








Silencing RNAs expressed from W-linked *PxyMasc* “retrocopies” target that gene during female sex determination in *Plutella xylostella*

Tim Harvey-Samuel^a, Xuejiao Xu^{b,1}, Michelle A. E. Anderson^{a,4} , Leonela Z. Carabajal Paladino^{a,4} , Deepak Purusothaman^{a,2} , Victoria C. Norman^{a,3}, Christine M. Reitmayer^a , Minsheng You^b, and Luke Alphey^{a,4,5} 

Edited by Trudi Schüpbach, Princeton University, Princeton, NJ; received April 7, 2022; accepted August 29, 2022

The Lepidoptera are an insect order of cultural, economic, and environmental importance, representing ~10% of all described living species. Yet, for all but one of these species (silkworm, *Bombyx mori*), the molecular genetics of how sexual fate is determined remains unknown. We investigated this in the diamondback moth (*Plutella xylostella*), a globally important, highly invasive, and economically damaging pest of cruciferous crops. Our previous work uncovered a regulator of male sex determination in *P. xylostella*—*PxyMasc*, a homolog of *B. mori* *Masculinizer*—which, although initially expressed in embryos of both sexes, is then reduced in female embryos, leading to female-specific splicing of *doublesex*. Here, through sequencing small RNA libraries generated from early embryos and sexed larval pools, we identified a variety of small silencing RNAs (predominantly Piwi-interacting RNAs [piRNAs]) complementary to *PxyMasc*, whose temporal expression correlated with the reduction in *PxyMasc* transcript observed previously in females. Analysis of these small RNAs showed that they are expressed from tandemly arranged, multicopy arrays found exclusively on the W (female-specific) chromosome, which we term “*Pxyfem*”. Analysis of the *Pxyfem* sequences showed that they are partial complementary DNAs (cDNAs) of *PxyMasc* messenger RNA (mRNA) transcripts, likely integrated into transposable element graveyards by the noncanonical action of retrotransposons (retrocopies), and that their apparent similarity to *B. mori* *feminizer* more probably represents convergent evolution. Our study helps elucidate the sex determination cascade in this globally important pest and highlights the “shortcuts” that retrotransposition events can facilitate in the evolution of complex molecular cascades, including sex determination.

Insects have evolved a diverse range of mechanisms to determine sex, which vary widely both between and within orders (1). Even within groups where the broad pattern of sex determination is similar (e.g., XY males/XX females), this can be achieved through nonhomologous molecular mechanisms, for example, X-linked gene dosage or dominant male-determining loci (2). The study of these mechanisms is of fundamental interest but also provides tools and targets for genetic pest management strategies, including gene drives (3, 4).

Much of what is known regarding insect sex-determination systems is based on examples from a few orders of economic or human-health importance, primarily the Diptera and Hymenoptera (2). The Lepidoptera, which contain ~10% of all described living species, including many of economic, cultural, and environmental importance (5), are understudied in this regard—only the domestic silkworm (*Bombyx mori*) has had the molecular mechanism by which it determines sex uncovered (6). All lepidopterans exhibit female heterogamety, with most species, including *B. mori*, showing a WZ/ZZ female/male sex-chromosome complement (though the presence of the female-specific W chromosome is derived, with the ancestral state being ZO/ZZ—a trait still observed in some early-diverging moth lineages) (7). Recent work identified the master regulator of sex in *B. mori* to be a single Piwi-interacting RNA (piRNA) (known as *feminizer* [*fem*]) complementary to the Z-linked male determining gene *Masculinizer* (6). This 29-base-pair (bp) “*fem*” piRNA exists in multiple copies, exclusively on the female-specific W chromosome. In the presence of the W chromosome (i.e., in females), *Masculinizer* messenger RNA (mRNA) is silenced by *fem*, allowing the female sex-determination cascade to engage. In the absence of the W chromosome (i.e., in ZZ males), *Masculinizer* is translated, initiating the male cascade (6).

Being an example of an RNA interference (RNAi)-controlled sex-determination cascade, this study was of great significance. However, questions regarding its generality

Significance

Uncovering how species determine sex is of fundamental interest and provides avenues for genetic pest management. Much of what is known regarding insect sex determination is concentrated within the Diptera and Hymenoptera, despite other orders (e.g., Lepidoptera) being of great ecological and economic importance. Here, using small RNA sequencing of embryonic and early larval samples, we uncover an RNA interference (RNAi)-based sex determination system that silences the male-determining gene *PxyMasc* in the diamondback moth (*Plutella xylostella*). We track these silencing RNAs back to the W chromosome, where they are expressed from partial complementary DNA (cDNA) copies of *PxyMasc*. Analyses suggest these are *PxyMasc* “retrocopies”, integrated via long terminal repeat (LTR) retrotransposons, and that similarities between this system and the *feminizer* system in *Bombyx mori* represent convergent evolution.

This article is a PNAS Direct Submission.

Copyright © 2022 the Author(s). Published by PNAS. This open access article is distributed under [Creative Commons Attribution License 4.0 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/).

¹Present address: School of Life Sciences, Peking University, Beijing 100871, China.

²Present address: MRC-University of Glasgow Centre for Virus Research, University of Glasgow, Glasgow G12 8QQ, United Kingdom.

³Present address: Oxitec Ltd., Abingdon OX14 4RQ, United Kingdom.

⁴Present address: Biology Department, University of York, YO10 5DD, York, United Kingdom.

⁵To whom correspondence may be addressed. Email: luke.alphey@york.ac.uk.

This article contains supporting information online at <http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2206025119/-DCSupplemental>.

Published November 7, 2022.

remain. For example, the W chromosome is known to have evolved independently at least twice within the Lepidoptera (ZO/ZZ being the ancestral arrangement) with multiple examples of neo-Z fusions in the upper ditrysian Lepidoptera (8), suggesting that the true No. of independent W-generating events may be higher (9, 10). It thus remains unclear to what extent the W-linked *feminizer* system uncovered in *B. mori* is conserved, especially given this species' long history of domestication (~5,000 y of artificial rearing) and the fact that at least one wild silkworm species (*Samia cynthia*), although possessing a W chromosome, does not require it for female sex determination (11). Furthermore, although the function and Z linkage of *Masculinizer* appear to be deeply conserved, examples identified so far show that the amino acid sequence of this gene has diverged significantly over evolutionary time (12), to the extent that it is unclear how a multicopy, nucleotide-specific silencing RNA would have been able to maintain sequence complementarity. Indeed, the nucleotide sequence targeted by *fem* is not conserved, even in the closely related *Trilocho varians* *Masculinizer* homolog (13). Finally, to our knowledge, no hypotheses have been proposed for how such a unique mechanism of sex determination may have initially evolved.

We chose to explore these questions in the diamondback moth *Plutella xylostella*. In addition to being one of the world's most damaging agricultural pests (14–16) and one in which genetic pest-management strategies are being developed (17–20), *P. xylostella* possesses characteristics that are beneficial in addressing these questions. It is a member of the superfamily Yponomeutoidea, an early-diverging Ditrysian lineage evolutionarily distant from the more-derived Bombycoidea (which includes *B. mori*) and thus important in exploring early lepidopteran evolution (9). It possesses the ancestral lepidopteran chromosome No. ($n = 31$), with no large-scale chromosome rearrangements and a well-differentiated WZ/ZZ sex-chromosome system (9, 21, 22). Finally, it has been subjected to an unusually high degree of molecular genetic research for a lepidopteran, with publicly available male- and female-derived draft genome sequences (21, 23), restriction site-associated DNA-based chromosome-level maps (24), extensive karyotyping analysis, and crucially, a recently identified and characterized *Masculinizer* homolog (*PxyMasc*) (25).

Here, using the *PxyMasc* mRNA sequence as a guide, we identified a functionally consistent yet likely independently derived *fem*-like system in *P. xylostella*. We observed a range of small RNAs, expressed from early embryonic stages onwards, mapping to both the sense and antisense strands of *PxyMasc* exons four, five, and six, including size ranges corresponding to both small interfering RNAs (siRNAs)/microRNAs (miRNAs) and piRNAs. Rapid amplification of complementary DNA (cDNA) ends (RACE) conducted against these antisense piRNAs identified transcripts consisting of contiguous reverse-complement *PxyMasc* and truncated retrotransposon open reading frame (ORF) sequences. These putative *P. xylostella fem* ("*Pxyfem*") precursors mapped to multicopy loci within "transposable element graveyards" found exclusively in a female genome assembly and which genomic PCR confirmed were female specific (i.e., W linked). Intriguingly, analysis of the genomic context of these sequences suggests that *Pxyfem* evolved through a template-switching event involving long terminal repeat (LTR) retrotransposons and a *PxyMasc* mRNA transcript, followed by retroposition of the chimeric RNA onto the W chromosome. Analysis of the "fossilized" W-linked LTR remnants and the *Pxyfem* sequences themselves suggests that this event took place after the divergence of *P. xylostella* from other ditrysians whose genomes have been sequenced, suggesting independent

evolution of this female *Masculinizer* regulator relative to that of *B. mori*.

Results and Discussion

A Range of Small RNAs Map to *PxyMasc*. If a system similar to *B. mori fem* existed in *P. xylostella*, we would expect small RNA libraries (reads of 18–40 nucleotides in length) to produce sequences mapping to the antisense strand of *PxyMasc* in female-derived, but not male-derived, pools. Out of 3,328,300 female L1 larvae-derived small RNA reads, 38 were found to map to the *PxyMasc* mRNA sequence, with 12 of these mapping to the sense and 26 to the antisense strand. For the male L1 larvae-derived RNA, of 2,369,543 reads, only two mapped to *PxyMasc*, with neither of these mapping to the antisense strand (Fig. 1).

For the female L1-derived library, the majority of reads mapped to *PxyMasc* exons 5 and 6, with a particular concentration of reads mapping to the antisense strand of the exon 5–6 junction (Fig. 1C), overlapping with the highly conserved cysteine–cysteine domain—a hallmark of identified *Masculinizer* homologs and an area required for functionality in *B. mori Masc* (26). The two male reads mapped to independent sequences within the putative 3' untranslated region (UTR) of *PxyMasc*. These two reads were not present in the female-derived library (Fig. 1D).

Interestingly, unlike the *B. mori fem* transcript, which produces a single, 29-bp piRNA, in the female L1-derived sample, we observed multiple different *PxyMasc*-mapping reads whose size ranges corresponded to both siRNA/miRNAs (~21–24 bp) and piRNAs (~25–30 bp), although those belonging to the "piRNA" size class predominated (Fig. 1A). We did not observe the classic "ping-pong" signal associated with these reads (3' overlap of 10 bp between forward and reverse reads); however, they did display a bias toward a 5' end uridine (1U bias) in antisense strand-mapping reads, a signal of piRNA biogenesis (27) (Fig. 1G).

Subsequently, we set out to establish whether these *PxyMasc*-mapping reads were present in early embryos, as would be expected of a *fem*-like sex-determination system. Previously, we established that sex in *P. xylostella* is determined in the early embryo (6–24 h after laying) by assessing the temporal expression pattern of *PxyMasc* and *doublesex* during early embryos (25). There, *PxyMasc* mRNA was not observed at early time points (3 h postoviposition). However, by 6 h postoviposition, *PxyMasc* mRNA was present in both sexes and being translated, judging by observed male-form splicing of *doublesex* (a function of *PxyMasc* or its downstream cascade). By 24 h postoviposition, *PxyMasc* mRNA was no longer detectable in female embryos but remained present in males. This implied that the silencing of *PxyMasc* by any putative *fem*-like system would likely begin and peak during the 6- to 24-h period postoviposition. To assess this, we sequenced pooled embryonic small RNA libraries at 3, 6, 9, 12, and 24 h postoviposition (Fig. 2). As expected, the 3-h library did not contain *PxyMasc*-mapping reads. However, from 6 h onwards, such reads were observed, gradually increasing to the 12-h time point before dropping away by 24 h (*SI Appendix, Table 1* for total and normalized read Nos.). The position of these reads generally concurred with those observed in the female L1-derived sample, and again, read lengths corresponded both to siRNA/miRNAs as well as piRNAs. As with the female L1 sample, we observed a 1U bias in antisense-mapping reads only. However, we now also observed a 10A bias in sense-mapping reads only (see Fig. 2 for demonstration of this in the most-common size class of read in each time point). Additionally,

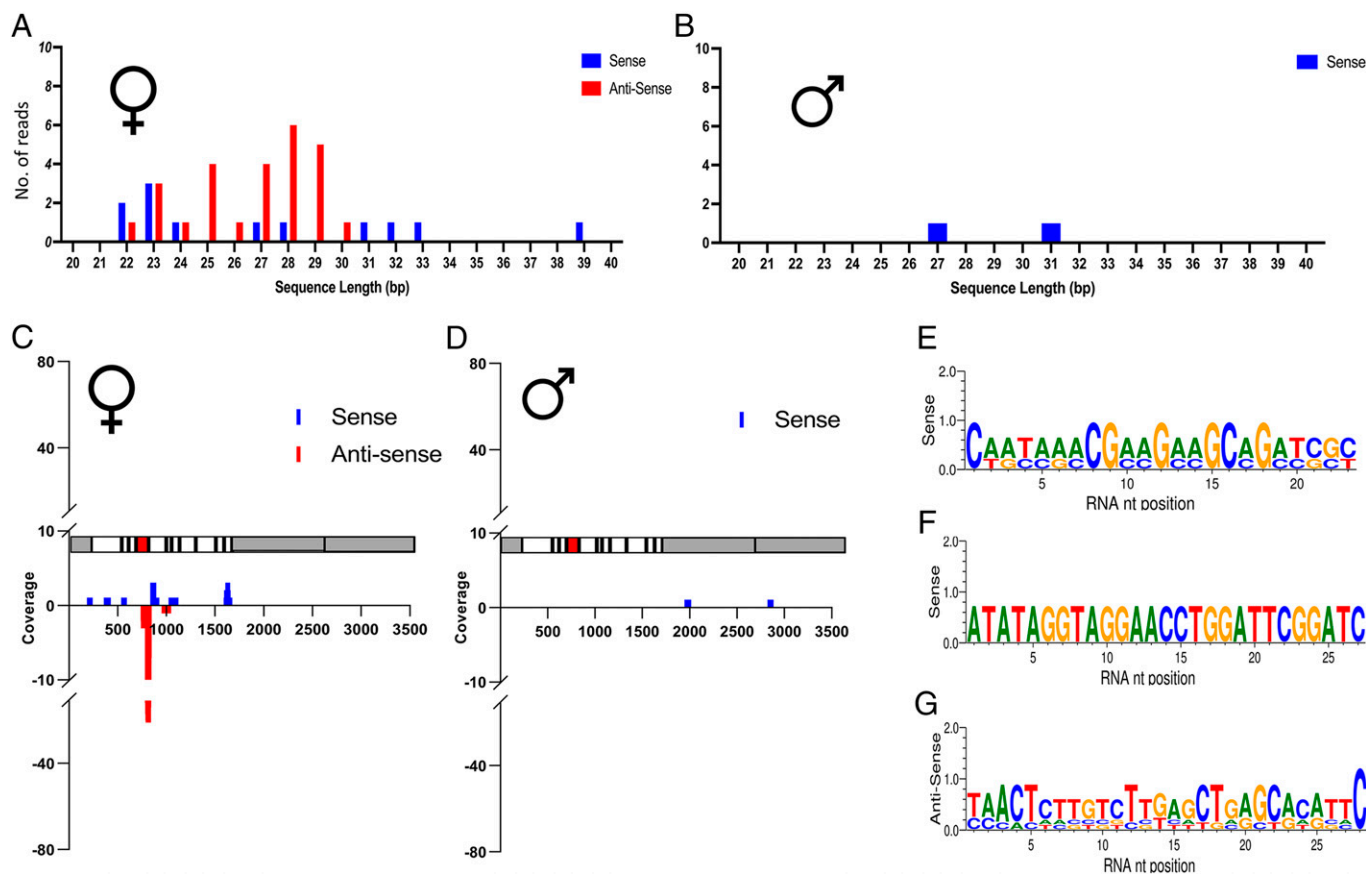


Fig. 1. Small RNA deep sequencing of female or male pooled L1 samples. (A, C, E, and G) Data deriving from L1 female pools. (B, D, and F) Data deriving from L1 male pools. (A and B) No., size, and orientation of reads mapping to the *PxyMasc* mRNA sequence. (C and D) Location along *PxyMasc* mRNA sequence of mapped reads. A schematic of the *PxyMasc* mRNA is shown parallel with each x axis to show features against which these reads map. Gray areas designate UTRs, white areas designate coding exons, and the red shaded area designates the region of the transcript coding for the cysteine–cysteine motif—a domain functionally requisite in *Bombyx mori* Masc and conserved among other lepidopteran Masc homologs identified to date. (E, F, and G) Sequence logos for the most common length of sense and antisense reads identified in the two pools. Note: the male pool did not provide any antisense reads mapping to *PxyMasc*.

sense and antisense reads showing a 3′ 10-bp overlap ping-pong signal (28) were apparent—see Fig. 3 for closer examination of this in the 12-h sample. Taken together, these data indicate a ping-pong amplification cycle of piRNAs targeting *PxyMasc* early in *P. xylostella* embryonic development. Interestingly, secondary concentrations of sense and antisense reads mapped down and upstream, respectively, of the primary exon 5–6 junction concentration (see Fig. 3), which may be indicative of a “phased” piRNA biogenesis (29, 30)—where the primary transcript these piRNAs derived from is serially processed into downstream piRNAs after the initial cleavage event. Further interrogation of the *Zucchini*-dependent processing pathway would be required to verify this.

These trends concur with the observed patterns of *PxyMasc* and sex-specific *doublesex* mRNA presence in males and females described previously (25) and suggest a potential role for these small silencing RNAs (ssRNAs) in mediating that pattern.

Antisense ssRNAs Derive from Transcripts Associated with the W Chromosome. To identify the primary transcripts from which these *PxyMasc*-targeting ssRNAs were derived, 5′ and 3′ RACE gene-specific primers were designed in a region of exon 5 where the largest concentration of those reads were observed. Using gene-specific primers matching the sense sequence of *PxyMasc* for 5′ RACE (and the antisense sequence of *PxyMasc* for 3′ RACE) prevented the unintentional amplification of the

endogenous *PxyMasc* transcript. One 5′ and one 3′ transcript were amplified from ovarian RACE-ready cDNA. The 5′ RACE transcript consisted of a cDNA copy of *PxyMasc* extending upstream from the primer binding point (in exon 5) into exon 6 (i.e., a partial reverse complement [RC] of *PxyMasc* spliced mRNA). BLASTn analysis of the remainder of the transcript revealed a 490-bp sequence (henceforth multiply repeated region 1) showing high similarity (highest = 97%) to multiple (>80) hits in both the male-derived *P. xylostella* genome assembly (PRJNA277936) and female-derived *P. xylostella* genome assembly (PRJEB34571) (Fig. 4A). However, the full RACE transcript sequence was found only in the female-derived assembly. Similarly, the 3′ RACE product consisted of RC-*PxyMasc* exons 5 and 4 (Fig. 4B). The full-length RACE transcript sequence could again only be found in the female-derived assembly, whereas a shorter 513-bp region of the transcript showed high similarity (99.4%) to multiple repeated regions in both the male- and female-derived genome assemblies (henceforth multiply repeated region 2).

We hypothesized that, in both cases, the full RACE transcripts were identified only in the female-derived assembly as they were located on (expressed from) the W chromosome, whereas the multiply repeated regions may represent sections of transposable elements present on the W chromosome but also elsewhere in autosomal regions (and therefore also in the male-derived assembly). To test this, we performed a series of genomic PCRs across the 5′ RACE transcript sequence (Fig. 4A).

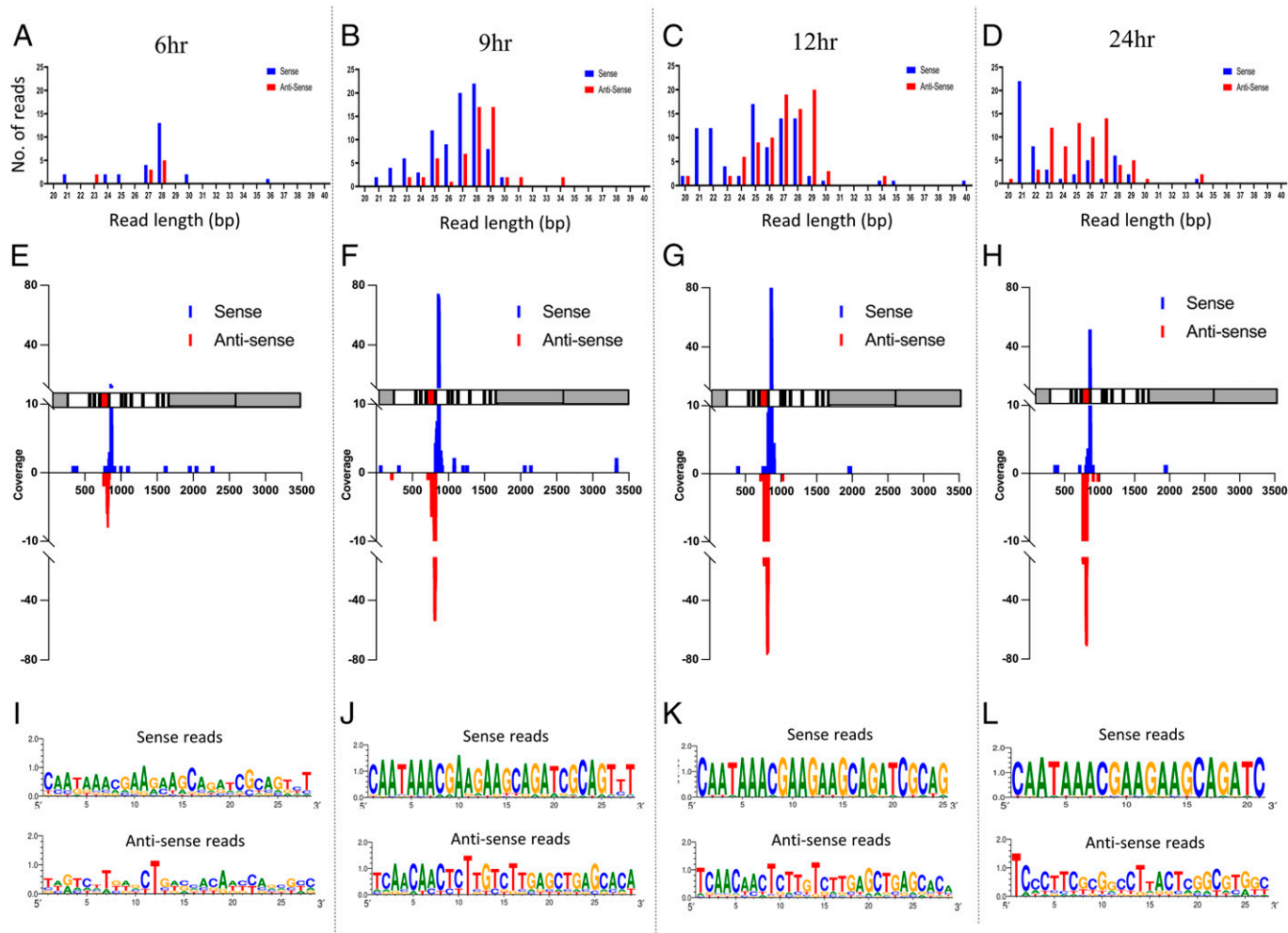


Fig. 2. Small RNA deep sequencing of 6, 9, 12, and 24 h pooled embryo samples. (A, E, and I) Data deriving from 6-h pooled embryos. (B, F, and J) Data deriving from 9-h pooled embryos. (C, G, and K) Data deriving from 12h pooled embryos. (D, H, and L) Data deriving from 24-h pooled embryos. For all time points, the No. of matching reads was normalized for total read No. derived from that pooled sample. (A–D) No., size, and orientation of reads mapping to the *PxyMasc* mRNA sequence. (E–H) Location along *PxyMasc* mRNA sequence of mapped reads. A schematic of the *PxyMasc* mRNA is shown parallel with each x axis to show features against which these reads map. Gray areas designate UTRs, white areas designate coding exons, and the red shaded area designates the region of the transcript coding for the cysteine–cysteine motif—a domain functionally requisite in *Bombyx mori* Masc and conserved among other lepidopteran Masc homologs identified to date. (I–L) Sequence logos for the most common length of sense and antisense reads identified in the two pools.

PCR amplicons were produced within the multiply repeated region 1 region when both male and female genomic DNA (gDNA) samples were used as templates. However, amplicons across the multiply repeated region 1 (MRR1)/*PxyMasc* fragment junction were only produced when female gDNA was used, implying that these regions are exclusive to the female genome and therefore W linked (Fig. 4C).

***Pxyfem* Occurs in Two Multicopy Tandemly Arranged Clusters within Transposable Element Graveyards.** BLASTn analysis showed that sequences highly similar to the full-length RACE transcripts mapped exclusively to a female-derived genomic scaffold (CABWKK01000004) with no matches in the male-derived genome assembly. Analysis of this scaffold uncovered 11 loci with sequences similar to the *PxyMasc* coding sequence (CDS) (henceforth *Pxyfem* loci) in two divergently orientated clusters (Fig. 5A). While all 11 *Pxyfem* loci included *PxyMasc* exons (partial) 4, 5, and 6, seven of these continued partway into exon 7 (henceforth “long” *Pxyfem* = 326 bp), while four ended partway through exon 6 (henceforth “short” *Pxyfem* = 192 bp) (Fig. 5B). These *Pxyfem* loci showed very high nucleotide sequence conservation both to *PxyMasc* (average similarity to *PxyMasc* = 93.46% ± 0.26%) and to each other (92.7% sequence conservation between

all loci), calculated over the core 192-bp region shared by both short and long *Pxyfem*. RT-PCR using ovarian-derived cDNA and primers which orientated “outwards” from each *Pxyfem* locus showed that at least some closely linked *Pxyfem* loci are expressed on the same transcript (SI Appendix, Fig. 2—sequencing of PCR bands and further analysis in SI Appendix, Fig. 3).

The structure of the wider regions surrounding each *Pxyfem* locus varied dramatically (Fig. 5B). These areas contained a range of partial ORFs similar to transposable elements as well as tandemly repeated elements that lacked discernible ORFs. Where present, these ORFs were orientated in the same direction as the *Pxyfem* loci direction of transcription (as deduced by RACE). Small RNA reads from the 12-h embryo timepoint were found to map widely across the entire area of a searched ~22-kbp genomic fragment that included *Pxyfem* loci 1–4 (SI Appendix, Fig. 1), though the density of antisense ssRNA reads mapping to this fragment increased dramatically in and around the four *Pxyfem* loci. This concurs with the situation in *B. mori*, where the highly repetitive W chromosome harbors large Nos. of transposable element graveyards, where repeated and truncated copies of transposon ORFs occur in a nested arrangement (6).

In contrast, the sequences immediately flanking the *Pxyfem* loci showed much higher conservation (Fig. 5B). At the 3′ end

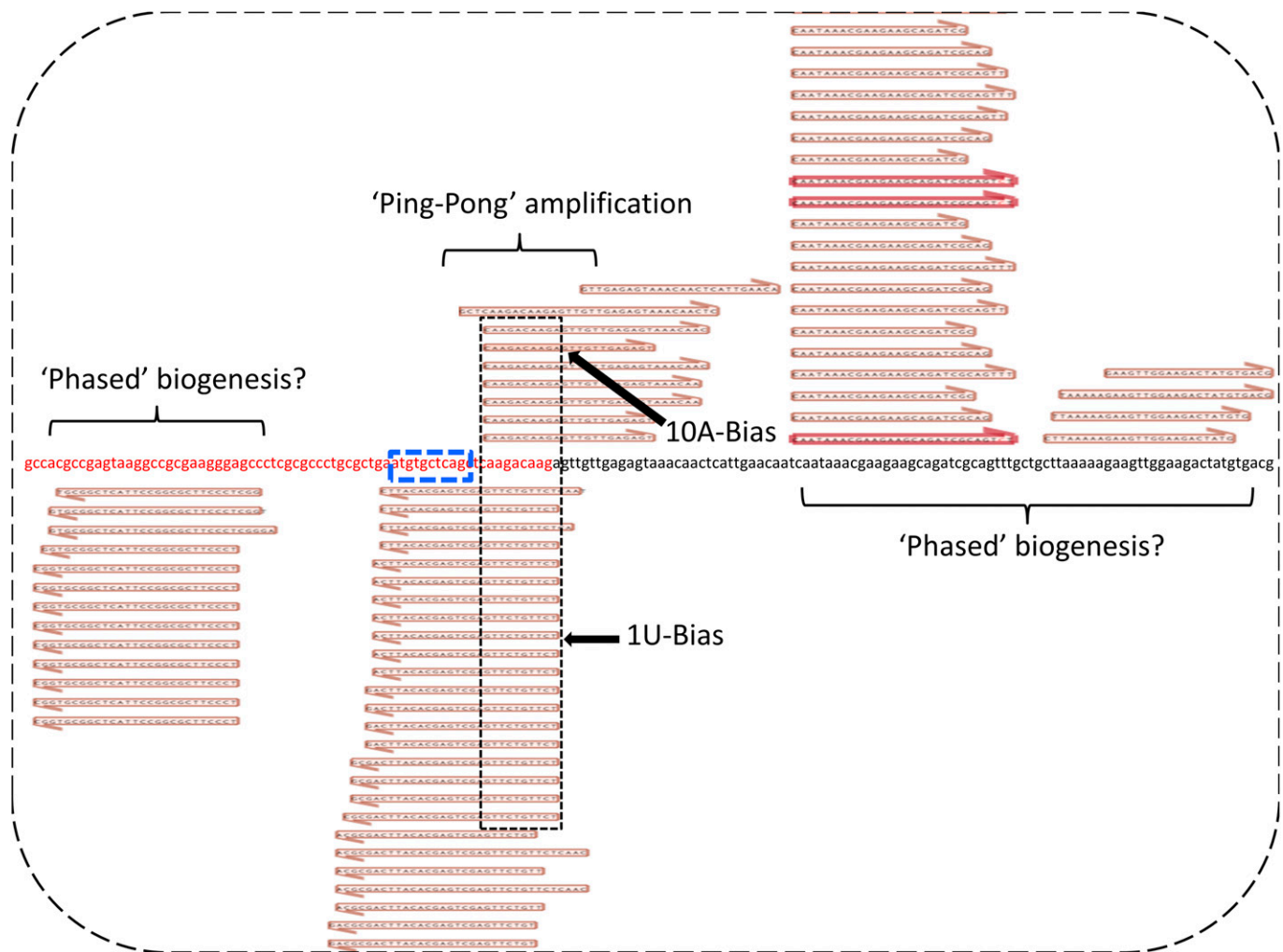


Fig. 3. Ping-pong amplification signal evident in pooled 12-h embryo *PxyMasc* mapping reads. Central nucleotide sequence is of the *PxyMasc* mRNA exon 5 (shown in red lettering) and exon 6 (shown in black lettering) junction. Blue dashed box designates the nucleotide sequence translated into the cysteine-cysteine domain. Sense and antisense reads mapping to this area show characteristic signs of the “ping-pong” amplification loop, overlapping on their 3' ends by 10 bp with a 1-U bias for antisense reads and 10-A bias for sense reads. Reads mapping away from this central area showed a strong bias for mapping in the same direction as those in the ping-pong cycle, e.g., reads shown mapping in exon 5 were all in an antisense orientation. This may be indicative of a “phased” biogenesis for these small silencing RNAs.

of each *Pxyfem* (i.e., abutting truncated *PxyMasc* exon 4), a 251-bp sequence was common to all 11 copies but absent elsewhere on this scaffold. These sequences contained an ORF that was aligned to generate an 82-amino acid (aa) consensus sequence for BLAST_p analysis. The consensus ORF showed closest similarity to the group-specific antigen (gag) polyproteins of LTR retrotransposons present in several lepidopteran families. The most similar hit (LOC105396011 with 51% aa identity over aligned region) and the highest No. of significant hits ($n = 13$) were found in the *P. xylostella* male-derived assembly. The female-derived assembly was deliberately excluded from analysis to prevent the confounding effects of identifying *Pxyfem*-linked loci; hits in the male-derived assembly must be autosomal or Z linked. Henceforth, we refer to *Pxyfem*-linked copies of this LTR retrotransposon as PLTR1, while the putative homologous endogenous gene we refer to as endogenous LTR1 (ELTR1).

Similarly, immediately flanking the 5' end of each short *Pxyfem* was a conserved ~256-bp sequence (truncated to 108 bp in one copy) whose translated ORF showed highest similarity (59.49% identity) to the gag polyprotein of a second LTR retrotransposon (LOC105389072) from the male-derived *P. xylostella* genome (henceforth *Pxyfem*-linked LTR2 = PLTR2,

putative endogenous homolog = ELTR2). Long *Pxyfem* also displayed an immediately adjacent conserved sequence, but this was much shorter (34 bp) and could not be identified. None of these conserved sequences were found elsewhere on this scaffold.

A phylogeny built using Clustal Omega (<https://www.ebi.ac.uk/Tools/msa/clustalo/>) that included each of the 11 *Pxyfem* copies as well as ~250 bp of up- and downstream genomic flanking sequence showed that these loci fell into two groups that could be differentiated by the presence of LTR2/absence of *PxyMasc* exon 7 or vice versa (Fig. 5C). Interestingly, while *Pxyfem* copies within the same genomic cluster matched with each other in terms of their orientation, they did not necessarily group together within the phylogeny.

***Pxyfem* and *fem* Are Functionally Consistent.** In *B. mori*, the *fem* piRNA targets *Masculinizer* transcripts for cleavage during embryogenesis (6). To assess whether *Pxyfem* plays a similar role in *P. xylostella*, we utilized a single-transcript, positive-readout system for identifying site-specific mRNA cleavage (31). This system uses two bacterial hairpins (CopT and CopA) placed in the 5' UTR and 3' UTR (respectively) of a luciferase expression construct, with an artificial polyA tail and downstream

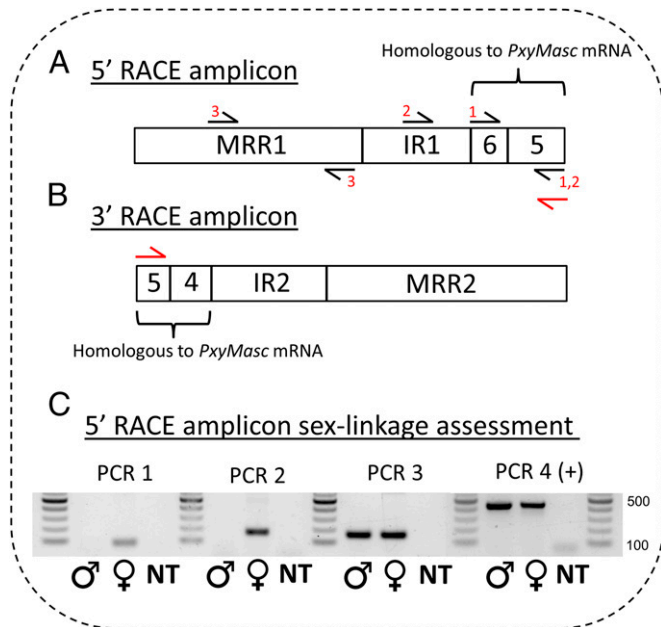


Fig. 4. Schematic of observed RACE transcripts and PCR assessment of RACE amplicon sex linkage. (A) Schematic of 5' RACE amplicon observed when using a gene-specific primer in the sense orientation to *PxyMasc* exon 5 (shown by red primer symbol). MRR1 = multiply repeated region 1; IR1 = intervening region 1; 5 and 6 = sequence homologous to *PxyMasc* exons 5 and 6. Black primer symbols and respective Nos. listed show the approximate locations and combinations of primers used in PCRs listed in C. (B) Schematic of 3' RACE amplicon observed when using a gene-specific primer in the antisense orientation to *PxyMasc* exon 5 (shown by red primer symbol). MRR2 = multiply repeated region 2; IR2 = intervening region 2; 4 and 5 = sequence homologous to *PxyMasc* exons 4 and 5. (C) PCR assessment of sex linkage of the genomic locus from which the 5' RACE amplicon was expressed. The same No. above wells denotes a consistent primer set used in that PCR. Position of primer sets 1, 2, and 3 shown in A. + control (PCR 4 +) shows amplification of the 17S gene. NT signifies a no-template (H_2O) control.

cleavage site placed upstream of CopA (*SI Appendix, Fig. 4A*). In the absence of mRNA cleavage, the two hairpins function to partially suppress translation of the synthetic mRNA. If the target site is cleaved (e.g., by a ssRNA), the CopA hairpin is released from the transcript, exposing the synthetic polyA tail and allowing more efficient translation. Cleavage of a target mRNA sequence can thus be identified by changes in luciferase readout between treatments.

Here, we adapted this design by replacing the miRNA targeted locus immediately downstream up the polyA tail with the ~325-bp section of *PxyMasc* against which *Pxyfem* matches—giving construct AGG2208 (*SI Appendix, Fig. 4*). We hypothesized that, in the presence of endogenous ssRNAs targeting this region (i.e., in female *P. xylostella*), mRNA transcripts expressed by AGG2208 would be cut, allowing for increased translation of nanoluciferase, and that this would be significantly reduced where these ssRNAs were absent (i.e., in male *P. xylostella*).

Prior to utilizing this system in *P. xylostella* embryos, however, we chose to test the AGG2208 construct in *Aedes aegypti* Aag2 cells to assess whether it functioned as expected. When cotransfected with a 'treatment' double-stranded RNA (dsRNA) targeting the *PxyMasc* section immediately downstream of the polyA tract, normalized levels of nanoluciferase increased significantly relative to when a "negative control" dsRNA (targeting a section of *AmCyan* CDS) was cotransfected (dsRNA = 2 ng/well, $t(14) = 12.96$, $P < 0.0001$; dsRNA = 50 ng/well, $t(13) = 3.629$, $P = 0.0031$). This demonstrated that cleavage of the AGG2208

transcript was sufficient to increase measured nanoluciferase output (*SI Appendix, Fig. 5*).

After this initial demonstration, we next used the AGG2208 construct to assess whether this section of *PxyMasc* against which *Pxyfem* matched was sufficient to label an mRNA transcript for

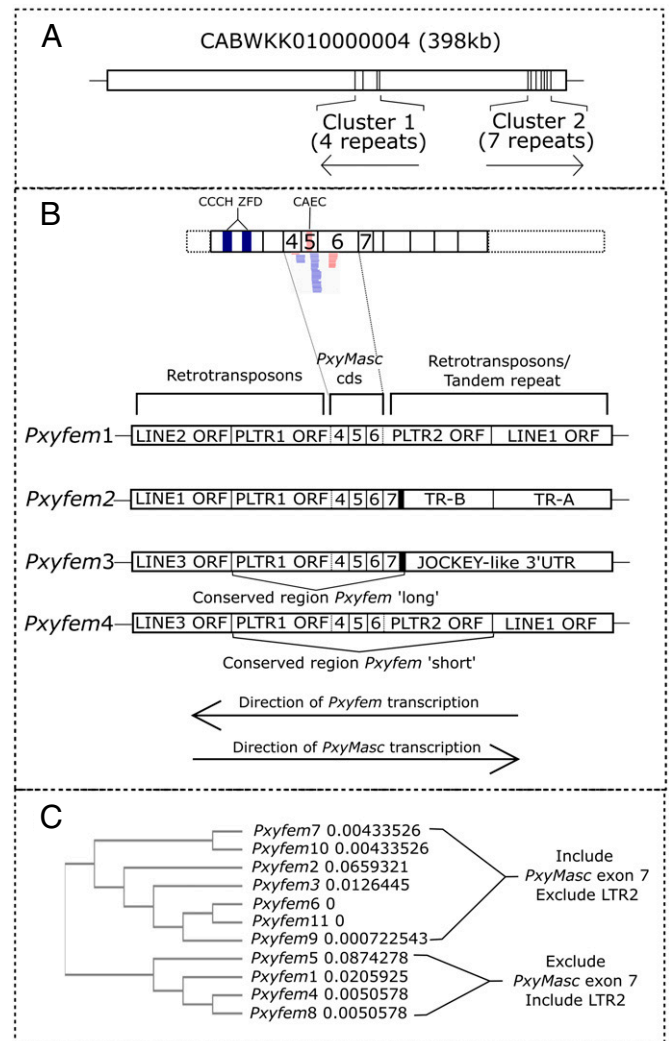


Fig. 5. RACE transcripts are derived from genomic loci nested within transposable element graveyards. (A) Full-length 5' and 3' RACE amplicon sequences mapped exclusively to a scaffold from the female-derived *P. xylostella* genome assembly (scaffold CABWKK010000004 from genome project PRJEB34571). These "*Pxyfem*" loci/copies occurred in two, differentially orientated clusters within each of which copies were tandemly arranged. Orientation of the copies shown by arrows under each cluster. (B) Schematic showing detail of the *Pxyfem* loci included in cluster 1 as well as C. Five hundred base pairs of up- and downstream genomic flanking sequence. Identity of truncated genes within these flanking sequences were elucidated through ORF analysis and BLASTp or BLASTn of the nucleotide sequence if no homologous ORFs were observed. Schematic above these four copies is of the full *PxyMasc* mRNA in its sense orientation, included to show the sequence homologous to this mRNA present in each *Pxyfem* copy. CCH ZFD = the N terminus zinc-finger domains characteristic of identified *Masculinizer* homologs; CAEC = the "masculinizing" domain characteristic of *Masculinizer* homologs. Lines cross over as *Pxyfem* copies are in the inverse orientation to the endogenous *PxyMasc* transcript. LINE = long interspersed element; LTR = long terminal repeat; TR = tandem repeat. Black box adjacent to exon 7 in *Pxyfem* long represents conserved 34-bp sequence of unknown identification. Direction of *Pxyfem* transcription taken from observed RACE products. Direction of endogenous gene transcription taken from the ORF orientation of *PxyMasc* homologous sequence and surrounding identified ORFs. (C) Phylogeny constructed using Clustal Omega including all 11 identified *Pxyfem* copies along with 250 bp of up- and downstream genomic flanking sequence. Sequence distance measures computed by Clustal are given as Nos. behind each terminal branch.

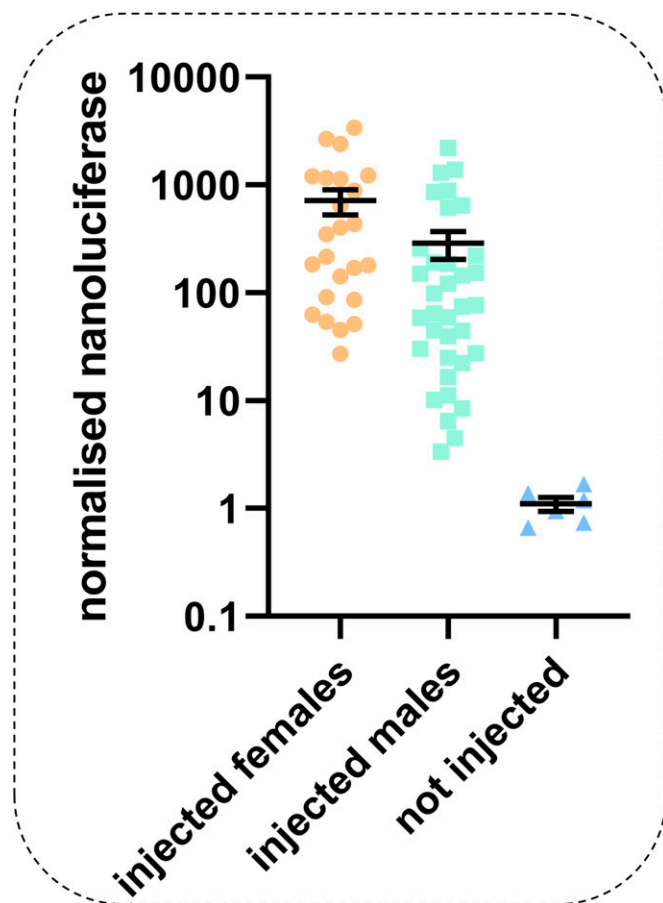


Fig. 6. Female embryos display higher levels of cleavage-induced positive readout signal than male embryos. x axis gives treatment, and y axis shows nanoluciferase readings in individual embryos (circles represent a single replicate embryo) normalized against firefly luciferase reading from the same embryo. Injected female embryos showed significantly higher levels of normalized nanoluciferase than injected males ($P = 0.0237$, $t = 2.325$, $df = 57$, $F = 24$, $M = 35$, error bars = 95% CI), aligning with the hypothesis that cleavage of transcripts in this sex by *Pxyfem*-derived ssRNAs increased subsequent translation.

cleavage in injected *P. xylostella* embryos and whether the sex of those embryos influenced this behavior. We observed a significant effect of embryo sex, with females showing a higher nanoluciferase output than males ($P = 0.0237$, $t = 2.325$, $df = 57$, Fig. 6). This result indicates that the presence of the W chromosome (and thus the *Pxyfem* loci/expressed ssRNAs) is sufficient to cause the cleavage of transcripts bearing the *Pxyfem*-targeting *PxyMasc* sequence. This observation, paired with our previous result showing that these *PxyMasc*-targeting ssRNAs are expressed in the early embryo at a time consistent with sex determination, strongly suggests the involvement of the *Pxyfem* system in determining female sexual fate in *P. xylostella* through silencing *PxyMasc*.

Origins of *Pxyfem*. The location of the *Pxyfem* loci on a large W-linked scaffold allows us to explore the possible origins of this system. Two interlinked questions include how and when such a mechanism might have arisen.

How did *Pxyfem* arise? *Pxyfem* loci consist of partial *PxyMasc* cDNAs extending over several exon–exon boundaries (i.e., excluding three *PxyMasc* introns). This implies that the initial *Pxyfem* W-linkage event involved a spliced *PxyMasc* mRNA intermediate rather than a duplication of the Z-linked *PxyMasc* gene itself. In contrast, the 29-bp *fem* in *B. mori* does not extend outside of *Masculinizer* exon 9, and thus, no such conclusion can be drawn there.

Both *Pxyfem* and *fem* are in wider nested transposable and repeat element graveyards. LTR retrotransposons can drive the duplication of endogenous genes through a process known as retrotransposition (32, 33). These LTR retrotransposons increase their frequency through a “copy and paste” mechanism, in which their mRNA is reverse transcribed into cDNA, made double stranded, and then integrated elsewhere in the genome, forming a second copy of the original element (34). However, analysis in a diverse range of eukaryotes has found that mRNA template-switching events may occur during the reverse transcription step, allowing chimeric LTR retrotransposons to be generated (35). If the nonretrotransposon mRNA is derived from an endogenous gene, this may result in the partial duplication of that sequence, a “retrocopy”, if integration is successful. That on the W-linked scaffold, the PLTR1 retrotransposon fragment was only found associated with *Pxyfem*, and its tight linkage to all 11 copies, suggests a role for an ancestral ELTR1 in the generation of *Pxyfem*. In *Drosophila melanogaster*, a hallmark of such template-switching retrotransposition events is short areas of microsimilarity at the “switch points” between the LTR and endogenous gene transcripts (35). Consistent with this, at the precise putative switch point between ELTR1 and *PxyMasc* exon 4, we observed a shared sequence (ATTT) in these two transcripts whose length fits within the variance of such switch points observed previously in *D. melanogaster* retrocopies (Fig. 7). We were unable to identify a sequence similar to ELTR1 on the opposite side of long *Pxyfem*, i.e., adjoining sequence homologous to *PxyMasc* exon 7, though a short, conserved sequence of unknown origin was identified in this position (Fig. 7A). Similarly, we were unable to identify the LTR sequences that would have been required for genomic integration of a chimeric retrotransposon. Sequence drift in these areas, as well as rounds of recombination and reduplication, may have removed these sequences or altered them beyond recognition. Previous studies exploring gene duplication through retrotransposition have utilized “young” retrocopies specifically to minimize these confounding effects (35). The severely truncated nature of the transposable element (TE) ORFs identified on this W chromosome scaffold implies that this degeneration and fragmentation would not have been unique to the hypothesized ancestral *Pxyfem*-PLTR1 integration.

The architecture of short *Pxyfem* provides a clue as to how these reduplication events may have occurred. Relative to long *Pxyfem*, the four short *Pxyfem* copies are truncated on their 5' ends and, in place of this lost sequence, contain the PLTR2 retrotransposon ORF fragment (Fig. 7B). The putative switch point between short *Pxyfem* and PLTR2 is more difficult to assess than for PLTR1 due to a 15-bp tandem duplication of the terminal *Pxyfem* sequence at this junction. However, an area of potential microsimilarity was identified in the *PxyMasc* CDS immediately downstream of this duplicated sequence and immediately upstream of the PLTR2 homologous region in ELTR2, with the putative switch point containing a combination of the single nucleotide polymorphisms in these two regions). Short *Pxyfem* may represent the result of a retrotransposition event involving an ancestor of the ELTR2 retrotransposon and the previously integrated PLTR1-long *Pxyfem* fusion. Consistent with the hypothesized more-recent origin of short *Pxyfem*, we observed a 10-bp element repeated at the junctions of PLTR1/PLTR2 and their respective flanking genomic regions, which may represent the target site duplication signal of a LTR retrotransposition event (34).

When did *Pxyfem* arise? There are two mutually exclusive scenarios regarding the evolutionary relationship between the *B. mori*

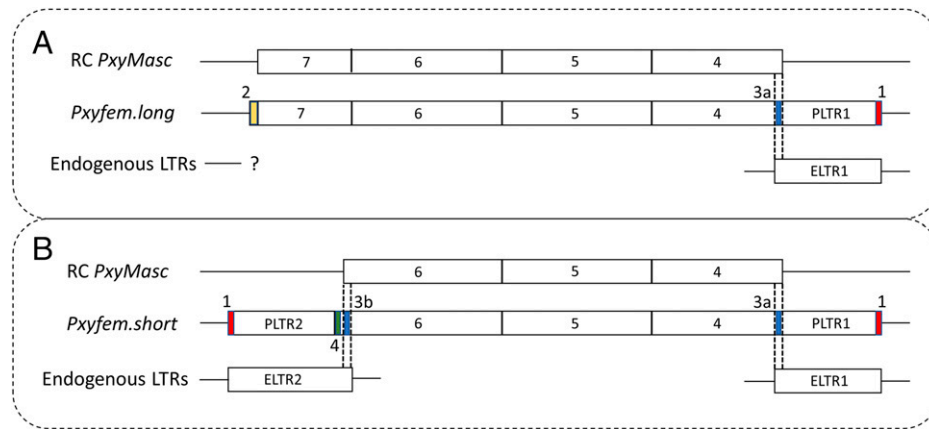


Fig. 7. Microsimilarity and repeat regions suggest a possible role for “retrotransposition” in the origins of *Pxyfem*. (A) Analysis of “*Pxyfem long*”. From top to bottom, the three sequences are RC *PxyMasc* = reverse complement of *PxyMasc* transcript; *Pxyfem long* = generalized schematic of the *Pxyfem long* copies, including 250 bp of up- and downstream genomic flanking sequence; and endogenous LTRs = the closest homolog to PLTR1 that flanks *Pxyfem long*. Nos. above colored segments in *Pxyfem long* denote conserved sequences as follows: 1 = 10-bp repeat: (ACACAAGGCT), 2 = 34 bp conserved but unidentified sequence: (ACCCAGCAGTTACATGTTGGGAAGCAGAAATTT), and 3a = *PxyMasc*/ELTR1 switch point: (ATTT). Dashed lines represent the putative switch points between RC *PxyMasc* and ELTR1, which gave *Pxyfem*. (B) Analysis of “*Pxyfem short*”. Three sequences same as in A but exchanging the generalized *Pxyfem long* sequence with the generalized *Pxyfem short* sequence. 3b = *PxyMasc*/ELTR2 switch point: *PxyMasc* sequence = AAGAAG, ELTR1 sequence = AACAG, putative switch point = AAGCAG. Dashed lines represent the putative switch points between RC *PxyMasc* and ELTR2, which gave *Pxyfem short*. 4 = 15 bp tandem duplication: (AGTCTGCTGGTTAAGAAGAAGTCTGCTGGTTAAG).

and *P. xylostella* “feminizer” systems. Scenario 1: *fem* and *Pxyfem* derive from a common ancestor (“proto-feminizer”) that evolved prior to the split of the *Plutellidae* from the rest of the Ditrysia, ca. 125 Mya. Scenario 2: *fem* and *Pxyfem* evolved independently after this split, and the functional similarities between them represent convergent evolution. These two scenarios are explored below.

Assuming that *Pxyfem* is a partial *PxyMasc* retrocopy, one way to assess the two scenarios is through comparison of the fossilized TEs involved—PLTR1 and PLTR2—with other lepidopteran LTR retrotransposons. Under scenario 1 (common origin) we would not expect these to show higher similarity to remaining ancestral sequences in the *P. xylostella* genome (i.e., ELTR1/ELTR2) than to homologs of those sequences in other higher lepidopterans. Conversely, if the original template-switching event took place after the radiation of the ditrysiids (scenario 2— independent origin), we would expect that PLTR1 and PLTR2 would resemble ELTR1/ELTR2 more than other lepidopteran homologs. Consistent with scenario 2, when BLASTed, translated PLTR1/PLTR2 consensus sequences show higher similarity to LTR retrotransposons in the male-derived *P. xylostella* genome (i.e., in *P. xylostella* autosomal or Z-linked regions) than to hits in other lepidopteran assemblies. Moreover, we could not identify sequences similar to any retrotransposable element ORFs in the 767-bp *B. mori fem*-precursor transcript. Though these LTR-related sequences (i.e., a hypothetical BLTR1/BLTR2) may have been lost from the *B. mori fem* transcript through sequence turnover, it does not explain why the remnants of these sequences in *P. xylostella* (PLTR1/PLTR2) contain ORFs more similar to those found in *P. xylostella* autosomal regions.

In principle, this greater similarity could be due to homogenization of paralogous sequences by nonallelic gene conversion (NAGC). However, the nucleotide sequences of the PLTR1 copies and ELTR1 (i.e., the sequences on which the homology-based NAGC would have theoretically acted) show relatively low similarity (PLTR1/ELTR1 = ~57% identity over ~251 bp, PLTR2/ELTR2 = ~60% identity over 236 bp). It would appear that gene conversion, which is known to convert runs of, on average, a few hundred base pairs at a time at a rate 10–100× faster than point mutation (36), has had little observable role in the concerted evolution of these PLTR/ELTR pairs.

The two potential origin scenarios should also have different implications for evolution of *PxyMasc* and *Pxyfem* sequences. The function of *Pxyfem* established here is to provide a source of ssRNAs targeting *PxyMasc*. We would expect, therefore, that selection pressure on differing regions of *Pxyfem* to maintain nucleotide agreement with *PxyMasc* would be in proportion to the relative contributions of those areas toward that function. We would further anticipate that the redundancy of the 11 *Pxyfem* copies may have allowed the sequences of “low-functionality” areas to drift considerably over evolutionary time, both within and between copies. Under the ancient origin of scenario 1, its high level of within-sequence agreement would thus suggest that there has been extremely strong selection pressure across the *Pxyfem* sequence to maintain this agreement. This expectation is difficult to match with the observation that the area of *Pxyfem* that our sequencing identified as producing ssRNAs against *PxyMasc* is relatively concentrated. Indeed, the first 46 bp of *Pxyfem short* and *Pxyfem long*, despite showing 89% nucleotide identity to exon 4 of *PxyMasc*, did not produce a single antisense read mapping to *PxyMasc*. Similarly, for *Pxyfem long*, the final 97 bp, which displayed 96% nucleotide identity with exons 6 and 7 of *PxyMasc*, produced only a single antisense read mapping to *PxyMasc*. Given that these two regions appear to contribute few or no ssRNAs to the silencing of *PxyMasc*, it is difficult to explain why such tight sequence agreement between them and the relevant regions of *PxyMasc* would have been maintained for >125 My.

Under scenario 2 (independent origin), this situation can be explained easily, as here, the sequence-specific *feminizer* systems of *P. xylostella* and *B. mori* arose after the sequence divergence of their two respective *Masculinizer* homologs. Under this scenario, both the observed within-system agreement and the retention of low-functionality areas of *Pxyfem* are a consequence of the relatively recent origins of this system. This seems to fit the available data better than scenario 1.

In conclusion, we have identified a W-linked multicopy ssRNA-generating system in *P. xylostella* that targets the male-determining gene *PxyMasc*. This system produces ssRNAs belonging to both the siRNA/miRNA and piRNA size classes, suggesting that the *Pxyfem* precursor transcripts are processed by multiple RNAi machineries in the early embryo. Consistent with the hypothesis that this system functions like *feminizer* in

B. mori to silence the *Masculinizer* gene and determine female sexual fate, we observed that it initiates expression during early embryogenesis, peaking 12 h postlaying. Previously, we established that the *PxyMasc* transcript—initially expressed (3–6 h postlaying) in both male and female embryos—can no longer be observed in female embryos by 24 h postlaying, correlating with a transition to female-specific splicing of *doublesex*. The peak expression of the *Pxyfem* system thus overlaps with the reduction in *PxyMasc* transcripts observed in female *P. xylostella* embryos and a transition toward female somatic differentiation (the function of female-form *doublesex*). We note that this evidence is correlative and our conclusions would be strengthened by causative analysis (e.g., silencing or knockout of the *Pxyfem* system). However, the highly redundant nature of this system (multiple loci producing multiple, different ssRNAs produced by multiple RNAi machineries) makes this extremely technically challenging, e.g., compared with functional interrogation of single-copy, protein-coding genes or a single, homogenous piRNA sequence (e.g., *fem* in *B. mori*). As such, we chose to employ a “mode-of-action”-based approach for functional assessment of the *Pxyfem* system. Further supporting the proposed mode of action of *Pxyfem*, experiments with exogenous transcripts “labeled” with the *Pxyfem* target sequence indicated that this 325-bp sequence is targeted for ssRNA-mediated cleavage in the early female embryo. Intriguingly, our analyses suggest the more likely origin scenario for this *Pxyfem* system is that it arose after the divergence of the basal *Plutellidae* from the higher Ditrysia and that similarities between it and *B. mori fem* represent convergent evolution. Our model for the origin and evolution of *Pxyfem* implicates involvement of LTR–transposon retrotransposition. Further work elucidating the evolutionary relationship between the lepidopteran *W* chromosomes and, especially, the sequencing of other members of the *Plutellidae* may further aid in this analysis.

Methods

Insect Details. *P. xylostella* (originally collected from Vero Beach, Florida, USA) were reared on beet armyworm artificial diet (Frontier Biosciences, Germantown, MD, USA) under a 16:8 h light: dark cycle, 25 °C, and 50% relative humidity.

Primers. Details of all primers used in this study can be found in *SI Appendix, Table 2*.

Generating *P. xylostella* Small RNA Libraries and Mapping to *PxyMasc*.

Larval L1 samples. Newly hatched (within 1 h) *P. xylostella* larvae were individually collected, placed in RNAlater (Thermo Fisher Scientific, Waltham, MA, USA), immediately frozen in liquid nitrogen, and stored at –80 °C. RNA was extracted from each sample using a miRNeasy Mini Kit and a RNeasy MinElute Cleanup Kit (both from Qiagen, Valencia, CA, USA), in order to isolate both >200-nucleotide (nt) and <200-nt fractions from each sample. Less than 200-nt isolates were immediately frozen at –80 °C, while >200-nt isolates were used as template for cDNA synthesis (LunaScript RT Mastermix Kit, New England Biolabs, MA, USA) and subsequent *doublesex* sexing PCR as previously described (25). Once the sex of each sample had been confirmed through observing *doublesex* splicing patterns, <200-nt samples were combined to make a male ($n = 7$) and female ($n = 5$) small RNA isolate. These two combined samples were ethanol precipitated and used to generate male and female L1 small RNA libraries using the NEBNext Multiplex Small RNA Library Prep Set for Illumina (New England Biolabs, MA, USA). Indexed libraries were size selected for small RNAs (sRNAs) (<150 nt) using AMPure XP magnetic beads (1.3× followed by 3.7×) (Beckman Coulter, Brea, CA, USA). The final size of each library was determined using a D100 High Sensitivity Screentape on a TapeStation 2200 (Agilent, Santa Clara, CA, USA). Libraries were quantified using the Qubit dsDNA BR assay (ThermoFisher, MA, USA), diluted to 20 nM and further quantified using the NEBNext Library Quant Kit for Illumina (New England Biolabs, MA, USA). Samples were pooled and diluted to 4 nM and run on an Illumina Miseq by the Bioinformatics, Sequencing, and Proteomics Facility at The Pirbright Institute.

Embryo samples. Adult *P. xylostella* were allowed to lay eggs on a cabbage juice-painted parafilm sheet for 5 min. Laid eggs were removed from the cage after this time and placed into a humidified Petri dish. At each time point (3, 6, 9, 12, and 24 h postlaying), 100 eggs were collected from the parafilm sheet, pooled, placed in RNAlater, and frozen in liquid nitrogen. Downstream processing of these five egg samples followed the same workflow outlined above; however, >200-nt isolates were collected, but not used, for *doublesex* sexing through RT-PCR, as these were mixed-sex collections/extractions. Library preparation and analysis were performed as for the L1 samples. The prepared sRNA libraries were sequenced using the Illumina Miseq platform as for the L1 samples.

Sequencing analysis. Sequenced reads were checked for quality using FastQC (37). The TruSeq sequencing adapters (RNA 5' adapter: GUUCAGAGUUCUACA-GUCCGACGAUC; (RA5); part No. 15013205, RNA 3' adapter: TGGAATCTCGGGTG CCAAGG; (RA3); part No. 15013207) and the poor-quality reads (Phred scores; $Q < 20$) were removed using Trimmomatic (38). Further, the reads that were not in the ssRNA size range (20–40 bp) were filtered out using the Seqkit toolkit (39). The preprocessed reads were then mapped to the *PxyMasc* sequence using GSNAP (40). The mapped alignment file was processed to identify sense and antisense reads using SAMtools (41). Mapped outcomes (size distribution and density plots) from different time-course experiments and male–female-specific libraries were then plotted using the Graphpad Prism 8 (GraphPad Software, Inc., USA). Integrated Genomics Viewer (42) was used to visualize the alignment files. The nucleotide usage (consensus sequences) from the identified similar size ssRNAs were plotted as sequence logos using Weblogo3 (43).

***Pxyfem* RACE.** From sRNA mapping of reads deriving from L1 samples, *PxyMasc* appeared to be targeted by two ssRNA groups being complementary to exons 4 and 5/6, respectively. These two areas were used to design 5' and 3' RACE primers (*SI Appendix, Table 2*) in order to identify mRNA sequence from the *Pxyfem* precursor transcript. 5' and 3' RACE ready cDNA was generated from a pooled female pupal RNA sample and also from a pooled adult female ovary sample (RNA extracted using RNeasy MinElute kit, RACE conducted using SMARTer 5'/3' RACE kit—Takara Bio, Kyoto, Japan). Visible bands were cloned using the NEB PCR cloning kit (New England Biolabs, MA, USA) and Sanger sequenced. Sequences that did not show evidence of mispriming (i.e., those that did show significant homology to *PxyMasc* in reverse complement) were further characterized through BLASTn and translated reading frames through BLASTp analysis (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>). For BLASTn, whole-genome shotgun contigs of organism “*P. xylostella*” were used as search subject. For BLASTp, non-redundant protein sequences of the same organism was used. In the latter case, significant protein matches were subsequently reblasted without species limitations in order to identify the families to which these belonged. Putative domains within protein homologs were identified using the NCBI conserved domain search (<https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>).

Assessment of *Pxyfem*–*W* Chromosome Linkage. gDNA from six male and five female pupae (sexed by eye) was extracted using the NucleoSpin Tissue kit (Macherey-Nagel, Düren, Germany), pooled by sex, and used as template for subsequent PCRs. Four PCRs were run on each template, informed by the sequencing results arising from 5' RACE analysis. PCR set 1 (primers LA4886 + LA4887) was designed to amplify across exons 5 and 6 of the region, showing sequence homology to *PxyMasc*; PCR set 2 (primers LA4888 + LA4889) was designed to amplify from the internal region of the transcript across the region showing sequence homology to exons 6 and 5 of *PxyMasc* (and thus has the same reverse primer as primer set 1); PCR set 3 (primers LA4890 + LA4891) was designed within the MRR1; and PCR set 4 (primers LA309 + LA310) was designed to amplify within the 17S genomic locus as a positive gDNA control. All PCRs were run using DreamTaq DNA polymerase (Thermo Fisher Scientific, MA, USA) and the program 98 °C–1 min, 35 cycles of 98 °C–30 s, 50 °C–30 s, 72 °C–30 s, and final extension 72 °C–2 min. A no template control was also run for each primer set with H₂O added instead of gDNA template. Primer sequences are detailed in *SI Appendix, Table 2*.

Positive-Readout mRNA Cleavage Experiment. The positive-readout cassette was based on a previously demonstrated design and was synthesized by Genewiz, Inc. In brief, the 325-bp *PxyMasc* sequence matching to *Pxyfem* was placed directly downstream of a polyA sequence and directly upstream of the CopA sequence. As previously, CopT was placed in the 5' UTR of nanoluciferase. This nanoluciferase transcript was placed under the control of the Hr5/le1

promoter, and this cassette was cloned into a pre-existing backbone (AGG1938) that contained an Op/le2-ZsGreen-Sv40 fluorescent marker cassette to give the positive readout construct AGG2208 (*SI Appendix, Fig. 4A*—GenBank accession No.: ON165680).

Initial testing of construct AGG2208 in AAG2 cells. dsRNA matching the 345-bp *PxyMasc* target region in AGG2208 ("treatment" dsRNA) and a 456-bp region of *AmCyan* CDS (negative control dsRNA) was generated using the T7 Megascript kit (Thermo Fisher Scientific, Waltham, MA, USA) according to the manufacturer's instructions, with subsequent cleanup of dsRNA product using the Megaclear kit (Thermo Fisher Scientific). Initial DNA samples for each dsRNA product were generated by PCR using AGG2208 as the template for treatment dsRNA and a standard *AmCyan* construct for the negative-control dsRNA. Primers used to generate these dsRNA products are listed in *SI Appendix, Table 2*. For the experimental testing, *Aedes aegypti* Aag2 cells were seeded in 96-well plates at a density of 50,000 cells/well. The following day, cells were cotransfected with AGG2208 plasmid and either treatment dsRNA or negative-control dsRNA or no dsRNA (dsRNA was used at two different concentrations: 2 ng and 50 ng/well). All cells were also cotransfected with a Hr5/le1-firefly-expressing plasmid (AGG1186—*SI Appendix, Fig. 4B*—GenBank accession no. MT119956 (44)) functioning as a normalization control. Each transfection treatment was carried out in eight technical replicates. After transfection, cells were incubated at 28 °C for 48 h and lysed with passive lysis buffer (Promega, WI, USA), and luciferase assays were performed using the Nano-Glo Dual Luciferase assay system (Promega). Luciferase values were measured using a GloMax-Multiplate reader (Promega) and results analyzed by performing Student's *t* tests (two tailed) using GraphPad Prism Version 9.3.1.

Use of construct AGG2208 in *P. xylostella* embryos. Prior to the experiment, AGG2208 and AGG1186 were combined with molecular-grade H₂O (final plasmid concentrations of 400 ng/μL and 300 ng/μL, respectively) to provide an injection mix. This mix was injected into <30-min-old *P. xylostella* embryos using techniques previously described (45). Forty-eight hours later, embryos were inspected under a fluorescence microscope, and eggs that did not show strong evidence of ZsGreen fluorescence (the fluorophore central marker on the injected plasmid that will fluoresce independently of any other effects) were excluded on the basis that these either had little to no mix injected or had died at an early stage after injection. Remaining eggs were transferred into individual PCR tubes and frozen at −20 °C.

Subsequently, eggs were individually homogenized in 5 μL diethyl pyrocarbonate water within the PCR tube and 1 μL was used for gDNA extraction using Phire Animal Tissue Direct PCR Kit (ThermoFisher, MA, USA) according to manufacturer's instructions, following the dilution protocol but adding 1 μL of the homogenized sample instead of tissue. Two different PCRs were performed for each egg sample: one for a confirmed autosomal locus (MRR1) to act as an extraction control (primers LA4890 + LA4891) and one for a confirmed W-specific locus (*Pxyfem*) to identify female embryos (primers LA4888 + LA4889). These reactions were confirmed in the *Assessment of Pxyfem-W Chromosome Linkage* section, above. PCRs were performed using Phire Animal Tissue Direct PCR Kit according to manufacturer's instructions using 2 μL of gDNA template for the genomic PCR and 3 μL of template for the female-specific PCR. PCRs were performed using a standard cycling heat block with the following program: 98 °C–5 min, 35 cycles of 98 °C–5 s, 64.9 (MRR1)/65.4 (*Pxyfem*) °C–5 s, 72 °C–20 s, and final extension 72 °C–1 min. Primer sequences are detailed in *SI Appendix, Table 2*.

To the remaining 4 μL of each egg homogenate sample, passive lysis buffer (Promega, WI, USA) was added to a 1× final dilution and incubated for 15 min

at room temperature. The resulting product was used to perform luciferase assays using Nano-Glo Dual Reporter Assay kit (Promega, WI, USA) according to manufacturer's instructions. Absorbance was measured using a GloMax Microplate reader (Promega, WI, USA). Both firefly and nanoluciferase values were measured, and nanoluciferase values of each sample were normalized against corresponding firefly values. Analysis was conducted on Prism (Graphpad, San Diego, CA, USA) using a two-tailed, unpaired *t* test. Female embryo replicates equaled 24, and male embryo replicates equaled 35.

Assessment of *Pxyfem* Operon Linkage. Samples included (1) cDNA generated from >200 nt RNA extracted previously from the 12-h embryo sample (LunaScript RT Mastermix Kit, New England Biolabs), (2) a no-RT control of the embryo sample, (3) molecular-grade H₂O, and (4) female pooled gDNA extracted previously for the *Assessment of Pxyfem-W Chromosome Linkage* experiment. Each of these samples was used as template for a PCR using primers LA2549 and LA4887 (*SI Appendix, Table 2*), which extended outwards from each *Pxyfem* locus. As such, amplicons produced by these reactions must occur between *Pxyfem* loci. PCRs were performed using Q5 polymerase (New England Biolabs, MA, USA) and the following thermocycler program: 98 °C–1 min, 35 cycles of 98 °C–30 s, 68 °C–30 s, 72 °C–3 min, and final extension 72 °C–2 min.

Data, Materials, and Software Availability. Raw sequencing reads are available at NCBI bioproject [PRJNA893552](https://doi.org/10.5061/dryad.mpg4f4r3g) (46). Luciferase values generated from the positive reporter experiments are available from <https://doi.org/10.5061/dryad.mpg4f4r3g> (47).

ACKNOWLEDGMENTS. D.P.'s PhD studentship was funded by The Pirbright Institute. L.Z.C.P. was supported by a Wellcome Trust grant made to L.A. (110117/Z/15/Z). For the purpose of open access, the author has applied a CC BY public copyright license to any author-accepted manuscript version arising from this submission. M.A.E.A. and L.A. were funded through a Defense Advanced Research Projects Agency award (N66001-17-2-4054) to Kevin Esvelt at Massachusetts Institute of Technology. The views, opinions, and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the US Government. V.C.N. and T.H.-S. were supported by European Union H2020 grant nEUROSTRESSPEP (634361) made to L.A.; X.X. was supported by the Chinese Scholarship Council Scholarship from the Chinese Government and a PhD student exchange program from Fujian Agriculture and Forestry University. T.H.-S. was supported by a UK Biotechnology and Biological Sciences Research Council (BBSRC) Impact Acceleration Account grant (BB/S506680/1) made to T.H.-S. and L.A.; L.A. and T.H.-S. were additionally supported by core funding from the BBSRC to The Pirbright Institute (BBS/E/I/00007033, BBS/E/I/00007038, and BBS/E/I/00007039). C.M.R. was supported by a Wellcome Trust grant made to L.A. (200171/Z/15/Z). The authors would like to acknowledge the assistance of Dr. Graham Freimanis at the Pirbright Institute Bioinformatics, Sequencing, and Proteomics Unit.

Author affiliations: ^aArthropod Genetics Group, The Pirbright Institute, Woking, Pirbright GU24 0NF, United Kingdom; and ^bState Key Laboratory of Ecological Pest Control for Fujian and Taiwan Crops, Institute of Applied Ecology, Fujian Agriculture and Forestry University, Fuzhou 350002, China

Author contributions: T.H.-S., L.Z.C.P., C.M.R., M.Y., and L.A. designed research; T.H.-S., X.X., M.A.E.A., L.Z.C.P., D.P., V.C.N., and C.M.R. performed research; T.H.-S., M.A.E.A., L.Z.C.P., D.P., and C.M.R. analyzed data; T.H.-S., X.X., M.A.E.A., L.Z.C.P., D.P., V.C.N., C.M.R., M.Y., and L.A. wrote the paper.

The authors declare no competing interest.

1. T. Gempe, M. Beye, Function and evolution of sex determination mechanisms, genes and pathways in insects. *BioEssays* **33**, 52–60 (2011).
2. H. Blackmon, L. Ross, D. Bachtrog, Sex determination, sex chromosomes, and karyotype evolution in insects. *J. Hered.* **108**, 78–93 (2017).
3. T. Dafa'alla, G. Fu, L. Alphey, Use of a regulatory mechanism of sex determination in pest insect control. *J. Genet.* **89**, 301–305 (2010).
4. C. Douglas, J. M. A. Turner, Advances and challenges in genetic technologies to produce single-sex litters. *PLoS Genet.* **16**, e1008898 (2020).
5. J. Mallet (2002) The Lepidoptera Taxome Project: Draft proposals and information. <https://www.ucl.ac.uk/taxome/>. Accessed 1 January 2021.
6. T. Kiuchi *et al.*, A single female-specific piRNA is the primary determiner of sex in the silkworm. *Nature* **509**, 633–636 (2014).
7. C. Fraïsse, M. A. L. Picard, B. Vicoso, The deep conservation of the Lepidoptera Z chromosome suggests a non-canonical origin of the W. *Nat. Commun.* **8**, 1486 (2017).
8. P. Nguyen, L. Carabajal Paladino, "On the neo-sex chromosomes of lepidoptera" in *Evolutionary Biology: Convergent Evolution, Evolution of Complex Traits, Concepts and Methods*, P. Pontarotti, Ed. (Springer, Switzerland, 2016), pp. 171–185.
9. M. Dalíková *et al.*, New insights into the evolution of the W chromosome in Lepidoptera. *J. Hered.* **108**, 709–719 (2017).
10. M. Hejníčková *et al.*, Absence of W chromosome in psychidae moths and implications for the theory of sex chromosome evolution in lepidoptera. *Genes (Basel)* **10**, 1016 (2019).
11. A. Yoshida, F. Marec, K. Sahara, The fate of W chromosomes in hybrids between wild silkworms, *Samia cynthia* ssp.: No role in sex determination and reproduction. *Heredity* **116**, 424–433 (2016).
12. S. Visser, A. Voleníková, P. Nguyen, E. C. Verhulst, F. Marec, A conserved role of the duplicated masculinizer gene in sex determination of the Mediterranean flour moth, *Ephesia kuehniella*. *PLoS Genet.* **17**, e1009420 (2021).

13. J. Lee, T. Kiuchi, M. Kawamoto, T. Shimada, S. Katsuma, Identification and functional analysis of a Masculinizer orthologue in *Trilocha varians* (Lepidoptera: Bombycidae). *Insect Mol. Biol.* **24**, 561–569 (2015).
14. N. S. Talekar, A. M. Shelton, Biology, ecology, and management of the diamondback moth. *Annu. Rev. Entomol.* **38**, 275–301 (1993).
15. M. J. Furlong, D. J. Wright, L. M. Dossdall, Diamondback moth ecology and management: Problems, progress, and prospects. *Annu. Rev. Entomol.* **58**, 517–541 (2013).
16. M. P. Zalucki *et al.*, Estimating the economic cost of one of the world's major insect pests, *Plutella xylostella* (Lepidoptera: Plutellidae): Just how long is a piece of string? *J. Econ. Entomol.* **105**, 1115–1129 (2012).
17. T. Harvey-Samuel *et al.*, Pest control and resistance management through release of insects carrying a male-selecting transgene. *BMC Biol.* **13**, 49 (2015).
18. L. Jin *et al.*, Engineered female-specific lethality for control of pest Lepidoptera. *ACS Synth. Biol.* **2**, 160–166 (2013).
19. T. Harvey-Samuel *et al.*, Engineered expression of the invertebrate-specific scorpion toxin AaHIT reduces adult longevity and female fecundity in the diamondback moth *Plutella xylostella*. *Pest Manag. Sci.* **77**, 3154–3164 (2021).
20. X. Xu *et al.*, Toward a CRISPR-Cas9-based gene drive in the diamondback moth *Plutella xylostella*. *CRISPR J.* **5**, 224–236 (2022).
21. C. M. Ward *et al.*, A haploid diamondback moth (*Plutella xylostella* L.) genome assembly resolves 31 chromosomes and identifies a diamide resistance mutation. *Insect Biochem. Mol. Biol.* **138**, 103622 (2021).
22. A. Kawazoe, The chromosome in the primitive or microlepidopterous moth-groups II. *Proc. Jpn. Acad., Ser. B, Phys. Biol. Sci.* **3**, 87–90 (1987).
23. M. You *et al.*, A heterozygous moth genome provides insights into herbivory and detoxification. *Nat. Genet.* **45**, 220–225 (2013).
24. S. W. Baxter *et al.*, Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS One* **6**, e19315 (2011).
25. T. Harvey-Samuel, V. C. Norman, R. Carter, E. Lovett, L. Alphey, Identification and characterization of a Masculinizer homologue in the diamondback moth, *Plutella xylostella*. *Insect Mol. Biol.* **29**, 231–240 (2020).
26. S. Katsuma, Y. Sugano, T. Kiuchi, T. Shimada, Two conserved cysteine residues are required for the masculinizing activity of the silkworm Masc protein. *J. Biol. Chem.* **290**, 26114–26124 (2015).
27. C. B. Stein *et al.*, Decoding the 5' nucleotide bias of PIWI-interacting RNAs. *Nat. Commun.* **10**, 828 (2019).
28. B. Czech, G. J. Hannon, One loop to rule them all: The ping-pong cycle and piRNA-guided silencing. *Trends Biochem. Sci.* **41**, 324–337 (2016).
29. I. Gainetdinov, C. Colpan, A. Arif, K. Cecchini, P. D. Zamore, A single mechanism of biogenesis, initiated and directed by PIWI proteins, explains piRNA production in most animals. *Mol. Cell* **71**, 775–790.e5 (2018).
30. B. W. Han, W. Wang, C. Li, Z. Weng, P. D. Zamore, Noncoding RNA. piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production. *Science* **348**, 817–821 (2015).
31. N. Kandul, M. Guo, B. A. Hay, A positive readout single transcript reporter for site-specific mRNA cleavage. *PeerJ* **5**, e3602 (2017).
32. J. Brosius, Retroposons—seeds of evolution. *Science* **251**, 753 (1991).
33. E. F. Vanin, Processed pseudogenes: Characteristics and evolution. *Annu. Rev. Genet.* **19**, 253–272 (1985).
34. E. R. Havecker, X. Gao, D. F. Voytas, The diversity of LTR retrotransposons. *Genome Biol.* **5**, 225 (2004).
35. S. Tan *et al.*, LTR-mediated retroposition as a mechanism of RNA-based duplication in metazoans. *Genome Res.* **26**, 1663–1675 (2016).
36. A. Harpak, X. Lan, Z. Gao, J. K. Pritchard, Frequent nonallelic gene conversion on the human lineage and its effect on the divergence of gene duplicates. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 12779–12784 (2017).
37. S. Andrews, *FastQC: A Quality Control Tool for High Throughput Sequence Data* (Babraham Institute, 2010).
38. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
39. W. Shen, S. Le, Y. Li, F. Hu, SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* **11**, e0163962 (2016).
40. T. D. Wu, S. Nacu, Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
41. H. Li *et al.*; 1000 Genome Project Data Processing Subgroup, The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
42. J. T. Robinson *et al.*, Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
43. G. E. Crooks, G. Hon, J. M. Chandonia, S. E. Brenner, WebLogo: A sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
44. P. Y. L. Tng *et al.*, Cas13b-dependent and Cas13b-independent RNA knockdown of viral sequences in mosquito cells following guide RNA expression. *Commun. Biol.* **3**, 413 (2020).
45. S. Martins *et al.*, Germline transformation of the diamondback moth, *Plutella xylostella* L., using the piggyBac transposable element. *Insect Mol. Biol.* **21**, 414–421 (2012).
46. T. Harvey-Samuel *et al.*, Silencing RNAs expressed from W-linked PxyMasc "retrocopies" target that gene during female sex determination in *Plutella xylostella*. NCBI: BioProject. <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA893552>. Deposited 24 October 2022.
47. T. Harvey-Samuel, Normalised luciferase assay data. Dryad. <https://doi.org/10.5061/dryad.mpg4f4r3g>. Accessed 26 October 2022.