nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-021-01592-z

Check for updates

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

Jeffrey Vedanayagam[1]✉, Ching-Jung Lin[1,2] and Eric C. Lai[1,2]✉

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.

While meiotic drive is widespread in plants, animals and fungi[1,2], we know little about its origins, molecular functions and short- and long-term persistence in wild populations[3]. 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 a biased progeny sex ratio (SR) of affected fathers that preferentially sire females[4]. SR drive systems occur broadly across eukaryotes, but are apparently lacking in some popular model systems. For example, strong SR drivers have not been identified in *Drosophila melanogaster* (*Dmel*), but its sister species *Drosophila simulans* (*Dsim*) harbours three different SR drive systems, termed Paris[5], Durham[6] 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 ago[9]. These closely related species are amenable to introgression genetics[10,11], yielding both SR drive and hybrid sterility factors that preferentially disrupt spermatogenesis[12]. The Durham drive system was uncovered during introgressions between *Dsim* and *Dsech*, where a minimal ~80 kb autosomal region was inferred to harbour 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 progeny[8]. 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 SR and normal spermatogenesis[7,13].

Nmy encodes retroposed *Dox* sequence forming an inverted repeat[9], 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. PacBio genomes from the *simulans* clade recently became available[14] and 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 that 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 (ORFs), their coding status and molecular origins were unknown[7].

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* testis[15], *Dox* encodes a 4.1 kb 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###*

[1]Developmental Biology Program, Sloan Kettering Institute, New York, NY, USA. [2]Weill Graduate School of Medical Sciences, Weill Cornell Medical College, New York, NY, USA. ✉e-mail: jeffreypratap@gmail.com; laie@mskcc.org

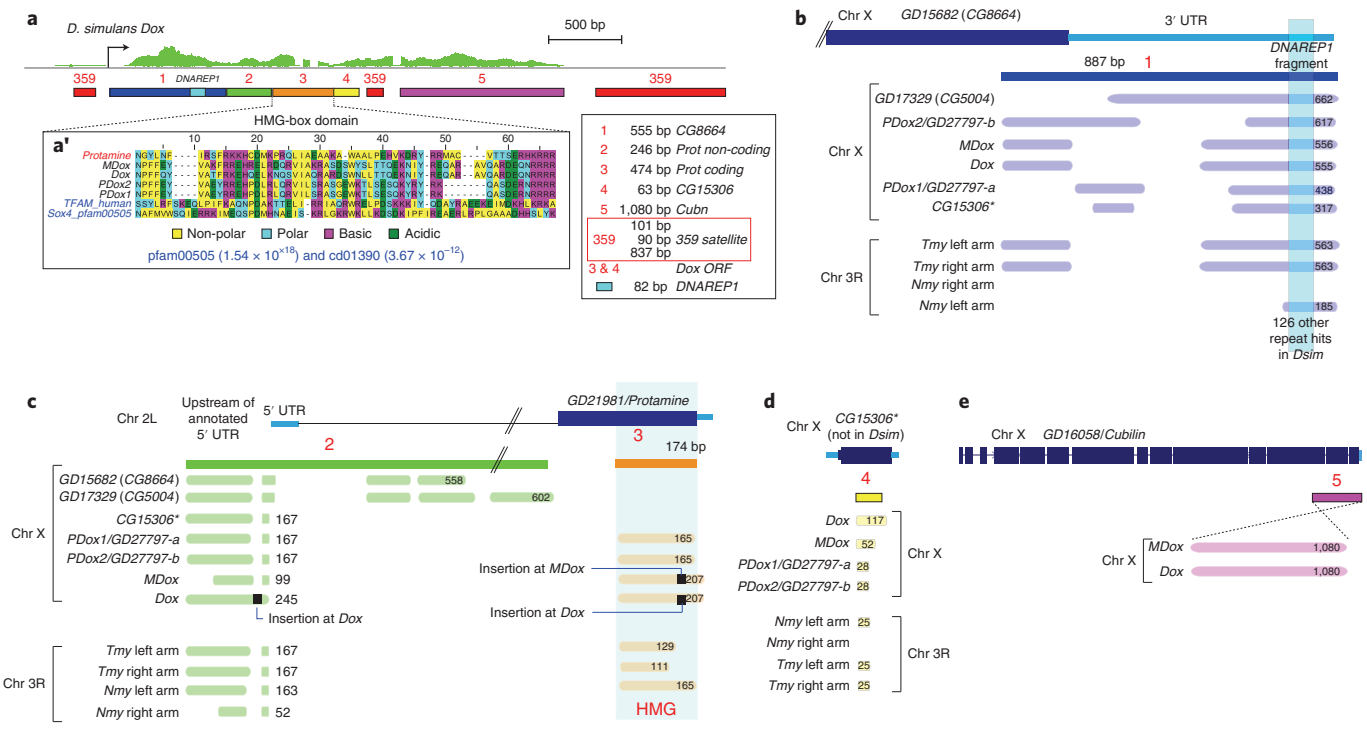

Fig. 1 | Structure of *Dox* transcript with segments acquired from various genes on the path to its origin. a, Testis RNA-seq data show a multi-exonic transcript from the *Dox* region, with several distinct segments acquired from protein-coding genes and repetitive elements. '359' corresponds to sequence 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 harbours a high mobility group (HMG)-box domain. Inset (a') highlights amino acid identity/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 887 bp from C-terminus and 3′ UTR of *GD15682* (*CG8664*). BLAST hits of various lengths to different genomic features on chr X and chr 3R 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 and 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*.

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). In *Drosophila*, 359 belongs to the complex 1.688 satellite repeat family. In *D. melanogaster*, a large block of 359 satellites resides in pericentromeric heterochromatin on the X chromosome[17], but *simulans* clade species harbour expanded 359 satellites, including a large block within euchromatic X (ref. [18]).

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 spermatids[19].

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 four 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*, with 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 high mobility group (HMG)-box domain that binds DNA[20]. While not recognized earlier[7], Conserved Domain Database (CDD v.3.19) retrieved significant hits ($e < 0.001$) that include signature residues of the general HMG-box domain (Fig. 1a, inset). *Dox* factors even exhibit homology to human HMG-box domains (Fig. 1a, inset), 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 ~1.1 kb homology to the C-terminal-encoding region of *Cubilin*. *Cubilin* matches to both *MDox* and *Dox*, but not other *Dox* family genes, indicating that 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 ORF (termed ORF13) in this region (Extended Data Fig. 2). In addition, copies of a potential ORF 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 (Fig. 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 (that is, 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 flux[21] (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,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 analysed 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 genomes[14].

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 that they form a distinct subfamily, which we term the *UnorthoDox* (*UDox*) genes. Although *UDox* and

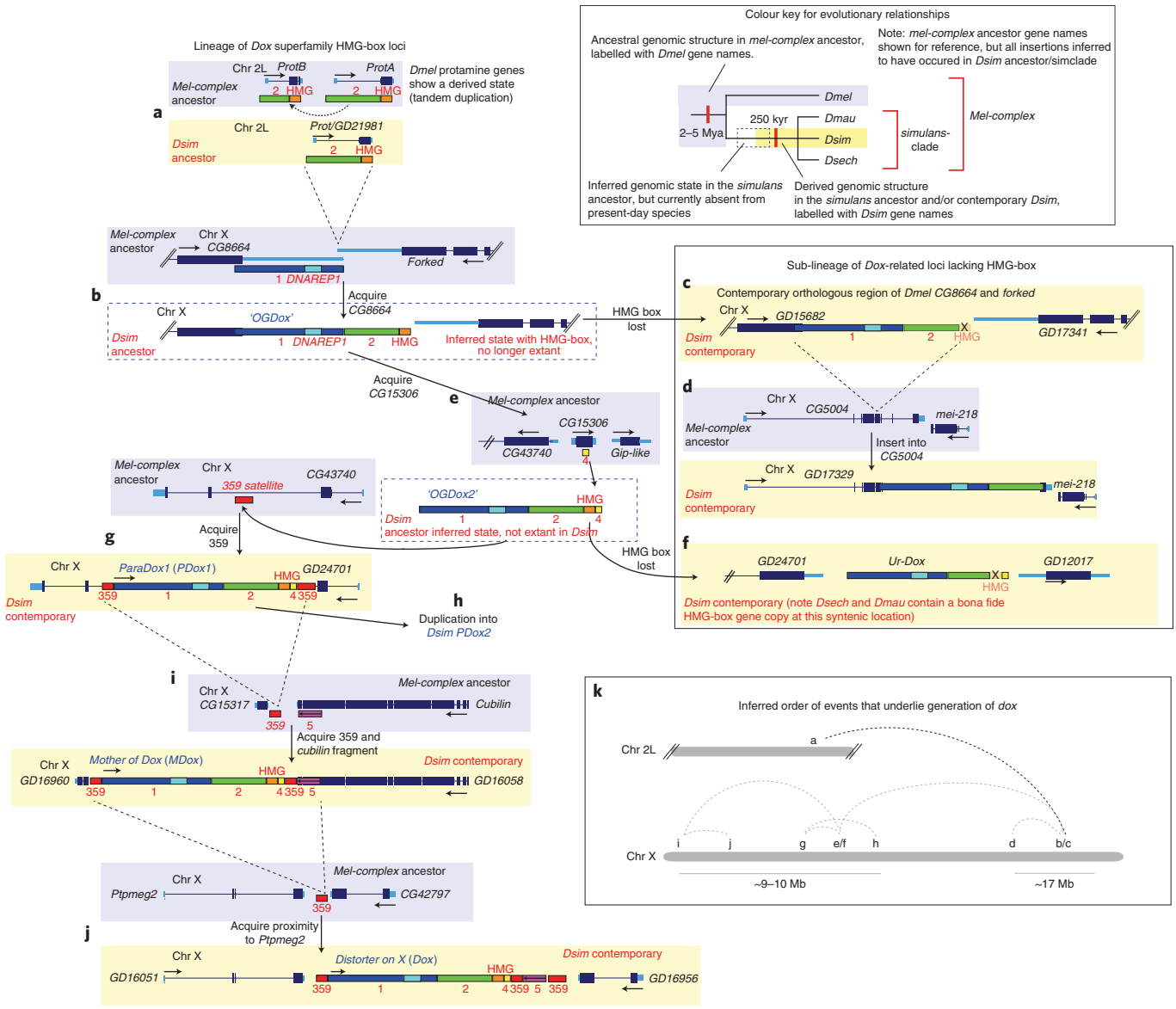

Fig. 2 | Stepwise origins of *Dox* from a *Protamine*-like ancestor with key for labelling of gene names and structures. Note that many regions correspond to extant genomic loci in *Dmel* (purple) and Dsim (yellow), but mobilizations occurred in a *simulans* clade ancestor. They are not meant to indicate that 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**, 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.

PDox families share higher homology to *Tmy* than *Nmy*, distinct sequences cluster these duplications separately (Fig. 3c). *Dsech* harbours 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

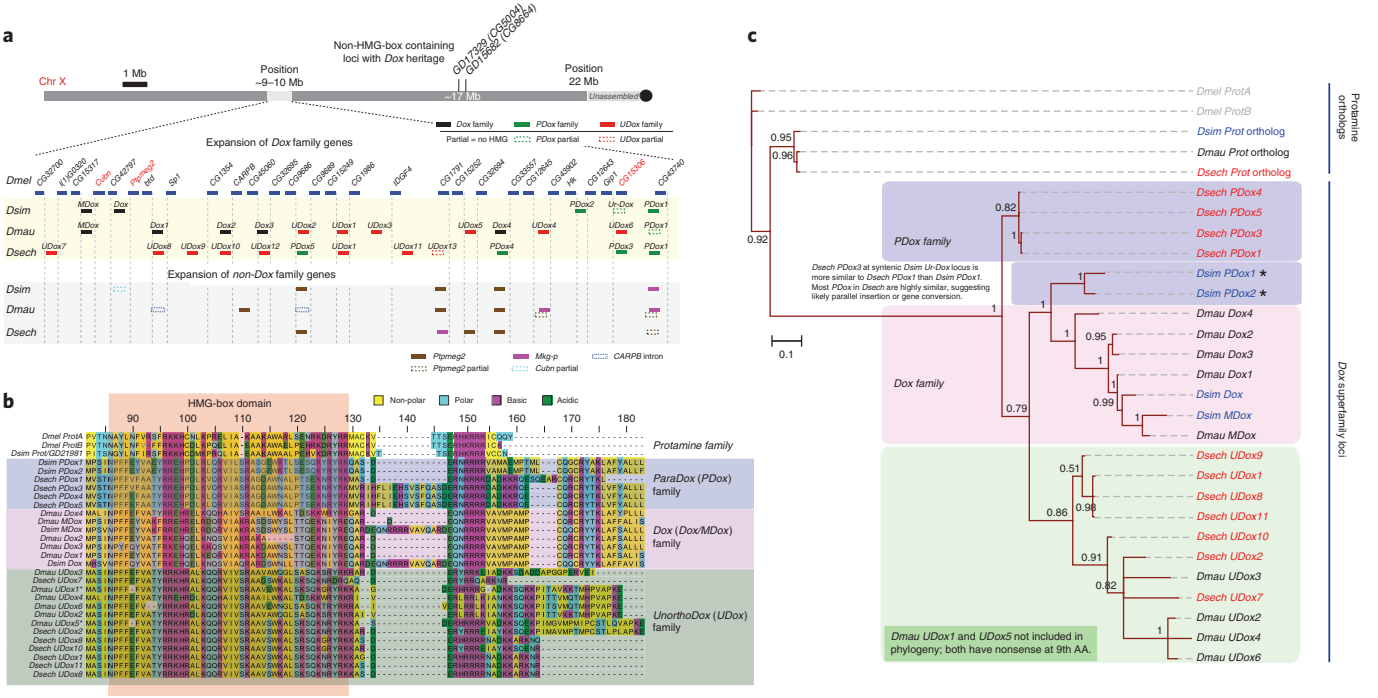

Fig. 3 | Evolution and diversification of the *Dox* superfamily in *simulans* clade species. a, Chromosomal view of expansion of *Dox* superfamily and non-*Dox* family expansions in ~1 Mb genomic window on the X. Blue tiles show flanking genes as genomic regions to orient expanded copies of *Dox* superfamily 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 amongst *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.

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 and Extended Data Fig. 5). In addition to this syntenic copy, there are three 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* clade[22], 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, which raises the question of 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 harbouring 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 harbour pre-existing 359

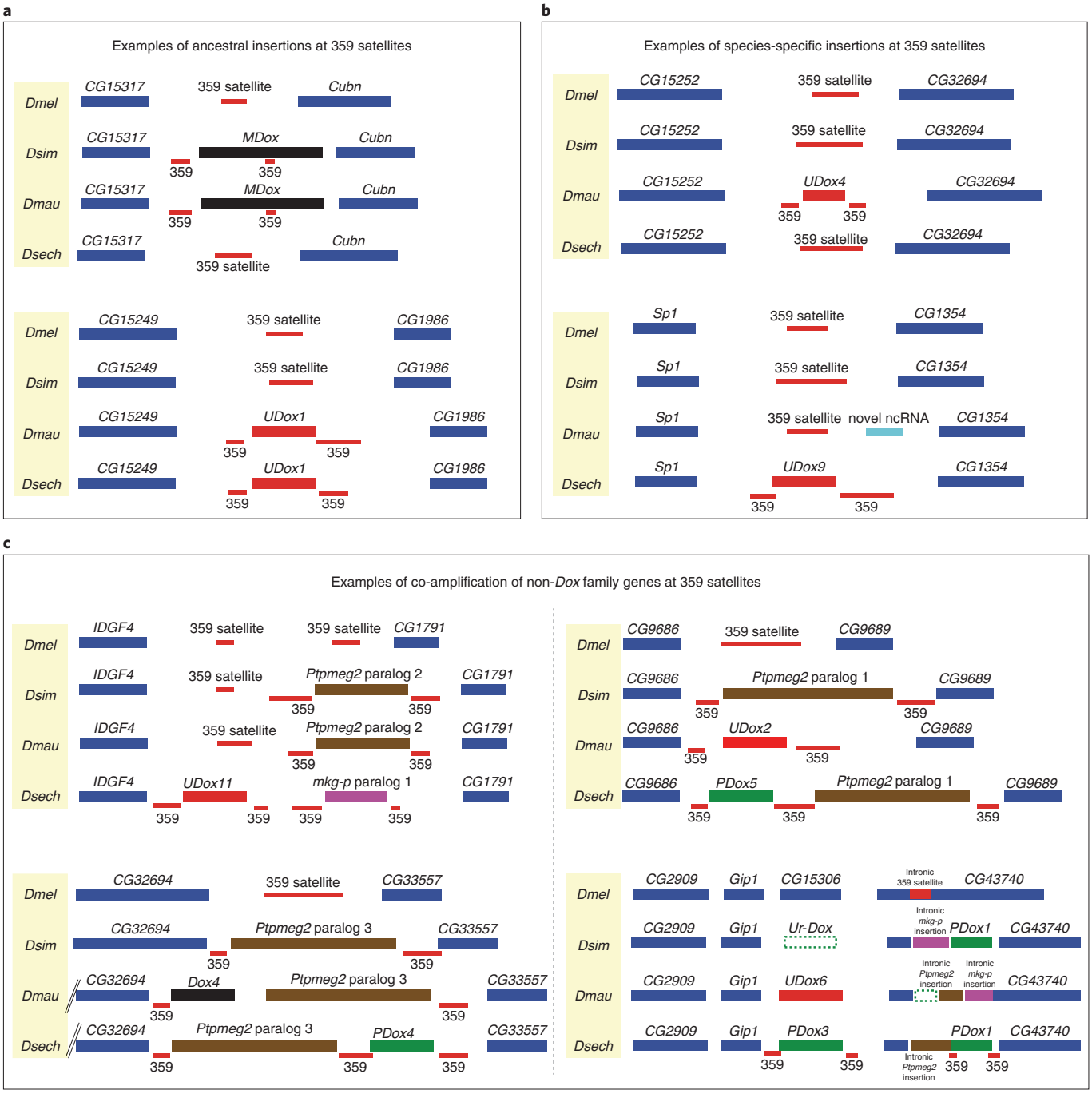

Fig. 4 | Examples of modes of Dox superfamily expansions in the *simulans* clade. a, Insertion of *Dox* superfamily loci is associated with 359 satellite repeats. Synteny analysis in the mel-complex revealed that 359 sequences at insertion sites are conserved as evidenced by their syntenic presence in *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 harbours 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 harbour independent insertions similar to *Dox* superfamily genes. Detailed synteny analyses of expansions of *Dox* superfamily with amplification of *Ptpmeg2* and *mkg-p* are shown in Extended Data Fig. 9.

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 conversion[27], or perhaps insertions via excised circular DNA[28].

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 targets[16]. *Dsim Nmy* is flanked by *GD26005*/*CG14369* and *GD20491*/*CG31337* (Fig. 5a). Synteny analysis shows that *Nmy* is also flanked by these genes in *Dmau*, but no such hpRNA exists at 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) implies another suppressor of these loci.

We next examined *Tmy*, located ~2 Mb 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 phenotypes[6]. Nevertheless, these genetic experiments alone do not actually mean that other species lack *Tmy*. One can only conclude that the syntenic region does not harbour 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 that *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 that the *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 ~10 Mb away from *Tmy2* in a more central location of 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 also to an uncharacterized autosomal region that proved to be the *Tmy* hpRNA[15]. 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 ~130 bp 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-C* 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 versus *Dmel*. However, as canonical *Protamine* genes duplicated in *Dmel* compared with the *simulans* clade (Fig. 2a), and exhibit signatures of positive selection[21], *Protamine* loci are subject to recurrent rapid evolution.

We examined the possibility of additional alterations in *Protamine* genes in *Dmel*. Interestingly, queries to *Dmel* Y (https://flybase.org) and the improved *Dmel* PacBio Y chromosome contigs[29] 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 chr Y[30], 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* (small ubiquitin-like modifier protease) and *Crg-1* (forkhead transcription factor)[30]. However, subsequent work clarified that CG46192 family and Mst77 family proteins contain the MST-HMG-box domain found in testis-restricted proteins[20]. Of note, both Mst77F and protamines replace histones during compaction of sperm chromatin[31]. 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 on 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 wild-type 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 revision[33]. 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.

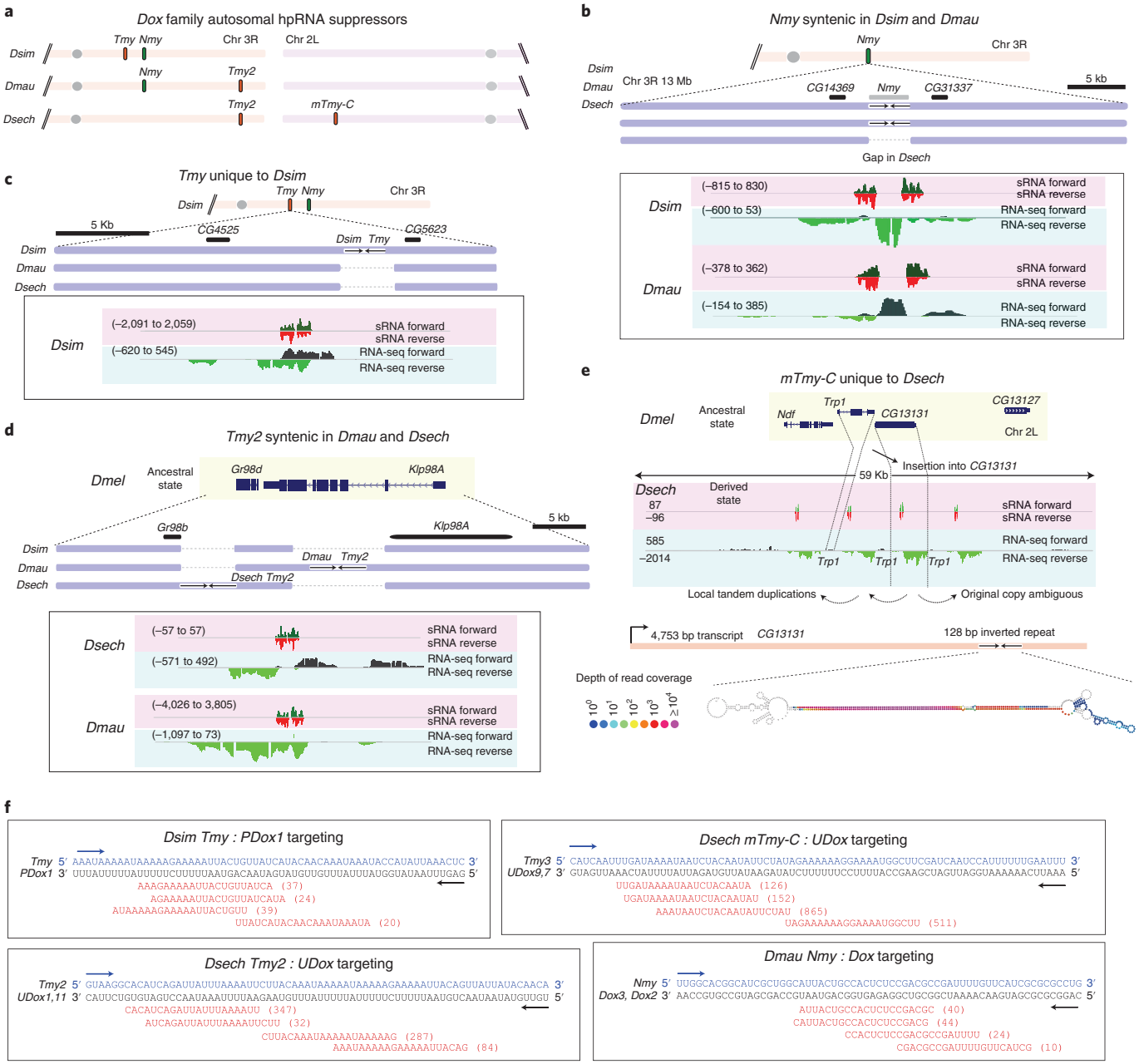

Fig. 5 | hpRNA:target evolution in the *Dox* superfamily network. a, Chromosome map of recently emerged autosomal hpRNAs targeting X-linked *Dox* superfamily in the *simulans* clade. **b**, *Nmy* hpRNA is syntenic only in *Dsim* and *Dmau* but not *Dsech*. Flanking sequences reveal a gap at the corresponding *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 *Trp1*. 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[XD1]* testis dataset is shown in parentheses. In **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 are shown in black, and reverse-strand RNA-seq reads in light green. *Y* axis values indicate normalized read counts for sRNA and RNA-seq.

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 ORFs that are shared with other genomic sources, and their syntenies 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

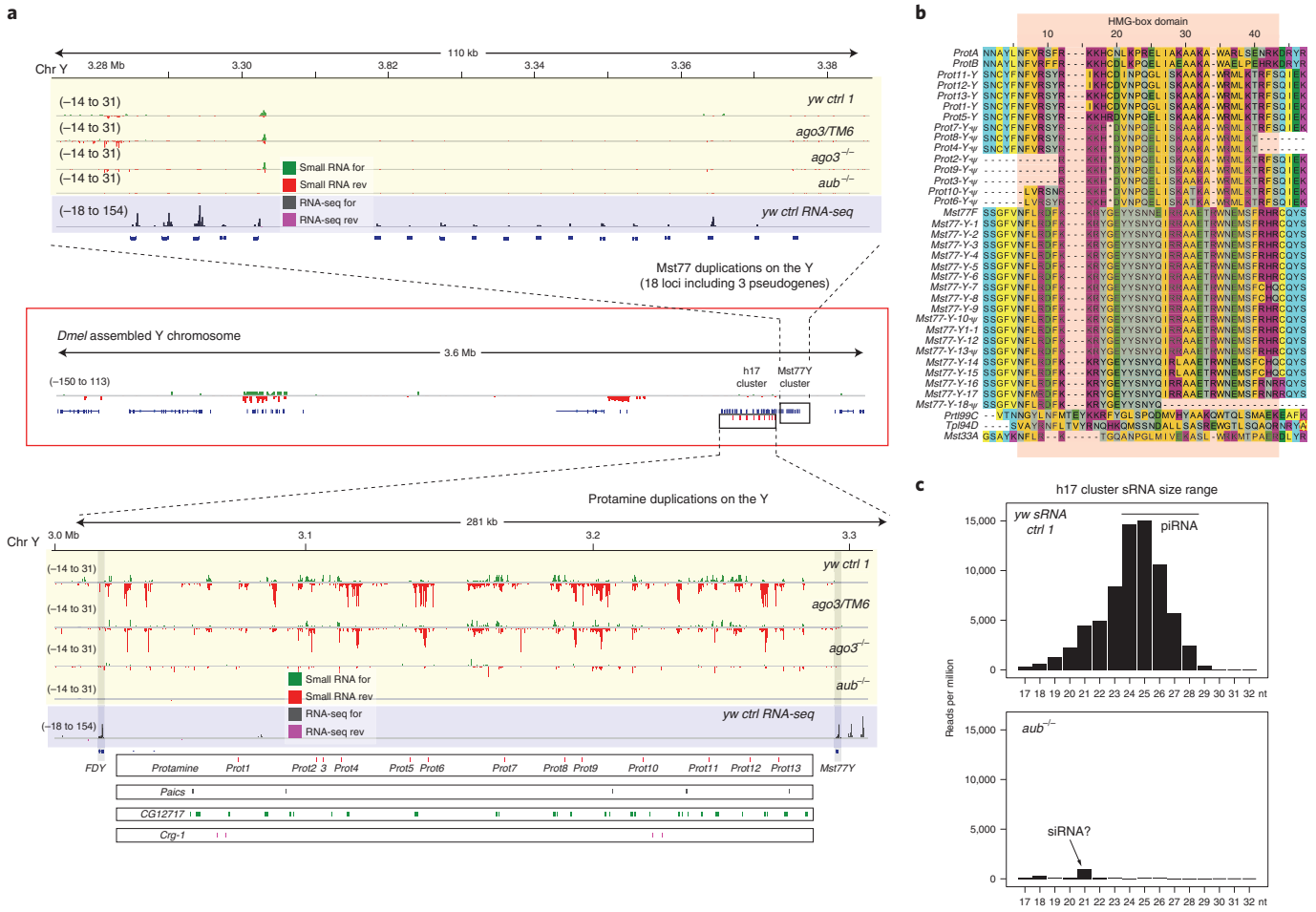

Fig. 6 | Genomic expansion of protamine genes in *D. melanogaster*. The autosomal *Protamine* locus is in a derived state in *Dmel*, as it is locally duplicated, unlike *simulans* clade and outgroup *Drosophila* species (Fig. 2a). **a**, Genome browser tracks of small RNA data (yellow background) and RNA-seq data (purple background) from control (*ctrl*; yellow white (*yw*)) 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. The h17 cluster is flanked by *flagrante delicto Y* (*FDY*) and *Mst77Y* genes highlighted in grey. **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).

into at least three recognizable subfamilies (and potentially more, depending on the granularity of subdivision), suggests that many members of this family may 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 hpRNAs[36]. By and large, our studies appear largely concordant, although they implement a single nomenclature for all novel *Dox* superfamily copies as *Dxl-1 to 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* (ref. [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 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 flies[37] and mammals[38]. 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 species[39]. 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 most abundant *Drosophila* satellite sequence[25,26]. Precise analyses of the genomic make-up 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 species[18]. 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.

Methods

Genome and transcriptome data. PacBio genome data for *simulans* clade species[14] 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* testis[15], 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). Total RNA (1 µg) was used to prepare small RNA libraries as described[40], 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 h containing 200 pmol 3′ linker, 1× 5′ DNA adenylation reaction buffer, 100 nM ATP and 200 pmol Mth RNA ligase, and the reaction is terminated by being heated to 85 °C for 5 min. Adenylated 3′ linker was then precipitated using ethanol and was used for 3′ ligation reaction containing 10% PEG8000, 1× RNA ligase buffer, 20 µM adenylated 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-polyacrylamide gel electrophoresis gel. The small RNA-3′ linker hybrid was then subjected to 5′ ligation reaction at 37 °C for 4 h containing 20% PEG8000, 1× 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 the following components to the ligated product: 2 µl 5× RT buffer, 0.75 µl 100 mM dithiothreitol, 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 onto ice. Superscript III RT enzyme (0.5 µL) and 0.5 µL RNase OUT were added to the RT mix, and the reaction was carried out at 50 °C for 1 h. cDNA libraries were amplified using 15 cycles of PCR with forward and illumina 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[1]* 14021-0241.60, which were used for PacBio genome sequencing[14], and *Dsech* 14021-0248.25 used for the Sanger assembly[41]. 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 artefacts caused by PCR amplification. All samples were pooled together, using the barcoded adapters provided by the manufacturer, over two 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 were mapped to PacBio genome assemblies for *Dsim*, *Dmau* and *Dsech* using hisat2 aligner with the 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 SAMTOOLS[42] with the following commands: (1) 'samtools view -bS file.sam > file.bam', (2) 'samtools sort file.bam > file_sorted.bam' and (3) 'samtools index file_sorted.bam'. Mapping statistics for the BAM alignment, and visualization BigWig files were obtained using the bam_stat.py and bam2wig.py scripts, respectively, from the RSeqQC package[43].

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 4 bp random-linker sequence inserted at 5′ and 3′ of the sRNA sequence (total 8 bp). After filtering ≤15 nt reads, we mapped the small RNA data to PacBio genome assemblies using Bowtie (with options '-v0 -best -strata'.) The resulting small RNA alignments in SAM format were converted to BED for downstream processing using the BEDops software and visualized on Integrative Genomics Viewer.

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+ (ref. [44]). Search for conserved protein domains in the *Dox* family genes was performed using both HMMER v.3.3.2 and National Center for Biotechnology Information Conserved Domain Database (CDD v.3.19).

Phylogenetic analysis. For phylogenetic analysis of *Dox* superfamily genes, we constructed an alignment of coding sequence 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 of 12, (2) gap extension penalty of 3 and (3) two refinement iterations. The resulted alignment was then manually curated to ensure proper alignment. Phylogenetic analysis on the nucleotide alignment was performed using the MrBAYES[45] plugin in Geneious software v.11.0.4. For this analysis, we used the HKY85 substitution model with *Dmel ProtA*/*B* as an outgroup. A gamma rate variation option was used with four gamma categories. For Monte Carlo Markov chain settings, we used the following parameters: (1) chain length ranging from 100,000 to 150,000 based on the trace file, (2) four heated chains, (3) heated chain temp of 0.2, (5) subsampling frequency of 100, (6) burn-in length of 1,000 and (7) a random seed of 7,826. For priors, we used the 'unconstrained branch lengths' option with GammaDir parameters of (1, 0.1, 1) and shape parameter of exponential (10).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Paired-end RNA-seq reads from *Dmau*, *Dsim* and *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.

Received: 30 April 2021; Accepted: 19 October 2021;

References

1. Zanders, S. E. & Unckless, R. L. Fertility costs of meiotic drivers. *Curr. Biol.* **29**, R512–R520 (2019).
2. Agren, J. A. & Clark, A. G. Selfish genetic elements. *PLoS Genet.* **14**, e1007700 (2018).
3. Lindholm, A. K. et al. The ecology and evolutionary dynamics of meiotic drive. *Trends Ecol. Evol.* **31**, 315–326 (2016).
4. Jaenike, J. Sex chromosome meiotic drive. *Annu Rev. Ecol. Syst.* **32**, 25–49 (2001).
5. Helleu, Q. et al. Rapid evolution of a Y-chromosome heterochromatin protein underlies sex chromosome meiotic drive. *Proc. Natl Acad. Sci. USA* **113**, 4110–4115 (2016).
6. Tao, Y., Hartl, D. L. & Laurie, C. C. Sex-ratio segregation distortion associated with reproductive isolation in *Drosophila*. *Proc. Natl Acad. Sci. USA* **98**, 13183–13188 (2001).

7. Tao, Y. et al. A sex-ratio meiotic drive system in *Drosophila simulans*. II: an X-linked distorter. *PLoS Biol.* **5**, e293 (2007).
8. Tao, Y., Masly, J. P., Araripe, L., Ke, Y. & Hartl, D. L. A sex-ratio meiotic drive system in *Drosophila simulans*. I: an autosomal suppressor. *PLoS Biol.* **5**, e292 (2007).
9. Garrigan, D. et al. Genome sequencing reveals complex speciation in the *Drosophila simulans* clade. *Genome Res* **22**, 1499–1511 (2012).
10. Masly, J. P. & Presgraves, D. C. High-resolution genome-wide dissection of the two rules of speciation in *Drosophila*. *PLoS Biol.* **5**, e243 (2007).
11. Presgraves, D. C., Gerard, P. R., Cherukuri, A. & Lyttle, T. W. Large-scale selective sweep among segregation distorter chromosomes in African populations of *Drosophila melanogaster*. *PLoS Genet.* **5**, e1000463 (2009).
12. Meiklejohn, C. D. & Tao, Y. Genetic conflict and sex chromosome evolution. *Trends Ecol. Evol.* **25**, 215–223 (2010).
13. Kingan, S. B., Garrigan, D. & Hartl, D. L. Recurrent selection on the Winters sex-ratio genes in *Drosophila simulans*. *Genetics* **184**, 253–265 (2010).
14. Chakraborty, M. et al. Evolution of genome structure in the *Drosophila simulans* species complex. *Genome Res* **31**, 380–396 (2021).
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).
16. Wen, J. et al. Adaptive regulation of testis gene expression and control of male fertility by the *Drosophila* hairpin RNA pathway. *Mol. Cell* **57**, 165–178 (2015).
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**, 1343–1349 (2007).
18. Garrigan, D., Kingan, S. B., Geneva, A. J., Vedanayagam, J. P. & Presgraves, D. C. Genome diversity and divergence in *Drosophila mauritiana*: multiple signatures of faster X evolution. *Genome Biol. Evol.* **6**, 2444–2458 (2014).
19. Rathke, C. et al. Transition from a nucleosome-based to a protamine-based chromatin configuration during spermiogenesis in *Drosophila*. *J. Cell Sci.* **120**, 1689–1700 (2007).
20. Doyen, C. M. et al. A testis-specific chaperone and the chromatin Rremodeler ISWI mediate repackaging of the paternal genome. *Cell Rep.* **13**, 1310–1318 (2015).
21. Dorus, S., Freeman, Z. N., Parker, E. R., Heath, B. D. & Karr, T. L. Recent origins of sperm genes in *Drosophila*. *Mol. Biol. Evol.* **25**, 2157–2166 (2008).
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**, 523–527 (2004).
23. Khost, D. E., Eickbush, D. G. & Larracuente, A. M. Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. *Genome Res* **27**, 709–721 (2017).
24. Sproul, J. S. 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).
25. Travaglini, E. C., Petrovic, J. & Schultz, J. Satellite DNAs in the embryos of various species of the genus *Drosophila*. *Genetics* **72**, 431–439 (1972).
26. de Lima, L. G., Hanlon, S. L. & Gerton, J. L. Origins and evolutionary patterns of the 1.688 satellite DNA family in *Drosophila* phylogeny. *G3 (Bethesda)* **10**, 4129–4146 (2020).
27. Haudry, A., Laurent, S. & Kapun, M. Population genomics on the fly: recent advances in *Drosophila*. *Methods Mol. Biol.* **2090**, 357–396 (2020).
28. Thomas, J., Phillips, C. D., Baker, R. J. & Pritham, E. J. Rolling-circle transposons catalyze genomic innovation in a mammalian lineage. *Genome Biol. Evol.* **6**, 2595–2610 (2014).
29. Chang, C. H. et al. Islands of retroelements are major components of *Drosophila* centromeres. *PLoS Biol.* **17**, e3000241 (2019).
30. Krsticevic, F. J., Schrago, C. G. & Carvalho, A. B. Long-read single molecule sequencing to resolve tandem gene copies: the Mst77Y region on the *Drosophila melanogaster* Y chromosome. *G3 (Bethesda)* **5**, 1145–1150 (2015).
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**, 6165–6177 (2005).
32. Nagao, A. et al. Biogenesis pathways of piRNAs loaded onto AGO3 in the *Drosophila* testis. *RNA* **16**, 2503–2515 (2010).
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).
34. Malone, C. D. et al. Specialized piRNA pathways act in germline and somatic tissues of the *Drosophila* ovary. *Cell* **137**, 522–535 (2009).
35. Lau, N. et al. Abundant primary piRNAs, endo-siRNAs and microRNAs in a *Drosophila* ovary cell line. *Genome Res* **19**, 1776–1785 (2009).
36. Muirhead, C. A. & Presgraves, D. C. Satellite DNA-mediated diversification of a sex-ratio meiotic drive gene family in *Drosophila*. *Nat. Ecol. Evol.* (in the press).
37. Rathke, C., Baarends, W. M., Awe, S. & Renkawitz-Pohl, R. Chromatin dynamics during spermiogenesis. *Biochim. Biophys. Acta* **1839**, 155–168 (2014).
38. Wang, T., Gao, H., Li, W. & Liu, C. Essential role of histone replacement and modifications in male fertility. *Front Genet* **10**, 962 (2019).
39. Meiklejohn, C. D. et al. Gene flow mediates the role of sex chromosome meiotic drive during complex speciation. *eLife* **7**, e35468 (2018).
40. Lee, J. E. & Yi, R. Highly efficient ligation of small RNA molecules for microRNA quantitation by high-throughput sequencing. *J. Vis. Exp.* **93**, e52095 (2014).
41. Clark, A. G. et al. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**, 203–218 (2007).
42. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
43. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).
44. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
45. Huelsenbeck, J. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).

Acknowledgements

The authors thank J. Wen for initial investigations into the domain structure of *Dox*, P. Smibert for the first recognition that *Dox*/*MDox* bear homology to protamine, A. Geneva for his inputs on phylogenetic analysis, D. 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, J.V.), US–Israel Binational Science Foundation (BSF-2015398, E.C.L.), National Institute of General Medical Sciences (R01-GM083300, E.C.L.) and National Institutes of Health MSK Core Grant (P30-CA008748).

Author contributions

J.V. and E.C.L. conceived and designed the study. J.V. performed all the computational analyses. J.V. and C.-J.L. collected testis tissue samples, and C.-J.L. generated libraries for RNA-seq and small RNA sequencing. J.V. analysed and interpreted the data with inputs from E.C.L. J.V. and E.C.L. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41559-021-01592-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-021-01592-z.

Correspondence and requests for materials should be addressed to Jeffrey Vedanayagam or Eric C. Lai.

Peer review information *Nature Ecology & Evolution* thanks Aaron Vogan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.