Supplementary Information

Satellite DNA-mediated diversification of a *sex-ratio* meiotic drive gene family in *Drosophila*

Christina A. Muirhead[1,2] and Daven C. Presgraves[1]

[1] Department of Biology, University of Rochester, Rochester, New York 14610, USA

[2] Ronin Institute, Montclair, New Jersey 07043

Text S1: Additional sequence insertions into *Dox*-region 359 clusters

The sat359 clusters in the *Dox* region (*Dmel* r6 X:9.4-10.4Mb) harbor several other non-native genes (**Extended Data Fig. 1**). There are at least three additional possible protein-coding genes. The sequence fragment referred to as "Ptpmeg2*"* is present in seven *Dxl* insertion loci (**Extended Data Fig. 1, 3**). It derives from genomic region *Dmel* r6 X:9628694-9631109, including most of the last six exons of the gene *Ptpmeg2*, a protein tyrosine phosphatase, and if transcribed is predicted to produce an ORF when spliced. As all "Ptpmeg2" fragments are flanked by sat359 sequence, these inserts may have been incidentally amplified along with *Dxl* insertions by a similar mechanism. The presence of "Ptpmeg2" in all three species at *Dxl-1*, *Dxl-4*, and *Dxl-6* suggests that "Ptpmeg2" sequence was inserted and amplified to multiple sites in the common ancestor of the *D. simulans* clade. A "mkg-p" fragment present in three *Dxl* insertion loci is an apparent retrotransposition of the lineage-specific gene *monkey-king protein* [53]. The mkg-p fragments are inserted into the 5' region in some copies of the "Ptpmeg2" insert. Lastly, in *D. sechellia*, there is one instance of what appears to be a transposase-like gene inserted into and flanked by sat359 repeats (**Extended Data Fig. 3**). In general, where there are multiple, different sequences inserted into a single sat359 cluster, they follow a consistent orientation, with the main, full-length, *Dxl* genes upstream (with respect to likely direction of transcription) of “Ptpmeg2” insertions. There are a few instances however, in which *Dxl* sequences are downstream of “Ptpmeg2” (*e.g.*, *Dxl-3* and *Dxl-4* in *D. mauritiana*; **Extended Data Fig. 1**). There is also at least one instance of a secondary inclusion of “Ptpmeg2” (*Dxl-1* in *D. sechellia*; **Extended Data Fig. 1**). The “Ptpmeg2”, “mkg-p”, and transposase-like sequence have inserted into *Dxl*-sat359 loci from elsewhere in the genome. Other inserted sequence, such as *CARPB* and the *cubn* sequence of *Dox* and *MDox*, derives from DNA that is immediately adjacent to sat359 sequence. These may have been incorporated as incidental passengers during the amplification process. One possible mechanism, movement through a circular DNA intermediate and re-incorporation via sat359 sequence homology, has some support in the “CARPB” intron sequence that has been incorporated in *D. mauritiana* (**Extended Data Fig. 4**).

Text S2: Origins of *Dox* sequence: regions X:17.1 and X:17.2Mb

In addition to the *Dxl* genes and the putative suppressors, there are at least two other genomic locations containing *Dox*-related sequence (*Dmel* r6, X:17.1Mb and X:17.2Mb; **Extended Data Fig. 2**). One of these may have provided the original source of the *Ur-Dox*-creating insertion into *CG15306*. In the *D. simulans* clade, both the X:17.1Mb and X:17.2Mb regions harbor large but now degraded insertions containing mixtures of non-native genomic DNA, transposable element fragments, and significant portions of at least two protein-coding genes, *tapas* and *krimper* (**Extended Data Fig. 2**). Although there is no extant protamine CDS sequence in X:17.1 or X:17.2 in any of the three species, the presence of *ProtA/B* upstream, *ProtA/B* intron sequence, and protamine-like coding sequence in *Dxl* genes suggests that protamine-like CDS was originally present at these sites but was subsequently deleted.

Region X:17.2 harbors an insertion in the 3'-UTR of *CG8664*, and now contains TE fragments, *ProtA/B* adjacent sequence, and a partial duplication of the gene *tapas*. Approximately 1.2 kb of this sequence at the proximal end, spanning the *ProtA/B*-adjacent sequence and part of the *CG8664* UTR and CDS, is readily alignable with *Dxl* sequences. The entire X:17.2 insert is also alignable with additional *simulans*-clade inserted sequence at X:17.1. The sequence inserted into X:17.1 contains the *tapas* duplication as well as some of the TE insertions found in X:17.2, but lacks some of the *CG8664* sequence at X:17.2 retained by *Dxl* genes. The X:17.1 insertion also contains additional TEs and fragments of the gene, *krimper*.

The functional significance of these insertions is unclear, but several features are suggestive. First, in *D. sechellia*, ~4.3kb consisting of the X:17.1 insert plus a small amount of flanking sequence has been duplicated tandemly for a total of nine copies of highly-similar sequence. This observation is consistent with a history of selection for increased gene dosage. Second, the presence of two piRNA pathway gene sequences, *tapas* and *krimper*, is either an extraordinary coincidence or suggest that these sequences had consequences for the small RNA regulation of *Dxl* genes. We and others [25] have only detected ~22-nt esiRNAs, but it is possible that *Dxl* genes were once regulated by piRNAs. Finally, we note that Tao et al. (2007) genetically mapped an *Enhancer of Dox* [*E(Dox)*]— a factor that increases the strength of *Dox*-mediated drive but does not itself drive— to a locus just proximal to the *forked* mutation, which corresponds to the location of *CG8664* but possesses sequence fragments with similarity to *Dox*. We suggest that this locus corresponds to the X:17.2 region described here and is therefore a strong candidate sequence for *E(Dox)*.

Text S3: Inference of *Dxl* protamine sequences

The protamine-encoding 157-codon ORF occurs in two of four discovered *Dox* transcripts [23], GenBank accessions EF596895.1 (functional *Dox*) and EF596897.1 (non-driving *dox*). We assumed conservation of those transcript structures for all lines which retained the relevant GT/AG splice junctions, and verified the existence of the spliced transcript forms in the RNA-seq samples by assessing the number of reads containing exact 20 bp matches for both spliced and unspliced (5'-end) forms (**Source Data Files** and **Supplementary Table 1**). Translations of the 157-codon ORF in spliced *Dxl* genes were aligned with several known *Drosophila* protamines. We aligned the putative HMG box domains from *Dxl* gene translations (amino acid positions 1-66) with those of *D. melanogaster ProtA* and *ProtB* (**Source Data Files**) and with their syntenically homologous amino acid sequences in the three *D. simulans* clade species and with *D. yakuba*, *D. erecta*, and *D. ananassae*. These alignments were used for phylogenetic analyses (**Fig. 3b**). A maximum likelihood tree was inferred using RAxML (v8.2.11), model "PROTGAMMAAUTO" with 1000 bootstrap runs, and drawn with FigTree (v1.4.4). Of the *Dxl* gene sequences not included in this analysis: in *D. mauritiana*, *Dxl*-5 and *Dxl*-9 have premature stop codons, whereas *Dxl-1* lacks the region entirely; in *D. sechellia*, *Dxl-6* and *Ur-Dox* lack start codons and *Dxl-1b* lacks the region entirely; and in *D. simulans*, *Ur-Dox* lacks the start codon and *Dxl-1b* lacks the region entirely (**Fig. 4a** and **Extended Data Fig. 2**).

Text S4: *Tmy* and *Tmyl* structure and flanking sequences

The *Tmy* and *Tmyl* genes share the same basic structure: long inverted repeats consisting of nearly full-length *Dxl* sequence and some apparently unrelated sequence (**Fig. 4b**). The *Dxl*-derived sequence includes sequence from *CG8664* and most of the *ProtA/B*-related sequence, but no sequence derived from the 3'-UTR of *CG15306.* In *D. mauritiana*, the two halves of the inverted repeats are disrupted by deletions and single nucleotide changes, while *Tmy* in *D. simulans* and *Tmyl* in *D. sechellia* appear to be relatively intact. While *Tmy* and *Tmyl* are very similar in the structure of their *Dxl*-matching segments, they differ in the flanking material included in the inverted repeats. In *D. simulans*, the unrelated sequence flanking *Dxl*-matching sequence consists of a partial sat359 repeat and ~277bp matching the inverted repeats of the TEs *transib3* and *invader2*. In *D. mauritiana* and *D. sechellia Tmyl* genes, the "unrelated sequence" in the inverted repeats consists of two partial sat359 repeats; sequence matching genomic sequence X: 9628297-9628645 (adjacent to the sequence incorporated as "Ptpmeg2" inserts); and a small fragment (~138bp) matching the inverted repeats of the TE *roo*.

Text S5: Expression studies

Sequence set. The "base" sequence set used for all mapping studies contains a curated set of transcripts from the Flybase *D. simulans* (r.2.02) genome, plus a set of transposable elements derived from RepBase (v. 4.0.5), rDNA sequences from the Eickbush group (http://blogs.rochester.edu/EickbushLab/?page_id=602) and satellite sequence from the Larracuente group [54]. The transcripts (both coding and non-coding) were curated so that only a single transcript (generally the longest) was retained for each annotated gene. Transposable elements and repeats were informally curated to focus on a *D. simulans* clade-specific subset. We further identified and removed several transcripts with partial BLAST homology to a *Dxl* consensus sequence. To this base, we added the specific *Dxl*-related sequences of interest for the particular study, such as the autosomal suppressor hpRNA sequence for small RNAs or *Dxl* consensus sequence for gene expression, and proceeded with the mapping analysis (see **Source Data Files**).

Diagnostics for gene expression. We attempted to detect copy-specific expression of *Dxl* genes from high throughput sequencing using several approaches. The main approach mapped RNA-seq data to a species-appropriate *Dxl* consensus sequence and assessed evidence of expression at diagnostic SNPs (see **Source Data Files** and **Supplementary Table 1**). A diagnostic SNP was one for which a particular variant was unique to a single *Dxl* copy. To ensure that we were considering all possible sources of *Dxl*-matching sequence, we included both the species-appropriate autosomal suppressors and the X:17.1Mb and X:17.2Mb regions in this analysis. To enrich the mapping regardless of indel or gene structure in the different *Dxl* copies, we mapped reads as unpaired. Any *Dxl* gene with at least one SNP variant exhibiting both frequency ≥0.01 and ≥ 10 supporting reads in both replicates (*D. simulans* and *D. sechellia*) or meeting that standard in at least two of three replicates (*D. mauritiana*) was considered to show evidence of expression. In some cases, we considered short segments of reads, where some combination of variants was unique to a single *Dxl* gene. Finally, we inferred expression of several additional *Dxl* genes by assessing reads directly as part of the verification of the conserved splicing structure in the ORF region (see **Source Data Files** and **Supplementary Table 1**).

Supplementary Data Files

1. Alignment of inserted sat359 clusters in Fig. S3.

2. *Dxl* protamines - positions 1-66, aligned with known protamines
3. *D. simulans* r2.02, curated transcript set (“base” sequence set)
4. Consensus hairpin sequences of autosomal suppressors
5. Species-specific consensus sequences, all *Dxl*-related sequence

Supplementary Table 1. Gene-specific evidence for *Dxl*-related expression in testes

Species	Gene/location	Expressed?	Evidence
D. simulans	***MDox***	Y	SNP
D. simulans	***Dox***	Y	SNP
D. simulans	***Dxl-2***	Y	SNP
D. simulans	***Ur-Dox***	Y	SNP
D. simulans	***Dxl-1a***	Y	SNP
D. simulans	***Dxl-1b***	*no evidence*	
D. simulans	***Nmy***	Y	SNP
D. simulans	***Tmy***	Y	SNP
D. simulans	***X:17.1***	Y	SNP
D. simulans	***X:17.2***	Y	SNP
D. mauritiana	***MDox***	Y	SNP
D. mauritiana	***Dxl-14***	*no evidence*	
D. mauritiana	***Dxl-12***	*no evidence*	
D. mauritiana	***Dxl-11***	Y	SNP, splice junction
D. mauritiana	***Dxl-10***	*no evidence*	
D. mauritiana	***Dxl-9***	Y	splice junction
D. mauritiana	***Dxl-8***	Y	splice junction
D. mauritiana	***Dxl-5***	Y	splice junction
D. mauritiana	***Dxl-4***	Y	SNP, splice junction
D. mauritiana	***Dxl-3***	Y	SNP, splice junction
D. mauritiana	***Dxl-1***	*no evidence*	
D. mauritiana	***Ur-Dox***	*no evidence*	
D. mauritiana	***Nmy***	*no evidence*	
D. mauritiana	***Tmyl_left***	Y	segment
D. mauritiana	***Tmyl_right***	Y	SNP
D. mauritiana	***X:17.1***	*no evidence*	
D. mauritiana	***X:17.2***	*no evidence*	
D. sechellia	***Dxl-15***	Y	SNP
D. sechellia	***Dxl-14***	Y	SNP
D. sechellia	***Dxl-13***	Y	SNP, splice junction
D. sechellia	***Dxl-12***	Y	splice junction
D. sechellia	***Dxl-11***	Y	splice junction
D. sechellia	***Dxl-10***	*no evidence*	
D. sechellia	***Dxl-9***	Y	segment
D. sechellia	***Dxl-7***	Y	SNP
D. sechellia	***Dxl-6***	*no evidence*	
D. sechellia	***Dxl-4***	*no evidence*	
D. sechellia	***Dx1a***	*no evidence*	
D. sechellia	***Dxl-1b***	*no evidence*	
D. sechellia	***Ur_Dox_a***	Y	SNP
D. sechellia	***Ur_Dox_b***	*no evidence*	
D. sechellia	***Tmyl_left***	Y	SNP
D. sechellia	***Tmyl_right***	Y	segment
D. sechellia	***Emy_1***	Y	SNP
D. sechellia	***Emy_2***	Y	SNP
D. sechellia	***X:17.1***	Y	SNP
D. sechellia	***X:17.2***	Y	SNP

Supplementary Table 2. ***Dxl*** **expression in testes, males, females, and early-stage embryoes in *D. simulans, D. mauritiana,* and *D. sechellia***

		Replicate	SRR	Dxl_consensus	Act5C	protamine	CG13131
					FBtr0210252	FBtr0221891	FBtr0222221
D. simulans							
	Testes	1	SRR14777836	52.41	587.83	2054.15	67.62
	Testes	2	SRR14777837	49.76	554.07	2066.17	56.22
	Males	1	SRR9025061	18.37	138.42	507.73	9.98
	Males	2	SRR9025062	21.64	114.75	482.26	1.73
	Males	3	SRR9025063	21.76	114.02	464.43	1.97
	Females	1	SRR9025064	0.21	260.41	0.10	0.00
	Females	2	SRR9025057	0.27	254.88	0.00	0.02
	Females	3	SRR9025058	0.35	265.46	0.00	0.00
	Embryo (stage 2)	1	SRR6968152	0.07	355.43	0.00	0.00
	Embryo (stage 2)	2	SRR6968153	0.00	357.98	0.00	0.02
	Embryo (stage 2)	3	SRR6968154	0.00	329.82	0.00	0.00
	Embryo (stage 5)	1	SRR6968155	0.09	811.52	0.00	0.00
	Embryo (stage 5)	2	SRR6968156	0.11	760.40	0.00	0.00
	Embryo (stage 5)	3	SRR6968157	0.04	742.14	0.00	0.00
D. mauritiana							
	Testes	1	SRR9025052	34.67	301.64	1777.50	13.57
	Testes	2	SRR9025053	31.85	454.18	1519.54	15.62
	Testes	3	SRR9025054	25.08	479.53	1237.63	17.13
	Males	1	SRR9025056	8.54	92.40	189.72	1.44
	Males	2	SRR9025047	11.84	126.59	374.75	4.40
	Males	3	SRR9025048	14.28	99.47	359.60	2.13
	Females	1	SRR9025049	0.05	250.58	0.00	0.01
	Females	2	SRR9025050	0.22	238.80	0.08	0.00
	Females	3	SRR9025051	0.17	256.07	0.00	0.00
	Embryo (stage 2)	1	SRR6968106	0.15	570.73	0.00	0.00
	Embryo (stage 2)	2	SRR6968107	0.19	543.68	0.00	0.00
	Embryo (stage 2)	3	SRR6968108	0.00	493.40	0.00	0.00
	Embryo (stage 5)	1	SRR6968109	1.15	1197.85	0.00	0.02
	Embryo (stage 5)	2	SRR6968110	1.13	1043.66	0.00	0.00
	Embryo (stage 5)	3	SRR6968111	1.16	1144.28	0.00	0.00
D. sechellia							
	Testes	1	SRR14777834	83.02	459.12	1539.09	388.24
	Testes	2	SRR14777835	113.63	655.03	1955.44	536.85
	Males	1	SRR9025045	46.03	182.69	681.40	125.69
	Females	1	SRR9025046	0.96	269.26	0.08	0.03
	Embryo (stage 2)	1	SRR6968146	0.05	328.26	0.00	0.00
	Embryo (stage 2)	2	SRR6968147	0.06	394.38	0.00	0.00
	Embryo (stage 2)	3	SRR6968148	0.04	416.36	0.00	0.00
	Embryo (stage 5)	1	SRR6968149	0.17	748.43	0.00	0.00
	Embryo (stage 5)	2	SRR6968150	0.34	801.06	0.00	0.00
	Embryo (stage 5)	3	SRR6968151	0.21	885.21	0.00	0.00

Supplementary Table 3. Average matching percentage of esiRNAs (weighted by esiRNA abundance) to putative target genes for each species.

Species	hpRNA	Putative Target Genes													
		Dox	***Mdox***	***Dxl-2***	***Dxl-1a***	***Dxl-1b***	***Ur-Dox***								
D. simulans	*Tmy*	0.54	0.51	0.92	0.94	0.00	0.11								
	Nmy	0.87	0.93	0.75	0.75	NA	0.16								
		MDox	***Dxl-14***	***Dxl-12***	***Dxl-11***	***Dxl-10***	***Dxl-9***	***Dxl-8***	***Dxl-5***	***Dxl-4***	***Dxl-3***	***Dxl-1***	***Ur-Dox***		
D. mauritiana	*Tmyl*	0.51	0.53	0.50	0.91	0.52	0.92	0.91	0.90	0.87	0.87	0.00	0.86		
	Nmy	0.91	0.93	0.64	0.48	0.71	0.53	0.49	0.57	0.57	0.47	NA	0.47		
		Dxl-15	***Dxl-14***	***Dxl-13***	***Dxl-12***	***Dxl-11***	***Dxl-10***	***Dxl-9***	***Dxl-7***	***Dxl-6***	***Dxl-4***	***Dxl-1a***	***Dxl-1b***	***Ur-Dox_a***	***Ur-Dox b***
D. sechellia	*Tmyl*	0.61	0.82	0.82	0.73	0.74	0.87	0.82	0.82	0.54	0.87	0.87	NA	0.87	0.09
	Emy	0.98	0.97	0.97	0.97	0.97	0.79	0.97	0.97	0.58	0.79	0.79	NA	0.79	NA

Supplementary Table 4. Datasets included in analysis can be found in SRA (https://www.ncbi.nlm.nih.gov/sra)

Description	**Accessions**
D. mauritiana PacBio genome assembly	SRR5574088
D. simulans PacBio genome assembly	SRR5491305
D. sechellia PacBio genome assembly	SRR5514394
D. mauritiana RNA-seq	SRR9025052, SRR9025053, SRR9025054
D. simulans RNA-seq	SRR14777836, SRR14777837
D. sechellia RNA-seq	SRR14777834, SRR14777835
D. mauritiana small RNA-seq	SRR7961897
D. simulans small RNA-seq	SRR410589, SRR410590
D. sechellia small RNA-seq	SRR6667444