in the *simulans*-clade²², and a duplicate of the autosomal *Mkg-p* gene. Finally, we also note loci with partial matches to *Cubilin*, or to an intronic region of *CARPB* (Fig. 3A).

Overall, the rapid proliferation and divergence of recently-emerged copies of the Dox superfamily, are atypical for conserved genes. Instead, they conform more closely to expectations for adaptively evolving genes engaged in conflict scenarios^{1,2}. Thus, we speculate that many members of the *simulans*-clade Dox superfamily may be meiotic

drivers, and raises the question whether other amplifying genes in this region may potentially have selfish activities.

Dox loci disseminate via insertions into satellite repeats

We were curious as to how the Dox superfamily is capable of such rapid expansion, going from none in *Dmel* to large and highly variable copy numbers in each of the three *simulans*-clade species. Many Dox superfamily members, and even other amplifying non-Dox loci, are flanked by 359 satellites (Fig. 3). In fact, diverse satellite repeats have highly dynamic numbers in both heterochromatin and euchromatin of *simulans*-clade species, and recently expanded on the X chromosome in the *simulans*-clade^{14,18,23,24}.

Amongst the diverse and rapidly evolving sets of *Drosophila* satellite elements, the most abundant and oldest-known class are the 359 element/1.688 satDNA^{25,26}. Strikingly, nearly all (n=28) of the amplified copies of *Dox* superfamily members in the *simulans*-clade are flanked on one or both sides by 359 repeats (Extended Data Fig. 9). Further inspection reveals distinct modes in the transposition of Dox superfamily genes. Many cases, as exemplified by the inferred movement of *Dsim MDox* to *Dox* (Fig. 2H, I), involve localized insertion into a pre-existing 359 satellite, resulting in flanking 359 sequences on both sides of *Dox* (Fig. 2I). We identified examples of such movements that are specific to *Dmau* or to *Dsech*, or that are shared by these species. *MDox* is syntenic and only shared between *Dmau* and *Dsim*, but at the insertion location, a block of 359 satellite repeat is found in both *Dsech* and *Dmel*, indicating insertion sites previously harboring 359 satellite (Fig. 4A). Similarly, *UDox1* is shared between *Dmau* and *Dsech* and flanked by 359, but in species where *UDox1* is absent, a 359 block is seen at the syntenic location (Fig. 4A). Details of flanking 359 satellite sequences, and their sequence feature at syntenic locations in *simulans*-clade and *Dmel* are provided in Supplementary Table 3.

We also find potentially independent insertions into pre-existing 359 satellite blocks. For example, *UDox4* is found only in *Dmau*, while *UDox9* appears to be an independent insertion in *Dsech*; all these independent insertion sites also harbor pre-existing 359 satellites (Fig. 4B). Finally, we note seemingly more complex trajectories in which there may have been independent or consecutive insertions into a given genomic locus, given that the three *simulans*-clade species can contain all different gene contents between genes syntenic with *Dmel* (Fig. 4C). It is challenging to determine these evolutionary scenarios unambiguously with current data, but recurrent associations with 359 satellites strongly imply they are causal players in Dox family dynamics. Potentially, they might facilitate gene conversion²⁷, or perhaps insertions via excised circular DNA²⁸.

Recurrent emergence of hpRNA suppressors of Dox family loci

With a fuller view of dynamic proliferation of Dox genes, we turned to evolutionary strategies for their suppression. In *Dsim*, *Nmy* was proposed to originate via retroposition of *Dox* on chr3R^{7,8}, consistent with our general model that hpRNAs emerge from their prospective targets¹⁶. *Dsim Nmy* is flanked by *GD26005/CG14369* and *GD20491/CG31337* (Fig. 5A). Synteny analysis shows *Nmy* is also flanked by these genes in *Dmau*, but no such hpRNA exists in the corresponding location in *Dsech* (Fig. 5B). The absence of *Dsech Nmy*

corresponds to our observation of absence of *Dox/MDox* homologs in *Dsech*. However, the highly abundant PDox copies in *Dsech* (Fig. 3A) implied another suppressor of these loci.

We next examined *Tmy*, located ~2Mb upstream of *Nmy* on chr3R in *Dsim*. *Dsim Tmy* is flanked by *GD19044/CG5614* and *GD20331/CG5623*; however, no hpRNA exists in the syntenic region of *Dmau* and *Dsech* (Fig. 5C). This is consistent with the original introgression genetics whereby replacement of the *Dsim Tmy* region with *Dmau* material unleashes meiotic drive phenotypes⁶. Nevertheless, these genetic experiments alone do not actually mean that other species lack *Tmy*, one can only conclude that the syntenic region does not harbor a *Tmy* equivalent. Indeed, we identified a hpRNA within a different region, syntenic between *Dmau* and *Dsech*, that is homologous to *Tmy* and generates abundant siRNAs (Fig. 5D).

Is this *Dmau/Dsech* hpRNA an ortholog, or paralog, of *Dsim Tmy*? We took note of the genes flanking these hpRNAs, and observed *Dmau/Dsech* contain duplicated sequences from a pair of genes, *Gr98d* and *Klp98A*. Interestingly, these genes reside adjacent to each other in the ancestral location shared with *Dmel* (Fig. 5D). In *Dsim* as well, these genes are adjacent to each other without any evidence for an aberration that could have resulted from ancestral insertion of Dox family members. This observation refutes a single UDoX/hpRNA progenitor inserted between *Klp98A~Gr98d* in a *simulans*-clade ancestor. One plausible scenario is UDoX/hpRNA progenitor emerged in either *Dmau* or *Dsech* and traversed species boundaries via gene flow. The observation that *Dmau* and *Dsech* hairpins are not in precisely syntenic order but instead reside on the left and right sides of a centrally aligned sequence that is common to *Dsim* and *Dmel* supports this view. Alternatively, it is possible that the hpRNA emerged in the ancestor to *Dmau* and *Dsech*, and the local duplication which generated a gene arrangement with hpRNA flanked by *Klp98A* and *Gr98d* was resolved via different paths as the species diverged into contemporary *Dmau* and *Dsech*. We call these “*Tmy2*” hpRNAs. *Dsim Tmy* resides ~10Mb away from *Tmy2* in a more central location in 3R flanked by *CG4525* and *CG5623*, and our observations support a likely independent origin of *Tmy* hpRNA in *Dsim* (Fig. 5A).

We noticed that some siRNAs mapped to *Dsech Tmy* also match other autosomal locations. This reminded us of our previous discovery of *Tmy* itself, which we recognized from siRNAs that originally mapped not only to the *Nmy* hpRNA as well as Dox loci on the X, but to an uncharacterized autosomal region that proved to be the *Tmy* hpRNA¹⁵. Closer examination revealed a repeated locus bearing four tandem copies in *Dsech*. The syntenic region in *Dmel* contains the adjacent *Trp1* and *CG13131* genes. These are still recognizable in *Dsech*, but the *CG13131* copies now contain a ~130bp inverted repeat within its 3' UTR, which generates siRNAs. The *CG13131~hpRNA~Trp1* multigene unit was subsequently duplicated locally, yielding the present-day disposition in *Dsech* (Fig. 5E). As these hpRNA inverted repeats are much smaller than *Tmy*, we refer to this as the *mini-Tmy Complex* (*mTmy-C*).

Alignments of *Nmy/Tmy/mTmy* loci with Dox superfamily genes in each species reveal preferred patterns of target complementarity with individual subfamilies (Extended Data Fig. 8). For example, the newly identified *mTmy-C* loci match well to a diversifying clade

of UDox genes in *Dsech*, and are well-positioned to serve as their functional suppressors. Phylogenetic analysis of hpRNAs and Dox superfamily targets support distinct clustering of preferred hpRNA targets (Extended Data Fig. 10). Some branches have weak node support, and certain evolutionary relationships are clouded by the fact that many syntenic loci may be the products of gene conversion or independent insertions. Nevertheless, we find high or indeed perfect antisense complementarity between mature siRNAs from various hpRNAs and individual members of the Dox/PDox/UDox subfamilies (Fig. 5G), consistent with the notion that individual hpRNAs preferentially target different Dox subfamilies.

Rapid evolution of protamine genes within *D. melanogaster*

Our analyses may lead to the impression of unilateral runaway evolutionary dynamics of protamine homologs in *simulans*-clade species vs. *Dmel*. However, as canonical protamine genes duplicated in *Dmel* compared to the *simulans*-clade (Fig. 2A), and exhibit signatures of positive selection²¹, protamine loci are subject to recurrent rapid evolution.

We examined the possibility of additional alterations in protamine genes in *Dmel*. Interestingly, queries of an improved *Dmel*/PacBio Y chromosome assembly²⁹ revealed multiple copies and pseudogenes of protamine within a genomic cluster (Fig. 6A). This region is adjacent to the 18-member Mst77F cluster located on chrY³⁰, and was in fact noted as a genomic region (h17 cytoband) containing multiple copies and fragments of several gene families. At the time, these were noted as copies of *CG46192*, *ade5* (*Paics*, purine biogenesis), *CG12717* (SUMO protease) and *Crg-1* (forkhead TF)³⁰. However, subsequent work clarified that *CG46192* family and Mst77 family proteins contain the MST-HMG Box domain found in testis-restricted proteins²⁰. Of note, both Mst77F and protamines replace histones during compaction of sperm chromatin³¹. We find that *CG46192*, along with its cluster copies and pseudogenes, are more similar to protamine than Mst77F/Mst77Y proteins (Fig. 6B), indicating that they represent a distinct amplification event. Moreover, there is a complex history to emergence of this cluster, since *Paics* and *CG12717* are adjacent X chromosome loci, while protamine (*Mst35Ba/b*) genes are located on chr2L (<https://flybase.org>). The assembled *Dsim* Y does not appear to contain copies of MST-HMG Box genes.

To assess relationships of the h17 cluster with small RNAs, we examined wildtype testis data with that of the piRNA factor *aubergine* (*aub*)³². Interestingly, abundant small RNAs map to the h17 chrY cluster (Fig. 6B), but not to the adjacent Mst77Y cluster (Fig. 6A). However, these are dominantly in the piRNA-sized range (Fig. 6C). Evidence that these are in fact piRNAs comes from the fact that their accumulation is strongly decreased in *aubergine* mutant testis (Fig. 6C). The observation of abundant testis piRNAs from the h17 cluster was independently reported while this work was in revision³³. Interestingly, the small amount of remaining h17 cluster small RNAs in these mutants are preferentially 21-nt long (Fig. 6C), suggesting a possible interplay of piRNA and siRNA biogenesis at this cluster, as seen for other piRNA clusters in *D. melanogaster*^{34,35}.

Discussion

Rapid evolutionary dynamics of Dox family meiotic drive genes

In this study, we reconstruct the ancestry and diversification of an expanded family of Dox genes and their presumed hpRNA/siRNA suppressor loci. These genes exhibit partly overlapping content amongst the three *simulans*-clade species, but exhibit numerous unique genomic copies and innovations within each species. Notably, all of these Dox family loci are absent from their closest sister species *D. melanogaster* and other species in the *Dmel* group. This implies the birth of a meiotic conflict in the *simulans*-clade ancestor, and subsequent cycles of proliferation of Dox family drivers and their subsequent suppression by hpRNA loci.

Until now, the presence of any distinctive nucleotide content of *Dox* was unknown, other than its homology to *MDox* and the hpRNA loci *Nmy* and *Tmy*^{7,8,15}. However, the recognition of multiple potential open reading frames that are shared with other genomic sources, and their synteny amongst *simulans*-clade species and *D. melanogaster*, allowed us to trace stepwise origins of an ancestral *Dox* gene from multiple genomic regions that remain identifiable in *D. melanogaster*. The rapid diversification of Dox family genes, which assort into at least three recognizable subfamilies (and potentially more, depending on the granularity of subdivision), suggests that many members of this family participate in meiotic drive.

While this work was in revision, Presgraves and colleagues independently reported their study of evolutionary dynamics of Dox superfamily loci and related hpRNAs³⁶. By and large, our studies appear largely concordant, although they implement a single nomenclature for all novel Dox superfamily copies as *Dxl-1~Dxl-15*, in order of their chromosomal positions, along with the designation of *Ur-dox* as the *simulans*-clade locus at the syntenic position of *Dmel CG15604*³⁶. We emphasize the logic and utility of Dox subfamily nomenclature in our study, since (1) the subfamilies exhibit characteristic sequence features suggesting potentially distinct activities, and (2) many syntenic copies actually exhibit clearly different sequences that assign them to different subfamilies (Fig. 3). As this complex topic ultimately requires future study and integration of the two studies, we sought to provide a side-by-side comparison of the *Dxl-##* nomenclature and the *Dox/PDox/UDox* nomenclature, alongside our naming rationale (Supplementary Table 2).

Amongst of the multiple fragments of ancestral genes detected at Dox loci, their homology to the HMG box domain of protamine provides a direct framework to interpret their impact on spermatogenesis. Sperm chromatin becomes highly condensed during maturation, coinciding with replacement of histones with protamines, in flies³⁷ and mammals³⁸. Since sex chromosome conflict is most apparent in the male germline, the homology of Dox family proteins to protamine provides a testable foundation for understanding their role in meiotic drive systems that distort fidelity and quality of spermatogenesis, namely Winters and Durham drive^{6,8}. Indeed, the intimate connection of protamines and sex chromosome conflict is bolstered by the independent expansion of euchromatic and Y chromosome protamine copies in *D. melanogaster*.

Repeat-mediated evolution of SR systems

It was recently documented that, despite an overall low amount of gene flow between *D. mauritiana* and *D. simulans*, including on the X, the *Dox/MDox* interval recently transferred between these species³⁹. This represents one mechanism for the spread of meiotic drive elements between related species. However, we are struck by the highly dynamic proliferation and diversity of Dox family loci amongst the three quite closely related *simulans*-clade species, which indicates that gene flow cannot account for their evolution. Our observation of near-universal existence of 359 satellite sequences flanking most Dox superfamily genes strongly suggests that these are involved in their evolutionary strategy of dispersal. This notion is further bolstered by the existence of satellite-flanked multigene units bearing a Dox family gene, and even their existence surrounding hpRNA genes.

The 359 satellite (also known as 1.688 satDNA) is the evolutionarily oldest and also the most abundant *Drosophila* satellite sequence^{25,26}. Precise analyses of the genomic makeup of repeat sequences, including satellites, are generally difficult due to their mis-assembly in short-read sequenced genomes; yet it was recognized some time ago that 359 satellites have recently expanded on the X chromosomes of *simulans*-clade species¹⁸. The advent of single-molecule long read sequencing has enabled much greater precision in documenting the high rate of evolutionary dynamics of 359 and other satellite sequences across *simulans*-clade species^{14,23,24}. Thus, Dox superfamily loci may potentially hijack the intrinsically elevated evolutionary dynamics of satellite sequence to fuel their spread and amplification, potentially involving exchanges to and from the extrachromosomal pool.

Materials and Methods

Genome and transcriptome data

PacBio genome data for *simulans*-clade species¹⁴ was obtained from SRA through the Bioproject ID: PRJNA383250. Individual genome assemblies for *D. simulans*, *D. mauritiana*, and *D. sechellia* are available through genome assembly IDs: ASM438218v1, ASM438214v1, and ASM438219v1 respectively. We used our previously reported transcriptome datasets from *Dsim* testis¹⁵, and prepared new RNA-seq data and small RNA data from *Dmau* and *Dsech* testis, as described below.

sRNA library preparation and sequencing

For small RNA analysis, we extracted RNA from testes and accessory glands of 7-day-old *Dsim w[XD1]*, *Dmau w[1]* 14021–0241.60, and *Dsech* 14021–0248.25 strains using Trizol (Invitrogen). 1 µg of total RNA was used to prepare small RNA libraries as described⁴⁰, with the addition of QIAseq miRNA Library QC Spike-ins for normalization (Qiagen). Adenylation of 3' linker was performed in a 40 µL reaction at 65°C for 1 hr containing 200 pmol 3' linker, 1X 5' DNA adenylation reaction buffer, 100 nM ATP and 200 pmol Mth RNA ligase and the reaction is terminated by heated to 85°C for 5 min. Adenylation of 3' linker was then precipitated using ethanol and was used for 3' ligation reaction containing 10% PEG8000, 1X RNA ligase buffer, 20 µM adenylation 3' linker and 100 U T4 RNA Ligase 2 truncated K227Q. The 3' ligation reaction was performed at 4°C overnight and

the products were purified using 15% Urea-PAGE gel. The small RNA-3' linker hybrid was then subjected to 5' ligation reaction at 37°C for 4 hr containing 20% PEG8000, 1X RNA ligase buffer, 1 mM ATP, 10 µM RNA oligo, 20 U RNaseOUT and 5 U T4 RNA ligase 1. cDNA synthesis reaction was then proceeded immediately by adding following components to the ligated product: 2 µl 5x RT buffer, 0.75 µl 100 mM DTT, 1 µl 1 µM Illumina RT Primer, and 0.5 µl 10 mM dNTPs. The RT mix was incubated at 65 C for 5 min and cooled to room temperature and transfer on to ice. 0.5 µL of superscript III RT enzyme and 0.5 µL RNase OUT were added to the RT mix and the reaction was carried out at 50°C for 1 hr. cDNA libraries were amplified using 15 cycles of PCR with forward and illumine index reverse primers and the amplified libraries were purified by 8% non-denaturing acrylamide gel. Purified libraries were sequenced on HiSeq2500 using SR50 at the New York Genome Center.

RNA-seq library preparation and sequencing

We used *Dsim w[XD1]* and *Dmau w[I]* 14021–0241.60, which were used for PacBio genome sequencing¹⁴, and *Dsech* 14021–0248.25 used for the short read genome⁴¹. We isolated total RNA from ~ 5-day-old flies and for *Dsim*, *Dmau* and *Dsech* samples. We extracted RNA from testes (dissected free of accessory glands) using Trizol (Invitrogen). We made two independent dissections to generate biologically replicate RNA samples, whose quality was assessed by Bioanalyzer. We used the Illumina TruSeq Total RNA library Prep Kit LT to make RNA-seq libraries from 650 ng of total RNA. Manufacturer's protocol was followed except for using 8 cycles of PCR to amplify the final library instead of the recommended 15 cycles, to minimize artifacts caused by PCR amplification. All samples were pooled together, using the barcoded adapters provided by the manufacturer, over 2 flow cells of a HiSeq2500 and sequenced using PE75 at the New York Genome Center.

Data analysis

RNA-seq data: Paired-end RNA-seq reads was mapped to both PacBio genome assemblies for *Dsim*, *Dmau*, and *Dsech* using hisat2 aligner with the following command `hisat2 -x indexed_genome_assembly -1 $ read1.fastq.gz -2 read2.fastq.gz -S file.sam`. The alignment file in SAM format was then converted to a compressed BAM file using SAMTOOLS⁴² with the following commands: (1) `samtools view -bS file.sam > file.bam` (2) `samtools sort file.bam > file_sorted.bam` (3) `samtools index file_sorted.bam`. Mapping statistics for the BAM alignment files were obtained using `bam_stat.py` script from the RSeqQC package⁴³.

small RNA data: sRNA reads were processed as follows: Raw sequence reads were adapter trimmed using Cutadapt software (<https://cutadapt.readthedocs.io/en/stable/>). After clipping the adapter sequence, we removed the 4bp random-linker sequence inserted at 5' and 3' of the sRNA sequence (total 8bp). After filtering 15 nt reads, we mapped the small RNA data to PacBio genome assemblies using Bowtie (with `-v0 -best -strata` options.) The resulting small RNA alignments in SAM format were converted to BED for downstream processing using the BEDops software and visualized on IGV.

Homology and domain searches

Sequence homology search for putative ORFs encoded in the Dox transcript and search for Dox like sequences in the PacBio assemblies were performed using command-line version of blastn and/or tblastn implemented in BLAST 2.2.31+⁴⁴. Search for conserved protein domains in the Dox family genes was performed using both HMMER v3.3.2 and NCBI conserved domain database (CDD v3.19).

Phylogenetic analysis

For phylogenetic analysis of Dox superfamily genes, we constructed an alignment of CDS for each ortholog using the translation align feature in Geneious (version 11.0.4). For this alignment, we excluded two UDox copies (*UDox1* and *UDox5*) in *Dmau*, which appear to carry premature stop codons. The alignment was performed using the Geneious multiple alignment sequence feature using the global alignment with free end gaps. For the alignment, a 65% similarity (5.0/−4.0) cost matrix was used with the following gap penalty parameters: (1) Gap open penalty: 12, (2) Gap extension penalty: 3, and (3) Refinement iterations: 2. The resulted alignment was then manually curated to ensure proper alignment. Phylogenetic analysis on the nucleotide alignment was performed using MrBAYES⁴⁵ plugin in Geneious software v11.0.4. For this analysis, we used the HKY85 substitution model with *Dmel*/ProtA/B as an outgroup. A gamma rate variation option was used with the following gamma categories: 4. For Monte Carlo Monte Chain (MCMC) settings, we used the following parameters: (1) chain length: ranging from 100000 to 150000 based on the trace file, (2) Heated chains: 4, (3) Heated chain Temp: 0.2, (5) Subsampling Frequency: 100, (6) Burn-in length: 1000, (7) Random seed: 7826. For priors, we used the Unconstrained Branch Lengths option with the following GammaDir parameters (1,0.1,1) and the Shape Parameter: Exponential (10).

Data availability

Paired-end RNA-seq reads from *Dmau*, *Dsim*, *Dsech* testis, and small RNA data from *Dmau* and *Dsech* are available from the GEO database: GSE185361.

Code availability

Codes for analyses in this manuscript are available at https://github.com/Lai-Lab-Sloan-Kettering/Dox_evolution.

Extended Data

A Identity of Segment 1 (from CG8664) to other Dox family loci

	CG15306	nmy_left_arm	GD17329	MDox	Dox	CG8664_gen...	PDox2	PDox1	tmy_left_arm	tmy_right_arm
CG15306		57.729%	51.684%	48.910%	48.910%	50.083%	96.250%	88.604%	48.558%	48.558%
nmy_left_arm	57.729%		27.341%	33.273%	33.333%	9.509%	29.011%	41.781%	31.972%	31.972%
GD17329	51.684%	27.341%		47.583%	47.583%	94.486%	55.891%	58.507%	48.338%	48.338%
MDox	48.910%	33.273%	47.583%		98.743%	60.714%	62.098%	45.865%	95.567%	95.567%
Dox	48.910%	33.333%	47.583%	98.743%		60.379%	61.958%	45.591%	95.382%	95.382%
CG8664_gene_region	50.083%	9.509%	94.486%	60.714%	60.379%		65.193%	59.740%	60.938%	60.938%
PDox2	96.250%	29.011%	55.891%	62.098%	61.958%	65.193%		89.569%	60.888%	60.888%
PDox1	88.604%	41.781%	58.507%	45.865%	45.591%	59.740%	89.569%		44.341%	44.341%
tmy_left_arm	48.558%	31.972%	48.338%	95.567%	95.382%	60.938%	60.888%	44.341%		100%
tmy_right_arm	48.558%	31.972%	48.338%	95.567%	95.382%	60.938%	60.888%	44.341%	100%	

B Identity of Segment 2 (from Prot/GD21981) to other Dox family loci (global alignment)

	Nmy_right_arm GD21981_ge...	CG8664_ove...	GD17329	CG15306	Dox	MDox	tmy_left_arm	nmy_left_arm	tmy_right_arm
Nmy_right_arm		83.636%	50.000%	0%	53.704%	52.727%	53.704%	51.852%	53.704%
GD21981_gene_region	83.636%		17.506%	18.601%	5.627%	8.384%	3.538%	18.909%	5.446%
CG8664_overlap	50.000%	17.506%		35.882%	23.793%	25.432%	14.481%	39.024%	22.719%
GD17329	0%	18.601%	35.882%		0.266%	3.865%	0%	1.550%	0.267%
CG15306	53.704%	5.627%	23.793%	0.266%		56.923%	96.939%	80.233%	92.814%
MDox	52.727%	8.384%	25.432%	3.865%	56.923%		43.750%	74.866%	55.556%
Dox	53.704%	3.538%	14.481%	0%	96.939%	43.750%		72.549%	98.990%
tmy_left_arm	51.852%	18.909%	39.024%	1.550%	80.233%	74.866%	72.549%		81.065%
nmy_left_arm	53.704%	5.446%	22.719%	0.267%	92.814%	55.556%	98.990%	77.326%	
tmy_right_arm	43.996%	5.739%	23.793%	1.762%	73.016%	52.434%	62.500%	81.065%	70.370%

B' Identity of Segment 2 (from Prot/GD21981) to other Dox family loci (local alignment)

	GD21981/Prot CG8664_ove...	GD17329	CG15306	PDox1	PDox2	tmy_right_arm	tmy_left_arm	Dox	nmy_left_arm
GD21981/Prot		94.074%	87.770%	92.593%	92.593%	91.852%	93.333%	93.333%	94.074%
CG8664_overlap	94.074%		93.525%	95.556%	95.556%	94.815%	96.296%	96.296%	97.037%
GD17329	87.770%	93.525%		90.647%	90.647%	89.928%	91.367%	91.367%	92.086%
CG15306	92.593%	95.556%	90.647%		100%	99.259%	96.296%	96.296%	97.037%
PDox1	92.593%	95.556%	90.647%	100%		99.259%	96.296%	96.296%	97.037%
PDox2	91.852%	94.815%	89.928%	99.259%	99.259%		95.556%	95.556%	96.296%
tmy_right_arm	93.333%	96.296%	91.367%	96.296%	96.296%	95.556%		100%	97.778%
tmy_left_arm	93.333%	96.296%	91.367%	96.296%	96.296%	95.556%	100%		97.778%
Dox	94.074%	97.037%	92.086%	97.037%	97.037%	96.296%	97.778%	97.778%	
nmy_left_arm	88.148%	91.111%	87.050%	91.111%	91.111%	90.370%	92.593%	92.593%	94.074%

C Identity of Segment 4 (from CG15306) to other Dox family loci

	MDox	Dox (reversed)	CG15306
MDox		92.308%	69.231%
Dox (reversed)	92.308%		71.154%
CG15306	69.231%	71.154%	

	CG15306	Tmy_left_arm
CG15306		84.000%
Tmy_left_arm	84.000%	

	CG15306	PDox2_last...
CG15306		85.714%
PDox2_last...	85.714%	

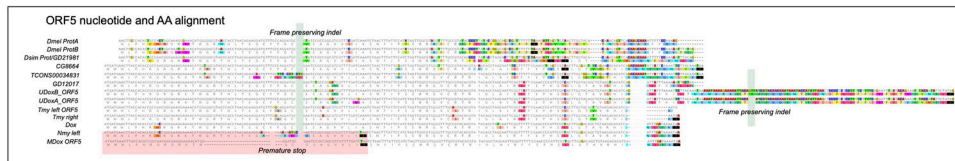
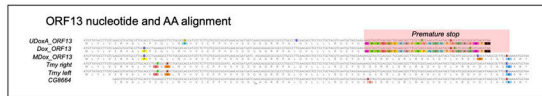
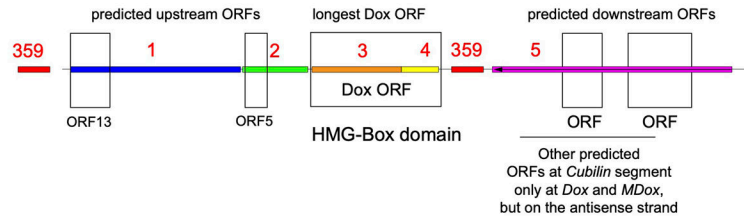
	CG15306	nmy_left_arm
CG15306		84.000%
nmy_left_arm	84.000%	

D Identity of Segment 5 (from Cubilin) to other Dox family loci

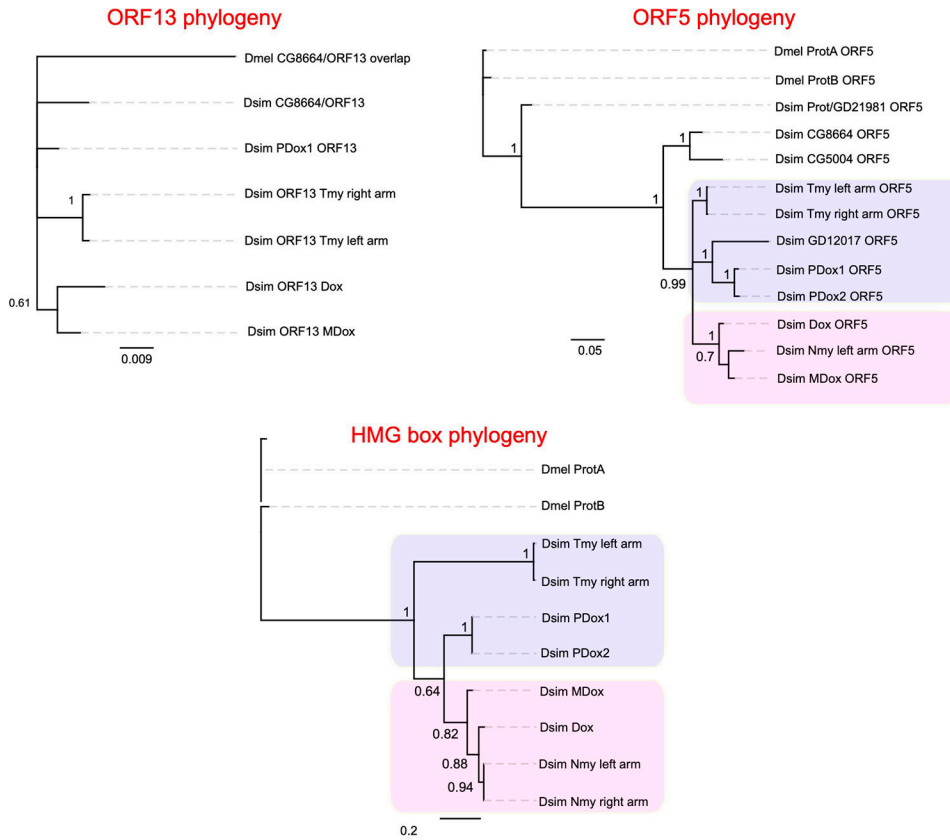
	cubn_gene...	MDox_3_pr...	Cubn_CDS	cubn_gene...	Dox_3_prim...
cubn_gene_region		93.832%		92.519%	88.765%
MDox_3_prime (reversed)	93.832%			88.765%	95.149%
cubn_gene_region					
Dox_3_prime (reversed)					

Extended Data Fig. 1 | BLASTn identity matrices for components of the Dox meiotic driver network.

(A) percent identity matrix for BLAST alignment for *CG8664* (segment 1) shown in main Figure 1B. An example key to interpret the table is highlighted in yellow, for example the identity of *MDox* to *CG8664* is 60.7%. (B) percent identity matrix for BLAST alignment for *Prot/GD21981* (segment 2) shown in Figure 1C. The % identity shown in B is from a global alignment comparing 2.9Kb *Prot/GD21981* gene to segments derived from protamine. (B') percent identity matrix for a local alignment where there is high homology between *Prot/GD21981* and segment 2 shown in Figure 1C. (C) percent identity matrix for BLAST alignment for *CG15306* (segment 4) shown in Figure 1D. (D) percent identify matrix for BLAST alignment shown for *Cubilin* (segment 5) in Figure 1E.

Domain structure of *Dox* and *MDox***Extended Data Fig. 2 |. Nucleotide and amino acid alignments for additional putative Dox ORFs (ORF13 and ORF5).**

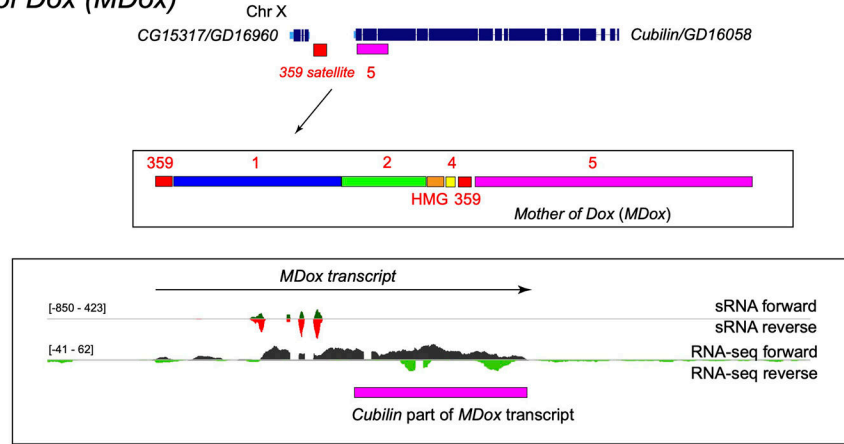
In two instances, ORF5 has accumulated premature stop codon at *Nmy* hpRNA left arm, and *MDox*. There are also two instances of frame-preserving indels at ORF5. Similarly, there are 2 instances of frame-shifting mutations, which resulted in a premature stop at *UDoxA* and *Dox* at ORF13. For ORF5 in *Dsim*, using a comparable ORF upstream of ProtA/B in *Dmel* as an outgroup, we found an excess of both non-synonymous fixed (7) and polymorphic sites (26) compared to synonymous fixed (1) and polymorphic (8) changes from MK test (MK test, P-value = 0.66). Similarly, for ORF13, we found an excess of both non-synonymous fixed (8) and polymorphic (3) sites compared to synonymous fixed (1) and polymorphic (2) changes (MK test, P-value = 0.51). While MK test does not support directional selection, frame preserving indels indicate likely functional roles for these ORFs.



Extended Data Fig. 3 |. Phylogeny of putative ORFs, ORF5 and ORF13, and the HMG-box containing ORF found at Dox.

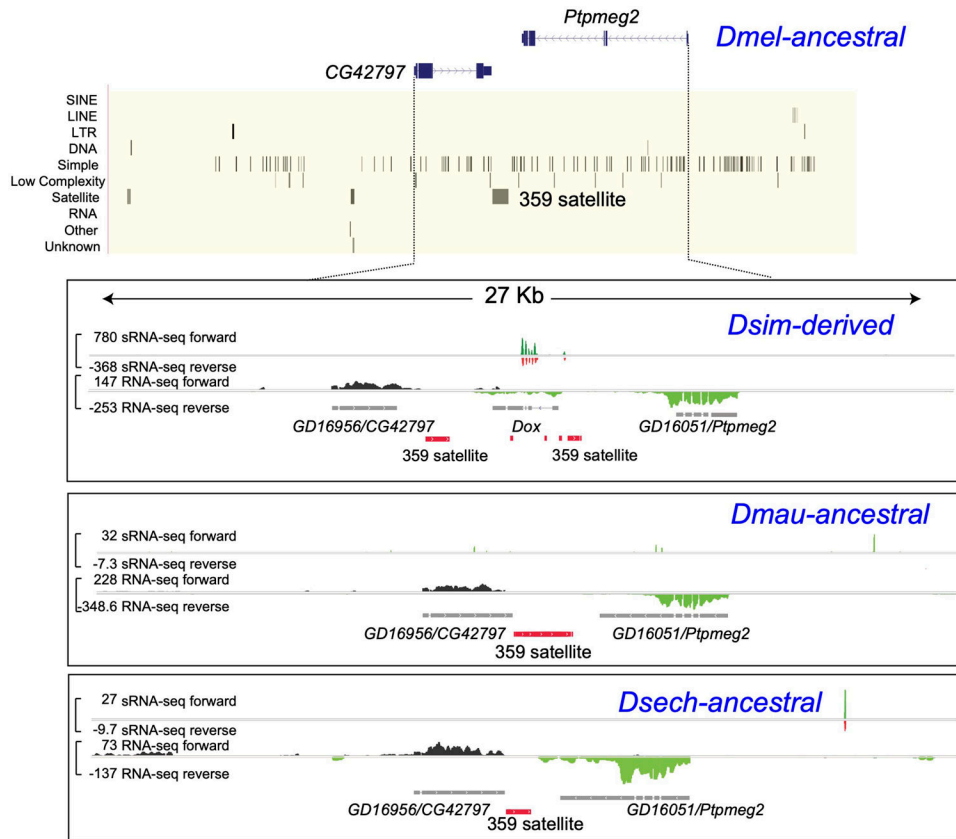
When *Dmel* is used as an outgroup, the alignable portion of ORF13 is 130nt, and the resulting tree is unresolved. However, the phylogeny of sequences bearing ORF5 (derived from non-coding portion of protamine) resolve the relationships between extant Dox family members and their hpRNA suppressors, such that Tmy is related to the PDox family while Nmy is related to the Dox/MDox genes. A similar relationship is recapitulated in the phylogeny constructed from HMG-box ORF (derived from protein-coding portion of protamine). Numbers in the tree branches show posterior probability from MrBayes analysis (outgroup is *Dmel* ORF13/ORF5/HMG-ORF, substitution model: GTR, rate variation: equal, chain-length: 10000, Heated chains: 4, burn-in length: 1000, random-seed: 13804, and we used the following unconstrained branch lengths as priors: Gamma (1,01,1,1). The dotted grey line is to align the branch labels.

Mother of Dox (MDox)



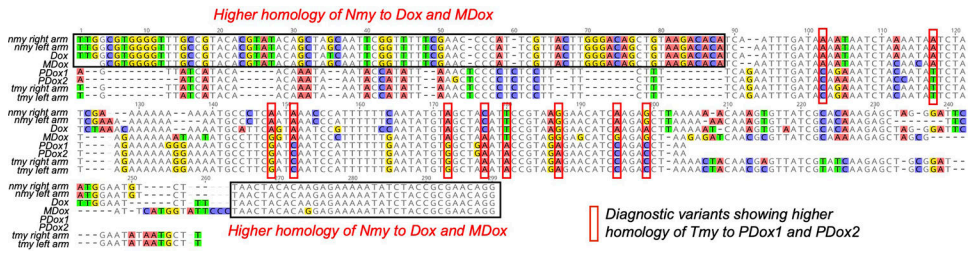
Extended Data Fig. 4 | Evolution and expression of MDox.

(A) Cubilin was acquired as part of the *MDox* transcript following insertion at a 359 satellite located between *CG15371/GD16960* and *Cubilin*. RNA-seq shows *MDox* transcript includes a portion of *Cubilin*, which forms the part of segment 5 seen at both *MDox* and *Dox*. The small RNAs (sRNAs) that map to *MDox* do not correspond to the *Cubilin* section.



Extended Data Fig. 5 | Example synteny analysis with RNA-seq, small RNA, and repeat annotation tracks.

For each Dox superfamily locus, we examined the *Dmel* ancestral State and their synteny in the simclade. In this example, the Dox locus is shown with reference to its flanking genes *CG42797* and *Ptpmeg2*. The *Dmel* ancestral State shows the flanking genes, and the corresponding repeat tracks in *Dmel*, which shows a block of 359 satellite repeat. In *Dsim*, there is an insertion of *Dox* in this site flanked by *GD16956/CG42797* and *GD16051/Ptpmeg2*. RNA-seq forward and reverse are shown in black and light green respectively at both flanking genes and the *Dox locus*. Small RNA mapping in the forward and reverse strands are shown in dark green and red respectively at *Dox*. In both *Dmau* and *Dsech*, the sequence in between *GD16956/CG42797* and *GD16051/Ptpmeg2* contains 359 satellite repeat similar to *Dmel*, but the *Dox* locus is absent. Y-axis values represent normalized RNA-seq and small RNA counts in their respective tracks.



Extended Data Fig. 6 |. Homology relationships between hpRNAs Nmy and Tmy and their Dox superfamily targets.

Alignment of hpRNA arms to targets reveals *Nmy* is highly homologous to *Dox* and *MDox*, while *Tmy* is highly homologous to *PDox1* and *PDox2*. Boxed segment in black shows stretches of nucleotide homology between *Dox/MDox* to *Nmy*. Boxed nucleotides in red shows diagnostic variants, which show higher homology of *Tmy* to *PDox1* and *PDox2*. These relationships suggest origins of hpRNA suppressors. *Tmy* likely emerged from a *PDox* copy, while *Nmy* emerged from *Dox/MDox*.

Author Manuscript

Author Manuscript

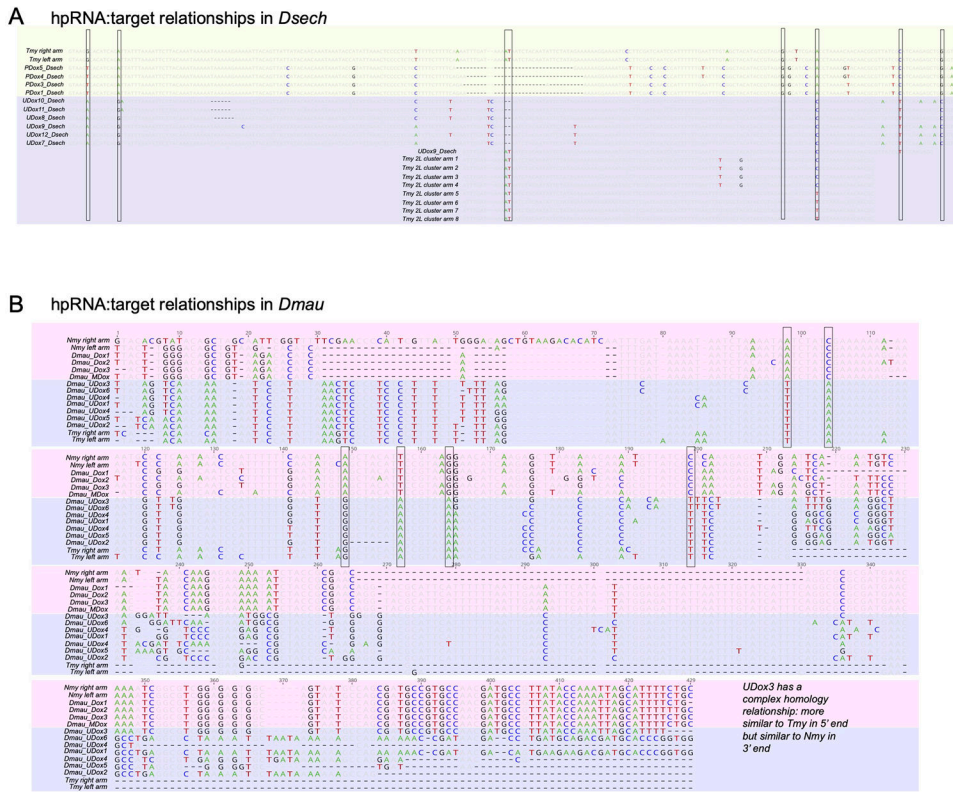
Author Manuscript

Author Manuscript

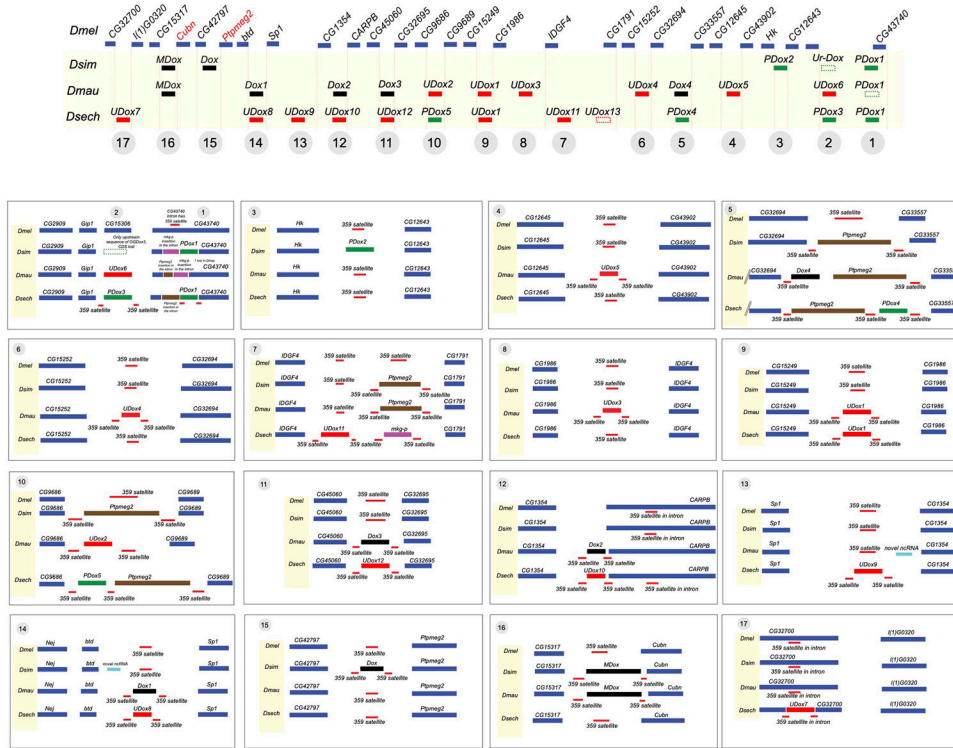


Extended Data Fig. 7. RNA-seq evidence for expression of Dox superfamily amplified genes in simulans clade species.

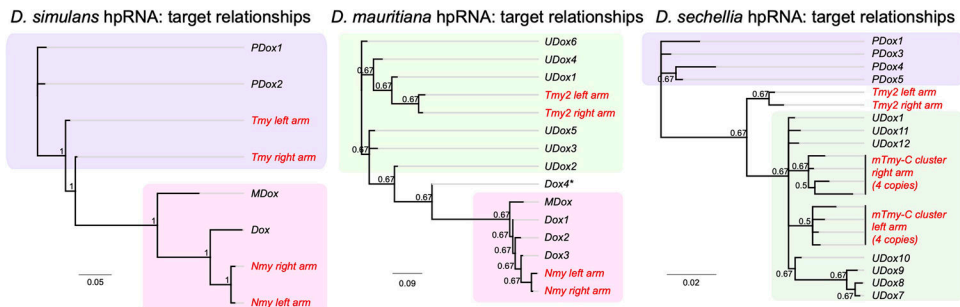
IGV screenshots show small RNA and RNA-seq mapping to Dox superfamily loci. The flanking 359 satellite sequences are shown in the annotation track in blue.



Extended Data Fig. 8 |. Nucleotide alignments showing relationships between hpRNAs and their targets. hpRNA/target complementarity in *Dsech* (A) and *Dmau* (B). Highlighted background shows subset of Dox superfamily loci related to a specific hpRNA. In *Dsech*, a subset of UDox family loci are highly similar to *Tmy* at 3R, while another subset of UDox family loci are similar to the novel *mini-Tmy* hpRNA cluster in 2L. Similarity defining polymorphisms are shown in boxes. Similarly, in *Dmau*, a subset of Dox family loci are similar to *Nmy*, while a subset of UDox family loci are homologous to *Tmy*. Interestingly, one locus *UDox3* shows higher homology to *Tmy* in its 5' end but greater homology to *Nmy* in its 3' end, indicating complex relationships and evolution of these hpRNAs and their suppressors.



Extended Data Fig. 9 |. Synteny relationships of all Dox superfamily loci.
 (Top) The genomic interval on chromosome X that contains Dox superfamily loci in the three simulans clade species (*Dsim*, *Dmau*, *Dsech*) is syntenic and preserves gene order in *Dmel*, which lacks Dox genes. The presence of Dox superfamily loci (subdivided into Dox/MDox, PDox and UDox families), is designated below. (Bottom) The syntenic genomic locations of each of the Dox superfamily insertions in the four mel-complex species is shown. Note that, as a rule, Dox superfamily loci are flanked by 359 satellite repeats. Moreover, all but one of these Dox superfamily insertion regions (#12) contains a 359 satellite block in the ancestral State (*Dmel*). Beyond Dox superfamily genes, there is also co-insertion of certain other gene families, notably *Ptpmeg2* and *mkg-p*.



Extended Data Fig. 10 |. Phylogenetic relationships between hpRNAs and their targets in *Dsim*, *Dmau*, and *Dsech*.
 Tmy and Nmy hpRNAs in *Dsim* target distinct Dox superfamily members PDox and Dox/MDox, respectively. In *Dmau* and *Dsech*, a non-syntenic Tmy2 targets a distinct UDox family. However, in *Dmau* and *Dsech*, the distinction of hpRNA: target relationships at the

phylogenetic level are not well resolved due to lack of appropriate outgroups, and potential gene conversion events confounding 1:1 relationships. The asterisk in *Dmau* hpRNA: target phylogeny is one such example of a likely gene conversion event which clouds separation of Tmy targets and Nmy targets into distinct clades. All phylogenetic trees were constructed using MrBayes plugin in Geneious software. We assessed MCMC convergence of trees with chain-lengths ranging from 100000-500000, and used a chain-length from the trace file, which represented converged tree. Values in the nodes represent posterior probability. We used the following MCMC settings [Chain length: 500000 (Dsim); 150000 (Dmau and Dsech); Subsampling frequency: 100000; Burn-in length: 1000] for the trees shown above.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Jiayu Wen for initial investigations into the domain structure of Dox, and Peter Smibert for the first recognition that Dox/MDox bear homology to protamine. We thank Anthony Geneva for his inputs on phylogenetic analysis. We also thank Daven Presgraves for communications on this topic prior to publication, and the referees for critical comments that improved this work. This work was supported by a Pathway to Independence award from the National Institute of General Medical Sciences (K99-GM137077, JV), US-Israel Binational Science Foundation (BSF-2015398, ECL), National Institute of General Medical Sciences (R01-GM083300, ECL) and National Institutes of Health MSK Core Grant (P30-CA008748).

References

1. Zanders SE & Unckless RL Fertility Costs of Meiotic Drivers. *Curr Biol* 29, R512–R520 (2019). [PubMed: 31163165]
2. Agren JA & Clark AG Selfish genetic elements. *PLoS Genet* 14, e1007700 (2018). [PubMed: 30439939]
3. Lindholm AK et al. The Ecology and Evolutionary Dynamics of Meiotic Drive. *Trends Ecol Evol* 31, 315–326 (2016). [PubMed: 26920473]
4. Jaenike J Sex chromosome meiotic drive. *Annu Rev Ecol Syst* 32, 25–49 (2001).
5. Hellegren Q et al. Rapid evolution of a Y-chromosome heterochromatin protein underlies sex chromosome meiotic drive. *Proc Natl Acad Sci U S A* 113, 4110–5 (2016). [PubMed: 26979956]
6. Tao Y, Hartl DL & Laurie CC Sex-ratio segregation distortion associated with reproductive isolation in *Drosophila*. *Proc Natl Acad Sci U S A* 98, 13183–8 (2001). [PubMed: 11687638]
7. Tao Y et al. A sex-ratio meiotic drive system in *Drosophila simulans*. II: an X-linked distorter. *PLoS Biol* 5, e293 (2007). [PubMed: 17988173]
8. Tao Y, Masly JP, Ararape L, Ke Y & Hartl DL A sex-ratio meiotic drive system in *Drosophila simulans*. I: an autosomal suppressor. *PLoS Biol* 5, e292 (2007). [PubMed: 17988172]
9. Garrigan D et al. Genome sequencing reveals complex speciation in the *Drosophila simulans* clade. *Genome Res* 22, 1499–511 (2012). [PubMed: 22534282]
10. Masly JP & Presgraves DC High-resolution genome-wide dissection of the two rules of speciation in *Drosophila*. *PLoS Biol* 5, e243 (2007). [PubMed: 17850182]
11. Presgraves DC, Gerard PR, Cherukuri A & Lyttle TW Large-scale selective sweep among Segregation Distorter chromosomes in African populations of *Drosophila melanogaster*. *PLoS Genet* 5, e1000463 (2009). [PubMed: 19412335]
12. Meiklejohn CD & Tao Y Genetic conflict and sex chromosome evolution. *Trends Ecol Evol* 25, 215–23 (2010). [PubMed: 19931208]
13. Kingan SB, Garrigan D & Hartl DL Recurrent selection on the Winters sex-ratio genes in *Drosophila simulans*. *Genetics* 184, 253–65 (2010). [PubMed: 19897749]

14. Chakraborty M et al. Evolution of genome structure in the *Drosophila simulans* species complex. *Genome Res* 31, 380–396 (2021). [PubMed: 33563718]
15. Lin C-J et al. The hpRNA/RNAi Pathway Is Essential to Resolve Intragenomic Conflict in the *Drosophila* Male Germline. *Dev Cell* 46, 316–326 e5 (2018). [PubMed: 30086302]
16. Wen J et al. Adaptive regulation of testis gene expression and control of male fertility by the *Drosophila* hairpin RNA pathway. [Corrected]. *Mol Cell* 57, 165–78 (2015). [PubMed: 25544562]
17. Usakin L et al. Transcription of the 1.688 satellite DNA family is under the control of RNA interference machinery in *Drosophila melanogaster* ovaries. *Genetics* 176, 1343–9 (2007). [PubMed: 17409066]
18. Garrigan D, Kingan SB, Geneva AJ, Vedanayagam JP & Presgraves DC Genome diversity and divergence in *Drosophila mauritiana*: multiple signatures of faster X evolution. *Genome Biol Evol* 6, 2444–58 (2014). [PubMed: 25193308]
19. Rathke C et al. Transition from a nucleosome-based to a protamine-based chromatin configuration during spermiogenesis in *Drosophila*. *J Cell Sci* 120, 1689–700 (2007). [PubMed: 17452629]
20. Doyen CM et al. A Testis-Specific Chaperone and the Chromatin Remodeler ISWI Mediate Repackaging of the Paternal Genome. *Cell Rep* 13, 1310–1318 (2015). [PubMed: 26549447]
21. Dorus S, Freeman ZN, Parker ER, Heath BD & Karr TL Recent origins of sperm genes in *Drosophila*. *Mol Biol Evol* 25, 2157–66 (2008). [PubMed: 18653731]
22. Wang W, Yu H & Long M Duplication-degeneration as a mechanism of gene fission and the origin of new genes in *Drosophila* species. *Nat Genet* 36, 523–7 (2004). [PubMed: 15064762]
23. Khost DE, Eickbush DG & Larracuent AM Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. *Genome Res* 27, 709–721 (2017). [PubMed: 28373483]
24. Sproul JS et al. Dynamic Evolution of Euchromatic Satellites on the X Chromosome in *Drosophila melanogaster* and the *simulans* Clade. *Mol Biol Evol* 37, 2241–2256 (2020). [PubMed: 32191304]
25. Travaglini EC, Petrovic J & Schultz J Satellite DNAs in the embryos of various species of the genus *Drosophila*. *Genetics* 72, 431–9 (1972). [PubMed: 4630029]
26. de Lima LG, Hanlon SL & Gerton JL Origins and Evolutionary Patterns of the 1.688 Satellite DNA Family in *Drosophila* Phylogeny. *G3 (Bethesda)* 10, 4129–4146 (2020). [PubMed: 32934018]
27. Haudry A, Laurent S & Kapun M Population Genomics on the Fly: Recent Advances in *Drosophila*. *Methods Mol Biol* 2090, 357–396 (2020). [PubMed: 31975175]
28. Thomas J, Phillips CD, Baker RJ & Pritham EJ Rolling-circle transposons catalyze genomic innovation in a mammalian lineage. *Genome Biol Evol* 6, 2595–610 (2014). [PubMed: 25223768]
29. Chang CH et al. Islands of retroelements are major components of *Drosophila* centromeres. *PLoS Biol* 17, e3000241 (2019). [PubMed: 31086362]
30. Krsticevic FJ, Schrago CG & Carvalho AB Long-Read Single Molecule Sequencing to Resolve Tandem Gene Copies: The Mst77Y Region on the *Drosophila melanogaster* Y Chromosome. *G3 (Bethesda)* 5, 1145–50 (2015). [PubMed: 25858959]
31. Jayaramaiah Raja S & Renkawitz-Pohl R Replacement by *Drosophila melanogaster* protamines and Mst77F of histones during chromatin condensation in late spermatids and role of sesame in the removal of these proteins from the male pronucleus. *Mol Cell Biol* 25, 6165–77 (2005). [PubMed: 15988027]
32. Nagao A et al. Biogenesis pathways of piRNAs loaded onto AGO3 in the *Drosophila* testis. *RNA* 16, 2503–15 (2010). [PubMed: 20980675]
33. Chen P et al. piRNA-mediated gene regulation and adaptation to sex-specific transposon expression in *D. melanogaster* male germline. *Genes Dev* 35, 914–935 (2021). [PubMed: 33985970]
34. Malone CD et al. Specialized piRNA pathways act in germline and somatic tissues of the *Drosophila* ovary. *Cell* 137, 522–35 (2009). [PubMed: 19395010]
35. Lau N et al. Abundant primary piRNAs, endo-siRNAs and microRNAs in a *Drosophila* ovary cell line. *Genome Res* 19, 1776–85 (2009). [PubMed: 19541914]

36. Muirhead CA & Presgraves DC Satellite DNA-mediated diversification of a sex-ratio meiotic drive gene family in *Drosophila*. *Nature Ecology & Evolution*, in press (2021).
37. Rathke C, Baarends WM, Awe S & Renkawitz-Pohl R Chromatin dynamics during spermiogenesis. *Biochim Biophys Acta* 1839, 155–68 (2014). [PubMed: 24091090]
38. Wang T, Gao H, Li W & Liu C Essential Role of Histone Replacement and Modifications in Male Fertility. *Front Genet* 10, 962 (2019). [PubMed: 31649732]
39. Meiklejohn CD et al. Gene flow mediates the role of sex chromosome meiotic drive during complex speciation. *Elife* 7(2018).
40. Lee JE & Yi R Highly efficient ligation of small RNA molecules for microRNA quantitation by high-throughput sequencing. *J Vis Exp*, e52095 (2014).
41. Clark AG et al. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450, 203–18 (2007). [PubMed: 17994087]
42. Li H et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–9 (2009). [PubMed: 19505943]
43. Wang L, Wang S & Li W RSeQC: quality control of RNA-seq experiments. *Bioinformatics* 28, 2184–5 (2012). [PubMed: 22743226]
44. Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ Basic local alignment search tool. *J Mol Biol* 215, 403–10 (1990). [PubMed: 2231712]
45. Huelsenbeck J & Ronquist F MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755 (2001). [PubMed: 11524383]

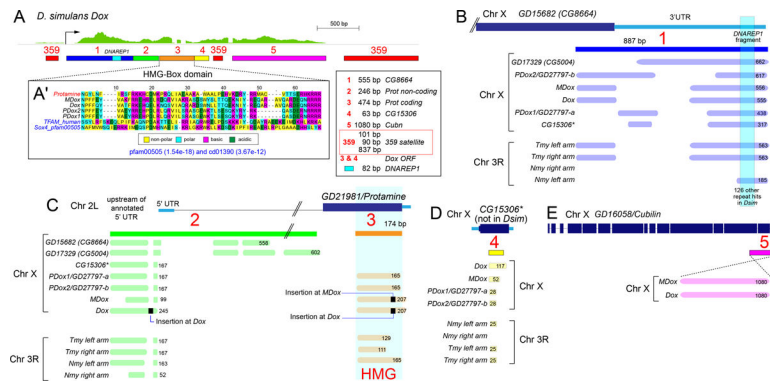


Figure 1. Structure of *Dox* transcript with segments acquired from various genes on the path to its origin. **(A)** Testis RNA-seq data shows a multi-exonic transcript from the *Dox* region, with several distinct segments acquired from protein-coding genes and repetitive elements. 359 corresponds to sequence with similarity to the 359 (also known as 1.688 family) satellite repeat. Segment 1 (blue) corresponds to sequence acquired from *GD15682* (*CG8664*); embedded within this segment is 82 bp derived from *DNAREP1* transposable element (turquoise). Segments 2 (green) and 3 (orange) correspond to sequences acquired from *GD21981* (*Protamine*). Segment 3 is from the protein-coding portion of *Protamine*, which harbors an HMG-box domain. Inset (**A1'**) highlight amino acid identify/similarities between *Dox* family genes and both the *Protamine* HMG-box domain, and more distant HMG-box sequences from human (pfam definition) and *Sox4* (pfam00505) as an outgroup. Segment 4 (yellow) was acquired from *CG15306*, and segment 5 (pink) derives from *Cubilin*. The key depicts *Dox* segment features, including their segment number, nucleotide length, and origin. **(B)** Overlap of various genomic regions to *GD15682* (*CG8664*) from BLAST search. Segment 1 (blue) corresponds to 887bp from C-terminus and 3' UTR of *GD15682* (*CG8664*). BLAST hits of various lengths to different genomic features on ChrX and Chr3R are shown as light blue bars with lengths of nucleotide homology indicated. 82 bp of Segment 1, which corresponds to *DNAREP1* transposable element, retrieves 126 BLAST hits in the *Dsim* PacBio genome **(C)** Genomic matches to the ancestral *protamine* (*GD21981*) gene, include regions with similarity to its upstream/5' UTR/intronic regions (green), and others bearing the HMG box domain (orange). **(D)** Segment 4 from *Dox* was acquired from *CG15306*. *CG15306* is no longer extant in *Dsim*, but relics from the insertion can be identified from BLAST search at *Dox* superfamily genes and their hpRNA suppressors. **(E)** Segment 5 from *Dox* was acquired from C terminus of *Cubilin* (pink). This segment is found only at *Dox* and *MDox*. Note that *Cubilin* matches to the antisense strands of *Dox* and *MDox*.

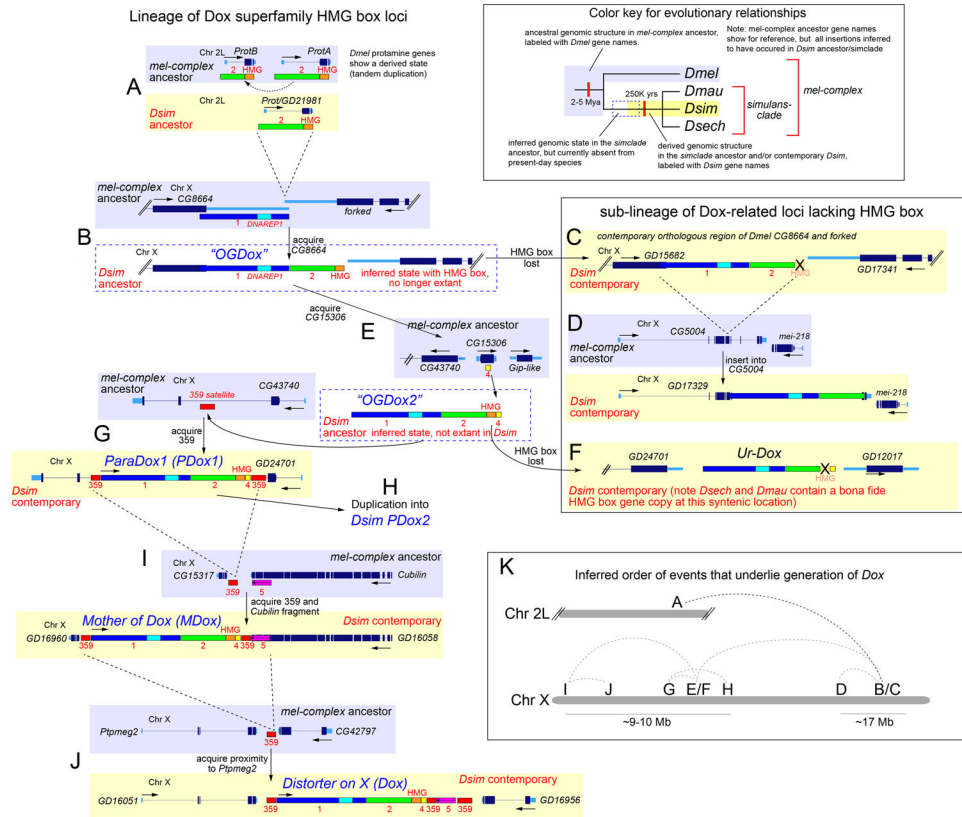


Figure 2. Stepwise origins of *Dox* from a *Protamine*-like ancestor. Upper right, key for the labeling of gene names and structures. Note that many regions correspond to extant genomic loci in *Dmel* (purple) and *Dsim* (yellow), but mobilizations occurred in a *simulans*-clade ancestor; they are not meant to indicate mobilizations occur between contemporary *Dsim* and *Dmel*. In some cases, the inferred events are no longer present in contemporary species (dotted blue boxes). (A) *Dmel* Protamine is tandemly duplicated (*ProtA/B*), while *Dsim* has a single copy (*GD21981*); *Dmel* is a derived state. Protamine segments acquired by *Dox* genes are green (segment 2) and orange (HMG box). (B) Juxtaposition of segments 1–2-HMG occurred upon insertion of Protamine between *CG8664* (turquoise box in *CG8664* 3' UTR corresponds to *DNAREP1* fragment) and *forked* genes, which we term the “original-*Dox*” (*ODox*). (C) *ODox* is inferred as an ancestral intermediate, since the contemporary *Dsim* *GD15682* locus lacks the HMG segment and only contains fused segments 1–2. (D) Another *Dsim* locus exhibits segments 1–2 without the HMG box, derived from insertion into *CG5004* (forming *Dsim* *GD17329*). (E) The *Dox* lineage with HMG acquired segment 4 by ancestral insertion of *ODox* into *CG15306*. We refer to HMG-bearing insertion into the *Dsim* ancestor as *ODox2*, again to reflect that it is not retained in present day *Dsim*. (F) The contemporary *Dsim* genome contains an unannotated gene referred to as *Ur-Dox*, which lacks the HMG box. However, ancestry to “*ODox2*” is reflected in the fact that the syntenic regions in *Dsech* and *Dmau* contain HMG box-containing *Dox* superfamily genes (see also Fig. 3). (G) Inferred insertion of *ODox2*, which juxtaposes segments 1–2-HMG-4 into a 359 segment, in the intron of *GD24701* (*CG43740*) yielded *ParaDox* (*PDox1*). (H) Duplication into *Dsim* *PDox2*. (I) acquire 359 and *Cubilin* fragment. (J) acquire proximity to *Ptpmeg2*. (K) Inferred order of events that underlie generation of *Dox*.

A nearly identical, dispersed copy (*PDox2*) is present in *Dsim*. **(I)** Mobilization of *PDox* between *Dsim* homologs of *CG15317* and *Cubilin* generated *MDox*. **(J)** *Dox* was generated by mobilization of *MDox* between *Dsim* ancestors of *Ptpmeg2* and *CG42797*. **(K)** Summary of mobilizations from an ancestral autosomal protamine copy through multiple regions of the X chromosome, ultimately yielding the contemporary *Dsim Dox* gene.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

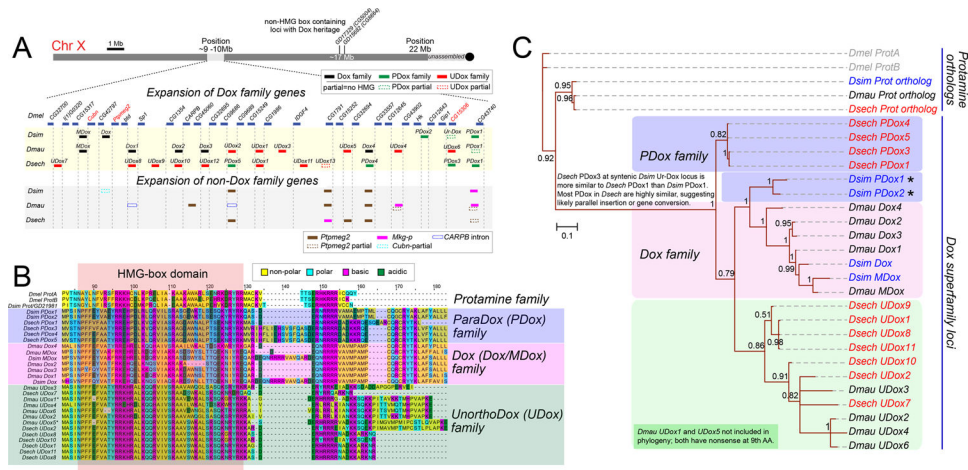
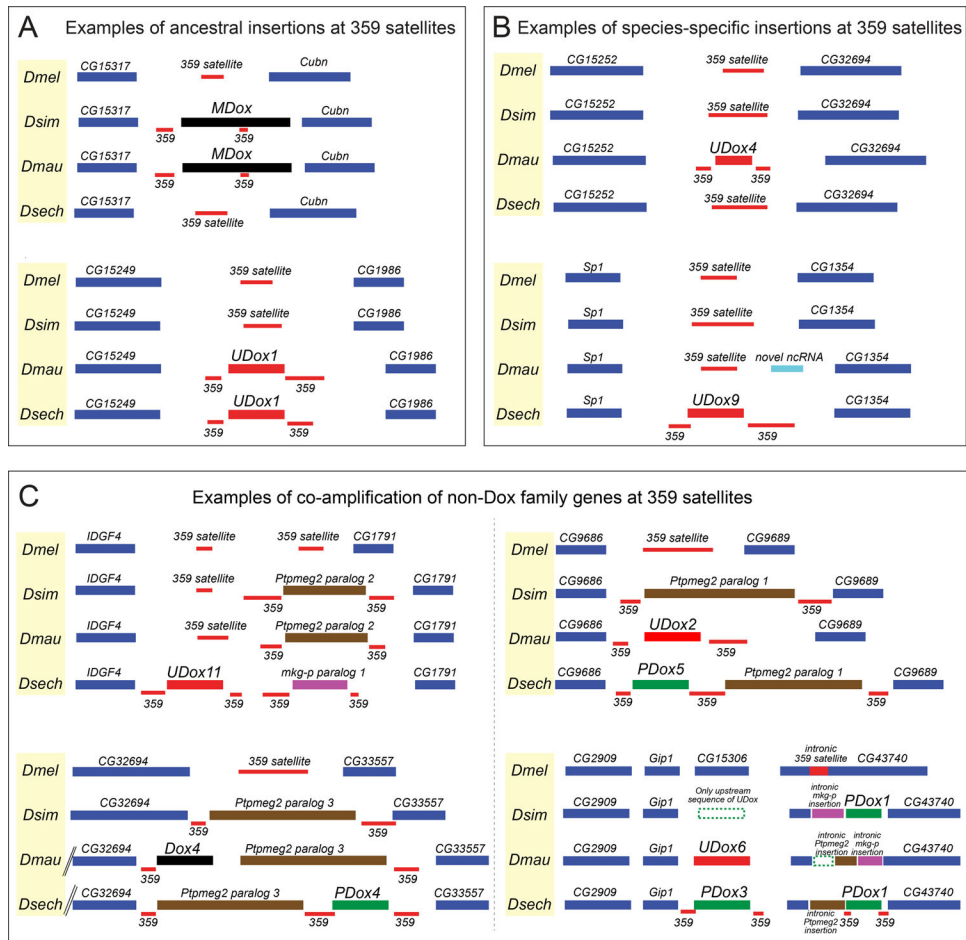


Figure 3. Evolution and diversification of the Dox superfamily in *simulans*-clade species. **(A)** Chromosomal view of expansion of Dox superfamily and non-Dox family expansions in ~ 1Mb genomic window on the X. Blue tiles show flanking genes as genomic regions to orient expanded copies of Dox superfamily members. **(B)** Classification of Dox superfamily loci into three subfamilies (PDox, Dox, and UDox) was based on amino acid similarity. The highlighted window within the protein alignment indicates the conserved HMG-box domain shared between Protamine and Dox family genes. **(C)** Phylogenetic tree showing similarity relationships among Dox superfamily loci. Highlighted boxes in the tree show clustering of Dox superfamily members into three subfamilies (PDox, Dox, and UDox) based on sequence similarity. Numbers in the tree nodes indicate posterior probability obtained from MrBayes analysis. *Dmel*/ProtA/B and *Dsim* Prot/GD21981 were used as outgroups.

**Figure 4.**

Examples of modes of Dox superfamily expansions in the *simulans*-clade. **(A)** Insertion of Dox superfamily loci is associated with 359 satellite repeats. Synteny analysis in the *mel*-complex revealed 359 sequences at insertion sites are conserved evidenced by their syntenic presence in *Dmel*. **(B)** Within *simulans*-clade examples of independent insertions of Dox superfamily members were found indicating their active spread within species. Novel, species-specific insertions/expansions are also associated with 359 satellite repeat at insertion sites, and the synteny of 359 repeat is preserved in other species that lack an insertion. **(C)** Spread of Dox superfamily loci is also linked to co-amplification of two non-Dox family genes on the X chromosome. *Ptpmeg2* and *mkg-p* gene amplification harbors signatures similar to Dox superfamily expansion, where these non-Dox family genes preferentially inserted at 359 satellite regions. *Ptpmeg2* and *mkg-p* co-amplification events show synteny at some instances (insertion between *CG32694* and *CG33557*), but also harbor independent insertions similar to Dox superfamily genes. A detailed synteny analyses of expansions of Dox superfamily with amplification of *Ptpmeg2* and *Mkg-p* are shown in Extended Data Fig. 9.

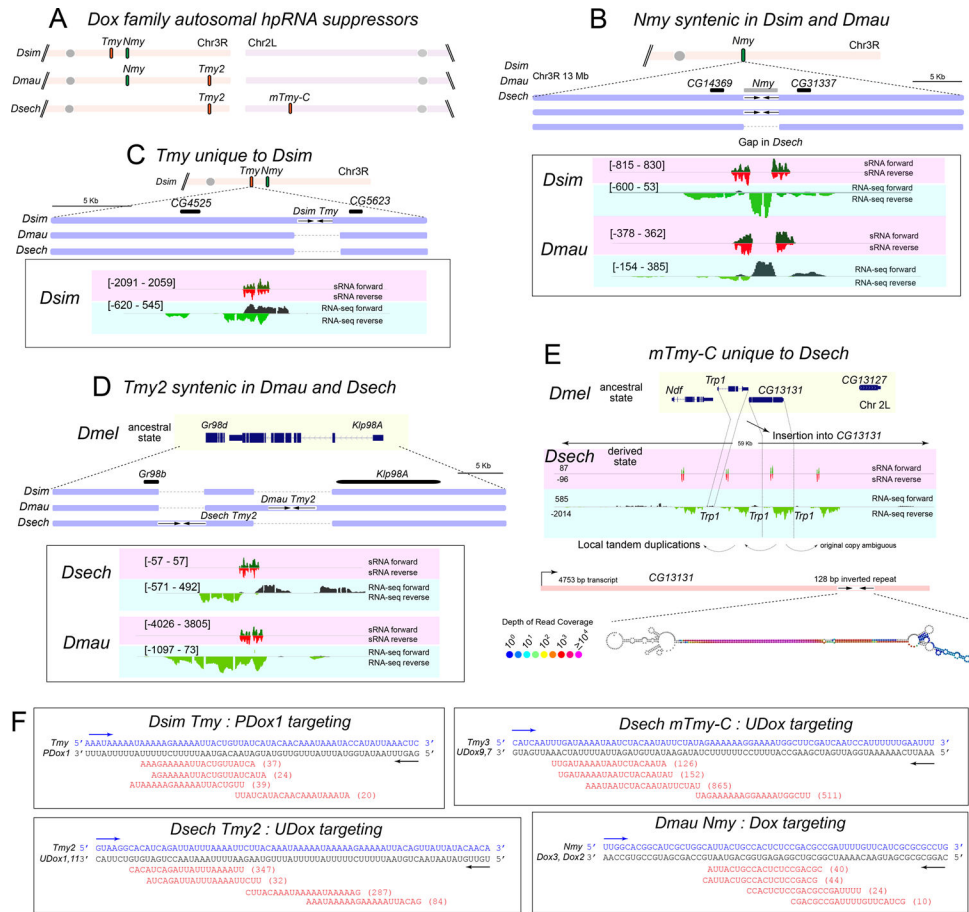


Figure 5. hpRNA:target evolution in the Dox superfamily network. (A) Chromosome map of recently-emerged autosomal hpRNAs targeting X-linked Dox superfamily in the *simulans*-clade. (B) Nmy hpRNA is syntenic only in *Dsim* and *Dmau*, but not *Dsech*. Flanking sequences reveal a gap at the corresponding *Dsech* region, while *Dsim* and *Dmau* Nmy share flanking genes *CG14369* and *CG31337*. (C) Tmy hpRNA is nearby Nmy on Chr3R and is unique to *Dsim*. Alignments show presence of *Tmy* only in *Dsim*, and gaps in *Dmau* and *Dsech*. However, the flanking genes *CG4525* and *CG5623* are preserved in all three species. (D) *Tmy2* hpRNA is syntenic between *Dmau* and *Dsech*. *Tmy2* emerged via insertion of a Dox superfamily member between *Gr98d* and *Klp98A*, followed by duplication to generate an hpRNA in the ancestral species. *Dsim* *Gr89b* and *Klp98A* exhibit *Dmel*-like ancestral state, but these genes are disrupted in *Dmau* and *Dsech* due to hpRNA birth at this locus. (E) The mini-Tmy-like hpRNA complex (*mTmy-C*) in *Dsech*. Gene models show *Dmel* ancestral state and location of the emergence of duplicated Tmy-like hpRNA cluster. The ancestor to the hpRNA cluster disrupted *CG13131* and in the contemporary state, this hpRNA is flanked by *Ndf* and *CG13127* genes. Local duplications also affect the flanking gene *Tip1*. Secondary structure for one hpRNA unit of the mTmy-C cluster. (F) Example of hpRNA:target relationships, with resulting small RNAs with antisense complementarity to targets. Number of small RNAs observed in *w[XDI]* testis dataset is shown in parentheses. For panels B, C, D, and E, normalized sRNA and RNA-seq tracks depict the structure and

expression from hpRNA. Forward strand sRNA reads are shown in dark green and reverse strand sRNAs in red. Similarly, forward strand RNA-seq reads shown in black, and reverse strand RNA-seq reads in light green. Y-axis values indicate normalized read counts for sRNA and RNA-seq.

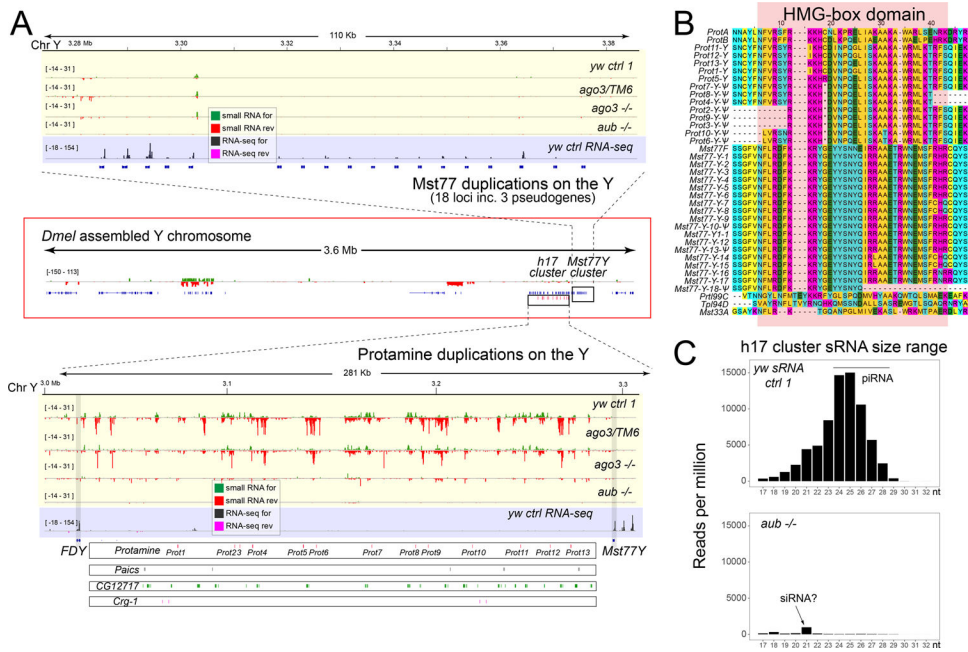


Figure 6.

Genomic expansion of protamine genes in *D. melanogaster*. As noted, the autosomal protamine locus is in a derived state in *Dmel*, as it is locally duplicated, unlike *simulans*-clade and outgroup *Drosophila* species (Figure 2A). (A) Genome browser tracks of small RNA data (yellow) and RNA-seq data (purple) from control (ctrl) and *ago3* heterozygous (over TM6) testis, as well as from piRNA pathway mutant testis (*ago3* and *aubergine/aub*). Two regions of the assembled Y chromosome (central red box) are shown as enlargements. Top, expansion of Mst77 genes. Protamine belongs to the MST-HMG box family, for which the autosomal member Mst77F was previously observed to have broadly expanded on the Y chromosome (Mst77Y cluster). Bottom, adjacent to the Mst77Y cluster, in the h17 cytoband, is another cluster bearing repeated portions of multiple protein-coding genes, including protamine. The annotated genes in the Mst77Y cluster are associated with testis RNA-seq evidence, but not small RNA data. By contrast, the h17 cluster is associated with abundant small RNA data, but not RNA-seq data. Most of these reads seem to be piRNAs, since they are depleted in piRNA mutant testes. (B) Alignment of MST-HMG box members indicates that the h17 copies on the Y are more closely related to Protamine than to Mst77 members or other MST-HMG box members. (C) Small RNAs from the h17 cluster are mostly piRNA-sized (~23–28 nt in these data) and are depleted in testis mutated for the piRNA factor *aubergine (aub)*. The minor population of h17 cluster small RNAs remaining in *aub* mutant testis appear siRNA-sized (21 nt).