

A widespread W-linked reservoir of masculinizing-gene homologs underlies recurrent evolution of feminizing piRNA loci in Lepidoptera

Suijie kuang

Hunan Agricultural University

zhuotian yao

Hunan Agricultural University

Yue Li

Hunan Agricultural University

yatao zhu

Hunan Agricultural University

youwei gong

Hunan Agricultural University

yuan qiu

Hunan Agricultural University

wenbing ding

Hunan Agricultural University

hualiang he

Hunan Agricultural University

hongshuai gao

Hunan Agricultural University

qiao Gao

Hunan Agricultural University

dong jing

Hunan Agricultural University

youzhi li

Hunan Agricultural University

lin Qiu

qiulin@hunau.edu.cn

Hunan Agricultural University

Keywords: piRNA, sex determination, W chromosome, Lepidoptera, Masculinizer, convergent evolution

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10825605/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A widespread W-linked reservoir of masculinizing-gene homologs underlies recurrent evolution of feminizing piRNA loci in Lepidoptera

Suijie Kuang¹, Zhuotian Yao¹, Yue Li¹, Yatao Zhu¹, Youwei Gong¹, Yuan Qiu¹, Wenbing Ding^{2,3}, Hualiang He¹, Hongshuai Gao¹, Qiao Gao¹, Dong Jing¹, Youzhi Li^{1,3,4,*}, Lin Qiu^{1,4,*}

¹ Hunan Provincial Key Laboratory for Biology and Control of Plant Diseases and Insect Pests, College of Plant Protection, Hunan Agricultural University, Changsha 410128, China.

² National Research Center of Engineering & Technology for Utilization of Botanical Functional Ingredients, Hunan Agricultural University, Changsha 410128, China.

³ Hunan Provincial Engineering & Technology Research Center for Biopesticide and Formulation Processing, Hunan Agricultural University, Changsha 410128, China.

⁴ Yuelushan Laboratory, Changsha 410128, China.

* Co-corresponding authors: Youzhi Li (liyoushi@hunau.edu.cn), Lin Qiu (qiulin@hunau.edu.cn)

Abstract

Background

In moths and butterflies (Lepidoptera), several distantly related lineages have independently arrived at a similar feminizing strategy: piRNAs from the W chromosome that silence masculinizing genes. Whether this convergence is coincidental or reflects a sequence-level predisposition of the W chromosome has remained unknown.

20 **Results**

21 Screening 266 lepidopteran genome assemblies shows that W chromosomes of 106 species
22 spanning 20 families harbor sequences homologous to Masc (Masculinizer) or znf-2 (zinc finger
23 protein 2), in three structural forms: intronless copies (42 species), intron-retaining copies (32
24 species) and short tandem arrays (61 species), with some species carrying more than one form.
25 Gene trees resolve these copies as independent, lineage-specific acquisitions. Transcriptome
26 screening detects transcriptional activity at W-linked loci in 12 of the 61 species with public
27 transcriptome data. Embryonic small-RNA and transcriptome profiling across four species from
28 three families identifies W-derived antisense piRNAs targeting Masc or znf-2 in three. In *Chilo*
29 *suppressalis*, knockdown of the W-linked Masc-derived locus *CsFem* or disruption of the piRNA
30 pathway shifts genetic females toward male-type doublesex splicing, establishing *CsFem* as a
31 feminizing piRNA precursor.

32 **Conclusions**

33 W chromosomes across Lepidoptera independently accumulate sequences homologous to
34 masculinizing genes, providing a sequence-level foundation that predisposes the W chromosome
35 toward the recurrent evolution of piRNA-based feminization. The parallel with small-RNA-based
36 sex determinants in dioecious plants suggests that repeat-rich, non-recombining sex-limited
37 chromosomes are a recurrent cradle for such systems across eukaryotes.

38 **Keywords**

39 piRNA, sex determination, W chromosome, Lepidoptera, Masculinizer, convergent evolution

40 Background

41 Primary sex-determining signals evolve rapidly across insects: although the downstream *doublesex*
42 (*dsx*) switch is deeply conserved, the upstream signals that direct its sex-specific splicing differ
43 markedly among lineages[1,2]. In Lepidoptera, a recurrent upstream mechanism involves piRNAs
44 derived from the W chromosome that silence masculinizing genes to direct female development[3–
45 12]. The principal known targets are *Masc* (*Masculinizer*), a CCCH-type zinc finger protein
46 required for male-type *dsx* splicing and dosage compensation, and *znf-2* (zinc finger protein 2), a
47 *Masc*-interacting factor whose repression likewise redirects the sex-determination cascade toward
48 female development[6,8,9]. Feminizing W-linked loci have now been molecularly characterized in
49 species spanning several families[5,7,8,11–14]. These loci vary markedly in architecture, from
50 transposon-chimeric sequences (*Fem* in *Bombyx mori*[5,13]) to retrocopies of masculinizing genes
51 (*PxyFem* in *Plutella xylostella*[7]), tandem-repeat arrays (*OfFem* in *Ostrinia furnacalis*[8], *EkMom*
52 and *PiMom* in the pyralids *Ephestia kuehniella* and *Plodia interpunctella*[11], *Fet-W* in *Lymantria*
53 *dispar*[12]), and intron-retaining duplications (*LdMasc-W* in *L. dispar*[14]), and share no cross-
54 family sequence homology. Among the loci that target *Masc*, the piRNA-binding sites on *Masc*
55 mRNA are non-overlapping across species[7,11,12]. Moreover, at least one locus (*OfFem*) targets
56 *znf-2* rather than *Masc*[8]. This diversity in locus architecture, target gene, and target site is
57 suggestive of repeated, independent origins, but whether this pattern reflects a systematic
58 predisposition of the W chromosome has not been tested across the order.

59 That a similar molecular solution, a W-linked piRNA precursor that silences a masculinizing gene
60 to direct female *dsx* splicing, has evolved independently in multiple lepidopteran lineages poses a
61 question that the characterization of individual systems cannot answer: does this convergence arise
62 by chance, or does the W chromosome provide a genomic predisposition? Two properties of

lepidopteran W chromosomes could favor such predisposition. Their repeat-rich, transposon-dense composition provides multiple routes by which gene-derived sequences can be captured and amplified[15]. Their lack of recombination promotes the retention of such sequences once integrated[15,16]. Because *Masc* and *znf-2* occupy conserved upstream positions in the sex-determination cascade[3–12], silencing either gene can shift *dsx* splicing toward the female pattern, making them high-value targets for any W-borne piRNA system. If sequences homologous to these genes accumulate on W chromosomes through the routes above, they would constitute a reservoir of precursor material from which feminizing piRNA loci can be independently recruited.

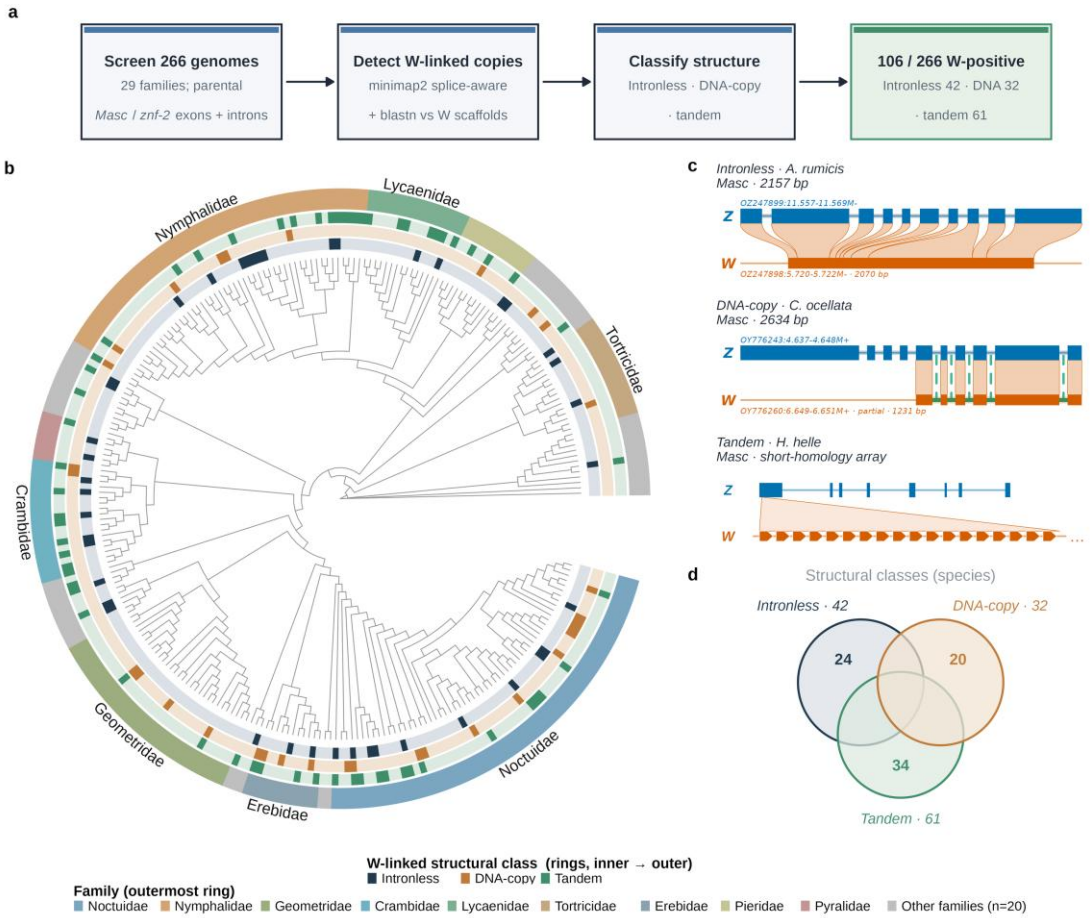
Here we survey 266 lepidopteran genome assemblies for W-linked sequences homologous to *Masc* and *znf-2* and complement this with transcriptome screening, embryonic small-RNA and transcriptome sequencing across four species from three families, and functional validation in *Chilo suppressalis*. We show that W chromosomes in diverse lineages across the order harbor a widespread, structurally diverse reservoir of sequences homologous to masculinizing genes from which feminizing piRNA loci can independently arise.

Results

W chromosomes across Lepidoptera harbor sequences homologous to masculinizing genes in diverse structural forms

A homology screen for W-linked copies of *Masc* and *znf-2* covered 266 lepidopteran genome assemblies across 29 families (Figure 1a). The sampling is almost entirely ditrysian (264 of 266 assemblies). W-linked *Masc*- or *znf-2*-homologous sequence was detected in 106 of the 266 assemblies (~40%), spanning 20 of the 29 sampled families (Figure 1b, Additional file 2). All 106 positive species are ditrysian, while the two non-ditrysian assemblies (*Ectoedemia decentella* and

85 *Tischeria ekebladella*) lacked detectable W-linked homologs. Of the 9 sampled families without any
 86 positive species, most were represented by fewer than 5 assemblies, so their absence may reflect
 87 limited sampling rather than true absence of the reservoir (Additional file 2).



89 **Figure 1.** A W-linked reservoir of *Masc*- and *znf-2*-homologous sequences across Lepidoptera.

90 (a) Screening and classification workflow yielding the 106 W-positive species.

91 (b) Circular phylogeny of all 266 screened lepidopteran genome assemblies (29 families).

92 Outermost band, family. Inner rings, the W-screening call for *Masc/znf-2* by structural class, from
 93 the inside out: intronless, intron-retaining, tandem array. W-linked *Masc*- or *znf-2*-homologous
 94 sequence was detected in 106 of the 266 species (~40%) across 20 families.

(c) One exemplar per structural class as two-chromosome ribbon ideograms (Z-linked parent top, W-linked copy bottom): intronless copy (*Acronicta rumicis Masc*), intron-retaining DNA-level copy (*Cosmorhoe ocellata Masc*) and tandem array (*Helleia helle Masc*, a ~38 bp *Masc*-derived unit reiterated head-to-tail, ~92% parental identity).

(d) Three-set Venn of structural classes by species: intronless copy (42), intron-retaining DNA-level copy (32) and tandem array (61). 34 species are tandem-only and 28 carry two or more classes.

W-linked copies were identified by aligning per-exon and per-intron query features of each parental gene to the corresponding W chromosome with BLASTN (Methods, Figure 1a). The recovered copies fall into several distinct structural forms (Figure 1c,d). Intronless copies span multiple parental exons but lack introns (42 species, 628 copies, 333 *Masc*-derived and 295 *znf-2*-derived, Additional file 1: Figure S1). Intron-retaining copies preserve parental intron-exon structure (32 species, 210 copies). Short-fragment tandem arrays consist of a brief parental-homologous unit, typically matching part of a single exon, reiterated head-to-tail (61 species). Copy number per species varies widely, from a single copy to nearly 100 in the most expanded loci. These categories are not mutually exclusive: 28 species carry more than one form, and even longer homologous fragments occasionally form tandem arrays, as in *S. exigua* (see below).

W-linked copies of all three structural classes have arisen independently across lineages

We next asked whether the W-linked copies share a common ancestry or have arisen independently in each lineage. If the reservoir originated once in an ancestor and was vertically transmitted, W-linked copies from different species should form a single clade separate from their parental genes. If instead copies were acquired independently, each should group with its own species' parental gene. Gene trees built separately for each structural class and each gene clearly favor the second scenario

117 (Figure 2a–f). In every tree, the W-linked copies of a species group with that species' own parental
118 coding sequence to the exclusion of all other species: 26 of 27 species for *Masc* intronless copies
119 and 14 of 17 for *znf-2*, 21 of 21 and 16 of 16 for *Masc* and *znf-2* intron-retaining copies, and 27 of
120 30 and 24 of 33 for *Masc* and *znf-2* tandem arrays, most at ultrafast bootstrap 100. The species that
121 depart from this pattern are congeners that share one ancestral W-copy event, such as *Boloria selene*
122 and *B. thore* for *znf-2* intronless copies (ultrafast bootstrap 82) and *B. titania* and *B. freija* for *Masc*,
123 together with three *Mechanitis* species and three Polyommata genera whose tandem units group by
124 locus rather than by species. No cross-family clade of W-linked copies was recovered in any tree,
125 ruling out a single shared origin across the order. The segment of the parental coding sequence
126 captured by each copy is also broadly distributed, spanning 4% to 94% of relative CDS position
127 across species and covering every decile (Figure 2g), as expected when capture events are
128 independent: a single ancestral capture followed by vertical transmission would fix one segment
129 across the order. The intronless structure of many W-linked copies further argues against ongoing
130 gene conversion from the parental locus, which would preserve intron-exon structure rather than
131 eliminating introns.

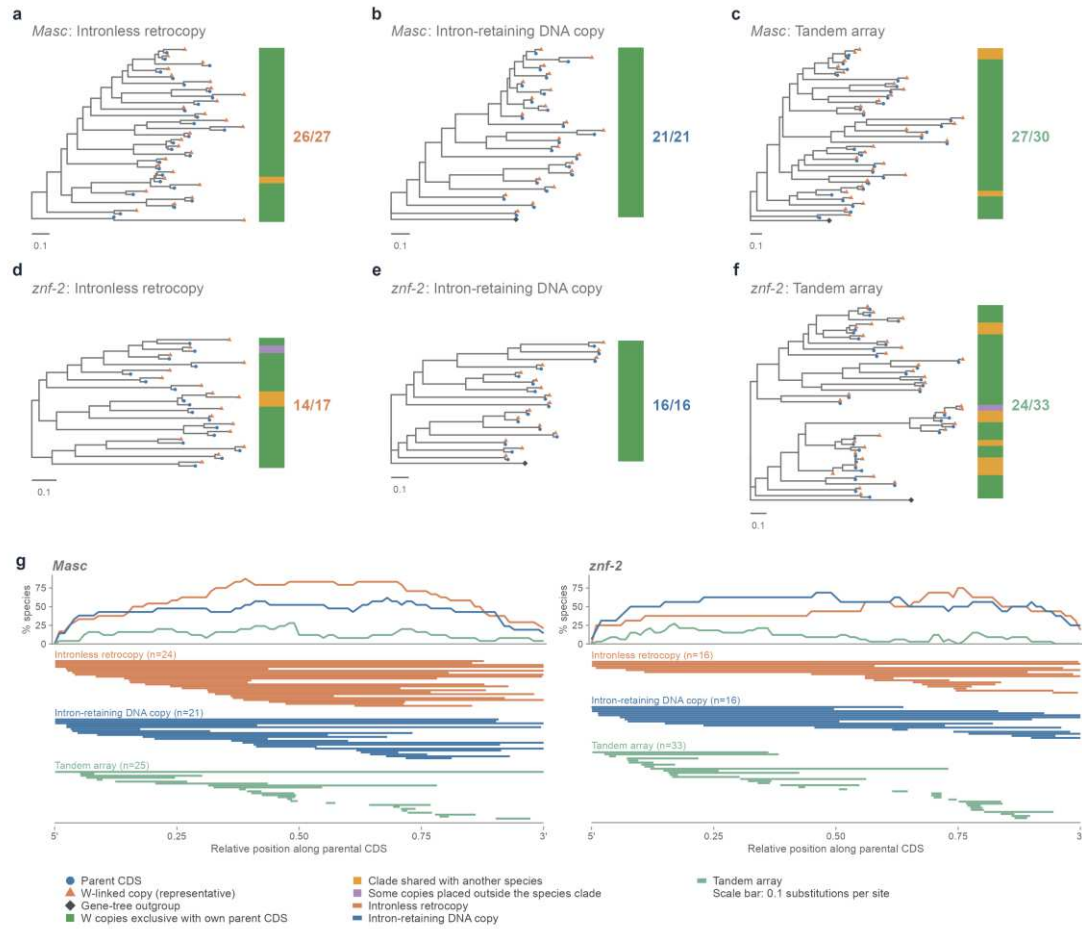


Figure 2. Independent, lineage-specific origins of W-linked *Masc* and *znf-2* copies.

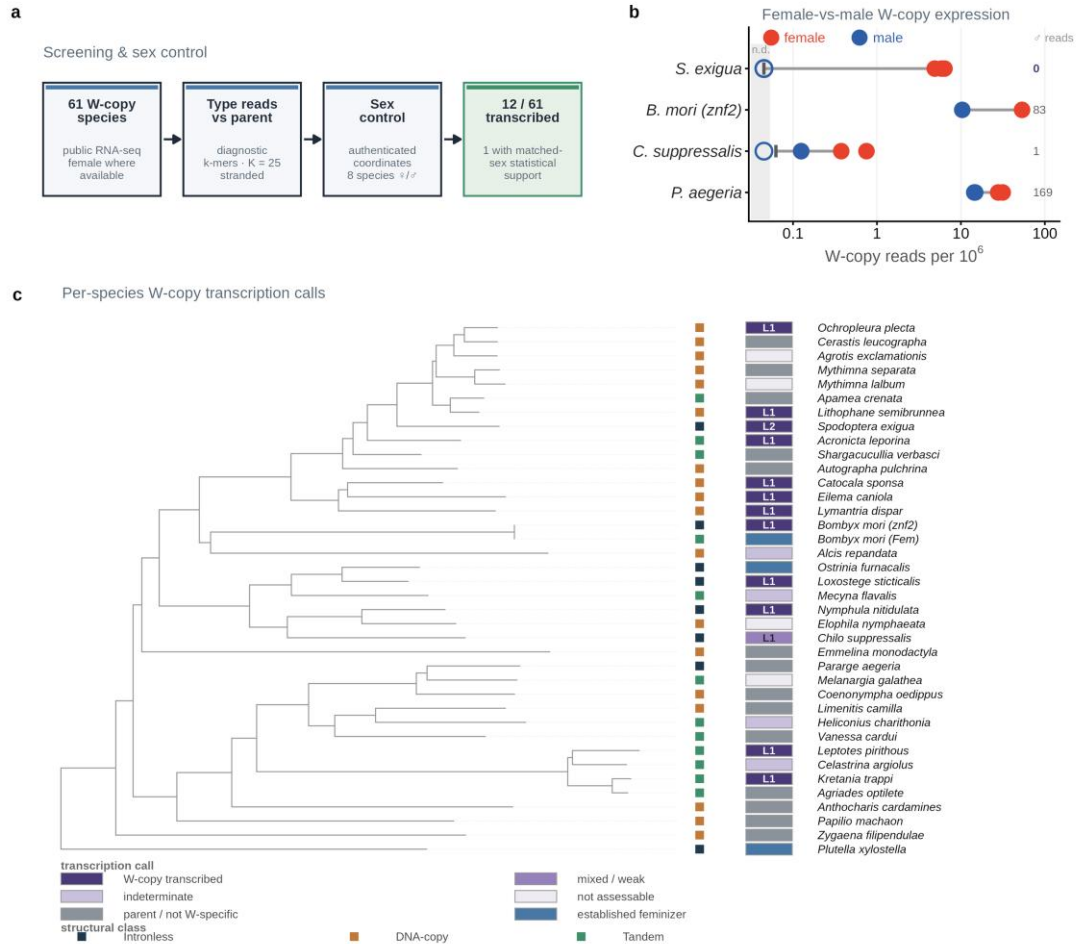
(a–f) Maximum-likelihood gene trees of W-linked copies and their parental coding sequences, one panel per gene and structural class: *Masc* (top row) and *znf-2* (bottom row) intronless retrocopies (a, d), intron-retaining DNA-level copies (b, e) and tandem-array units (c, f). Blue circles, parental coding sequences. Orange triangles, representative W-linked copies. Black diamond, outgroup (*Plutella xylostella* for *Masc* and *Argyresthia albistria* for *znf-2*, both confirmed as orthologs of the intended gene by reciprocal tblastx). The intronless trees in (a) and (d) are midpoint-rooted. Species assignments are unchanged under either rooting. Tip labels are omitted for legibility and the same trees with species names are given in Additional file 1: Figures S2 to S4. The bar beside each tree summarizes each tree species by species: green, the species' parental sequence and all of its W-

143 linked copies form a clade that excludes every other species. Amber, the clade is shared with
144 another species. Purple, individual copies fall outside the species clade. The fraction beside each bar
145 counts species meeting the exclusive criterion. Scale bars, 0.1 substitutions per site.

146 (g) Position of the parental coding sequence captured by each class of W-linked copy, plotted along
147 the parent from 5' to 3' for *Masc* (left) and *znf-2* (right). Upper panels, percentage of species whose
148 copies cover each position. Lower panels, one line per species-level record, grouped by structural
149 class. Independent capture events sample different parts of the coding sequence.

150 **A subset of W-linked loci are transcribed, with female-biased expression in *S. exigua***

151 The presence of target-homologous sequence on the W is only the first prerequisite for a feminizing
152 piRNA locus: to function, such sequence must also be transcribed. To ask whether this reservoir is
153 expressed, we screened the 61 species with public transcriptome data for W-copy transcription
154 using diagnostic k-mer typing (Methods, Figure 3a), spanning all structural categories (Figure 3).
155 This set of 61 species is distinct from the 61 tandem-array species in Figure 1d, and the identical
156 count is coincidental. The full per-species transcription screen, including the public sequencing
157 accessions used, is provided in Additional file 5.



168 = 6.70, $P_{\text{adj}} = 2.6 \times 10^{-13}$, male arm = 0 in 3/3 runs). In *B. mori* (*znf-2*), *C. suppressalis* and *P.*
 169 *aegeria* the male arm also carries W-copy reads, so these three are not scored as female-specific.
 170 The *B. mori* arm is the *znf-2* W-linked copy, not *Fem*. Open symbols, male values at or below
 171 detection (n.d.).

172 (c) Per-species calls on the constrained phylogeny (structural class, transcription call, evidence tier
 173 L1–L2, and literature-established feminizers in blue). *B. mori* appears as two lines, *Fem*[5] and the
 174 *znf-2* W-linked copy screened here. Calls: *mixed* / *weak*, W-copy reads below the positive-call
 175 threshold. *Parent* / *not W-specific*, reads assigned to the parent rather than the W-linked copy. *Not*
 176 *assessable*, read-level typing impossible in principle (e.g. *M. galathea*, whose W copies and parent
 177 share every diagnostic k-mer). Of the 61 species screened, transcription was detected in 12 (20%).
 178 The remaining 24 are in Additional file 5. *H. charithonia* is indeterminate, its array spanning the W
 179 and an autosomal repeat. Tiers: L1, transcription of the W-linked copy detected. L2, female-specific
 180 transcription confirmed with a matched male control. Structural-class squares mark the copy carried
 181 through the screen. Species carrying additional classes appear in Figures 1 and 4.

182 Twelve of the 61 screened species showed W-copy expression in available transcriptomes (20%,
 183 Figure 3c). This rate is constrained by the tissue and developmental-stage coverage of public
 184 transcriptome data, and the true proportion of species with W-copy transcription is likely higher.
 185 Among the loci carried through the screen, intronless copies showed the highest transcription rate (4
 186 of 8 loci, 50%), compared with intron-retaining copies (5 of 18, 28%) and tandem arrays (3 of 12,
 187 25%). These per-class denominators count loci rather than species and therefore do not sum to the
 188 61 species screened. The class of the copy screened in each species is given in Additional file 5.

189 Among the transcribed species, *Spodoptera exigua* showed the strongest female bias at its *znf-2*-
 190 derived W-linked copy (DESeq2 log₂ fold change = 6.70, $P_{\text{adj}} = 2.6 \times 10^{-13}$) with zero reads in all

three male replicates (Figure 3b). *Chilo suppressalis* showed a 6.7-fold female excess that did not reach significance in this public-data screen (exact $P = 0.060$, limited by the available sample size). Female-specificity of CsFem was subsequently confirmed by qRT-PCR on independently generated samples (Figure 5b) and by CQ analysis of genomic reads (Additional file 1: Figure S5). The remaining transcribed species showed signal without demonstrated female bias (details in Additional file 5).

W-linked loci of diverse structure produce antisense piRNAs targeting masculinizing genes

Transcription is necessary but not sufficient: to function as a feminizing signal, a W-linked transcript must be processed into antisense piRNAs that target and silence masculinizing gene mRNA, thereby redirecting *dsx* splicing toward the female pattern. We therefore profiled embryonic small RNAs and mRNAs in four species: *C. suppressalis* and *S. exigua*, the two species with female-biased W-linked transcription, *S. frugiperda*, a congener of *S. exigua* included for within-genus comparison, and *Pieris rapae*, a butterfly (Pieridae) included to test whether the mechanism extends beyond moths. Small-RNA reads were mapped to *de novo* transcriptome assemblies and processed through the W-linked piRNA Precursor Miner pipeline (Methods, Figure 4a), which integrates ping-pong piRNA detection, transcriptome mapping, and W-chromosome support annotation. We first confirmed that the pipeline recovers the known female-specific antisense piRNAs against *Ofznf-2* at the characterized *OfFem* locus of *Ostrinia furnacalis* (Additional file 1: Figure S6), then applied it to the four focal species. Candidate loci were validated by diagnostic PCR on sex-separated genomic DNA, confirming W-linkage (Figure 4b).



Figure 4. W-linked piRNA precursor miner pipeline and its application to four Lepidoptera species.

(a) Workflow of the W-linked piRNA precursor miner: small-RNA-seq (ping-pong piRNA reads) and transcript-seq (*de novo* Trinity precursors from female-enriched libraries) converge.

(b) PCR and RT-PCR validation in three species: genomic PCR for sex-linked amplification (genetic-sexing control), RT-PCR for female-specific transcription, and autosomal *EF1 α* as the amplification control (assay type and locus per species in the panel labels, primers in Additional file 4). Each track shows lanes cropped from the corresponding gel image (Methods).

219 (c) Small-RNA features at the three loci, one row per species (*C. suppressalis* *CsMasc* / *S.*
220 *frugiperda* *SfMasc* / *S. exigua* *Seznf-2*): sense (blue) and antisense (orange) coverage in RPM along
221 the parental CDS, read-length distribution (25–32 nt), and 5' nucleotide-frequency logos. The 5'-
222 uridine bias on antisense reads and tenth-position adenine bias on sense reads constitute the ping-
223 pong piRNA signature. Coverage, read-length and bits scales are independent per row.

224 (d) Z-linked parent (top), W-linked genomic copy (middle) and piRNA precursor (bottom), one
225 column per species. W-linked copies: two intronless *CsMasc* copies (*C. suppressalis*), a single 5'
226 *SfMasc*-homologous block (*S. frugiperda*) and a *Seznf-2*-derived tandem array (*S. exigua*). *SfFem*
227 embeds a lineage-specific PIF/Harbinger DNA-transposase ORF. *CsFem* carries a retroelement-
228 derived ORF on one flank. *SeFem* comprises ~3 tandem copies of the ~338 bp *znf-2*-derived unit
229 (precursor-transcript span). The primary W-linked array carries 10 units and a second W-linked
230 window 55 kb away carries 3 more (93–98% identity between units). Antisense homology of each
231 precursor was scored by minus-strand BLASTN (Additional file 3). The *CsFem*-flanking ORF is a
232 114-amino-acid non-catalytic fragment of a Pao/BEL-type retroelement, and the *SfFem* transposase
233 ORF is 726 amino acids. The *CsFem* RACE product is 1,912 nt including the poly(A) tail, of which
234 1,889 bp were cloned and used throughout.

235 Three of the four species yielded W-derived antisense piRNAs (Figure 4c, Additional file 1: Figure
236 S7). In *C. suppressalis*, dense *CsMasc*-antisense piRNAs (antisense-to-sense ratio ~42:1 at 48 hpo,
237 Additional file 1: Figure S8) traced to the W-linked *CsMasc* copy. In *S. exigua*, antisense piRNAs
238 against a *znf-2*-derived precursor formed a low-diversity pool dominated by a single 26 nt species
239 (68.2% of antisense reads). In *S. frugiperda*, antisense piRNAs mapped to a 373 bp *Masc*-
240 homologous block within *SfMasc* exon 1. Notably, *S. frugiperda* carries a partial single-exon *Masc*-
241 homologous sequence without tandem repeats that had fallen below the structural-screen threshold

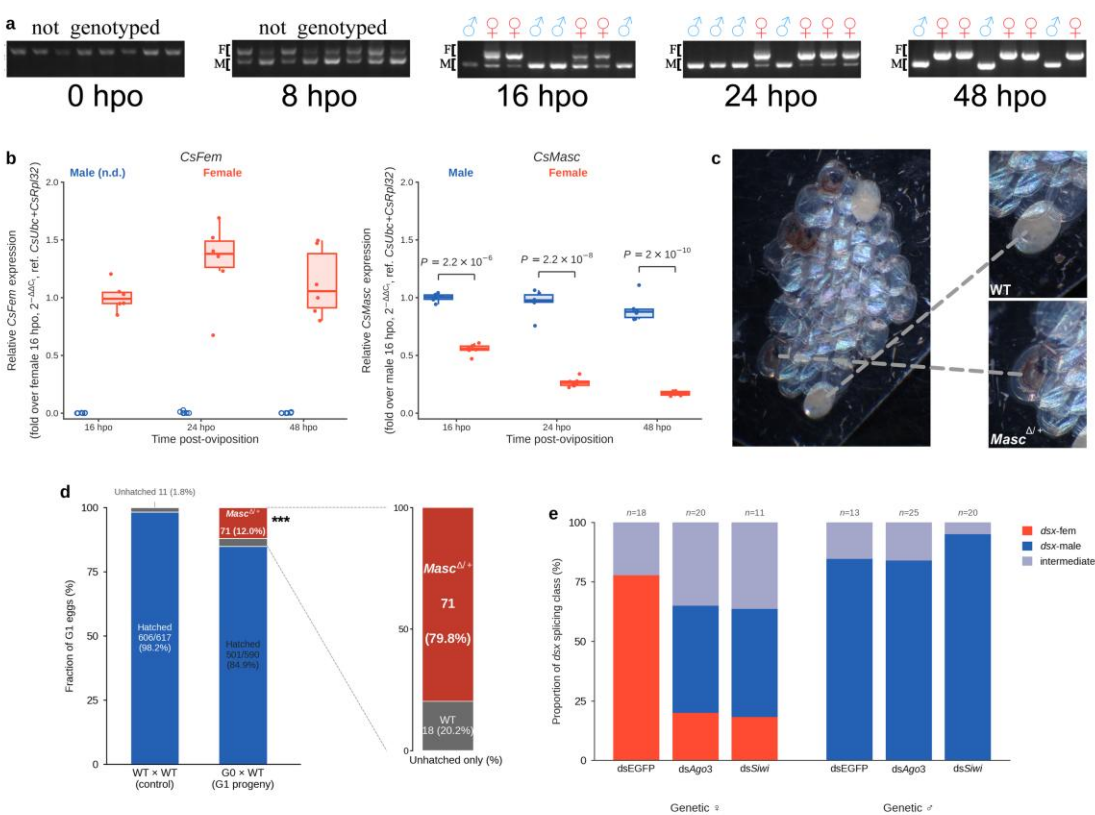
in the genomic survey, yet this locus still produced detectable piRNAs, suggesting that the genomic screen provides a conservative estimate of the reservoir. *P. rapae* was negative. All three positive species display hallmarks of ping-pong amplification: a 26–30 nt size mode, a 10-nt 5' overlap between sense and antisense reads, and a 5'-uridine bias on antisense reads (98.1%, 99.6% and 82.2% in *C. suppressalis*, *S. exigua* and *S. frugiperda*, respectively) with the complementary tenth-position adenine bias on sense reads. We designate the corresponding W-linked loci *CsFem*, *SeFem*, and *SfFem* on the basis of their W-linkage, female-specific expression, and antisense piRNA production.

The three loci differ markedly in architecture (Figure 4d): *CsFem* (~1.9 kb) is an intronless copy whose *Masc*-derived core is flanked on one side by a retroelement-derived ORF, *SeFem* is transcribed from a genomic locus whose primary array carries 10 tandem copies of a ~338 bp *znf-2*-derived repeat unit, of which the precursor transcript spans approximately 3, and *SfFem* (~5.7 kb) consists of a partial single-exon *Masc*-homologous sequence embedded within a PIF/Harbinger transposon transcript whose transposase ORF dominates the locus. Two of the three precursors match structural classes defined in the genomic survey (an intronless multi-exon copy and a short-fragment tandem array), whereas the third is a transposon-embedded partial-exon fragment that fell below the survey threshold, indicating that diverse structural forms, including some the genomic screen does not capture, can serve as substrates for piRNA production. The contrast within *Spodoptera* is particularly informative: *S. frugiperda* produces antisense piRNAs against *Masc* while its congener *S. exigua* targets *znf-2*, indicating independent recruitment of different masculinizing genes after species divergence.

Functional validation supports *CsFem* as a feminizing piRNA precursor in *C. suppressalis*

We carried out functional validation in *C. suppressalis*. CQ (chromosome quotient) analysis of genomic reads confirmed the female-specificity of *CsFem* (Additional file 1: Figure S5). PCR

265 across five embryonic time points (0, 8, 16, 24, 48 h post-oviposition) revealed that *Csdsx* splicing
 266 resolves into a clear female (F) or male (M) pattern by 48 hpo[17] (Figure 5a). The delayed
 267 resolution in females relative to males is consistent with the time required for progressive
 268 accumulation of *CsFem*-derived piRNAs to suppress *CsMasc*.



269
 270 **Figure 5.** *CsFem*, *CsMasc* and piRNA-pathway control of female-type *Csdsx* splicing in *C.*
 271 *suppressalis*.

272 (a) RT-PCR of *Csdsx* splicing in embryos at 0, 8, 16, 24 and 48 hpo. Genetic sexing by the W-
 273 marker primer, autosomal primer as amplification control. One embryo per lane.

274 (b) Sex-resolved *CsFem* (left) and *CsMasc* (right) qRT-PCR in genetically sexed embryos at 16, 24
 275 and 48 hpo ($n = 6$ per sex and stage). *CsFem* was female-specific, undetectable in males (n.d.).

276 *CsMasc* was male-biased, falling in females to $\sim 0.2\times$ the male level by 48 hpo (female versus male
277 at each stage, Welch's two-sided *t*-test on ΔCt , $P = 2.2 \times 10^{-6}$ to 2.0×10^{-10}). Boxes, median and
278 IQR. Points, individual embryos. Female red, male blue.

279 (c) Unhatched wild-type and heterozygous ZZ-male (*CsMasc*-edited) embryos.

280 (d) Hatch rate and genotypes of unhatched *Masc* CRISPR-edited G1 embryos. Seven G0 $\times$ wild-
281 type broods gave 590 eggs, of which 501 hatched (84.9%). Of the 89 unhatched embryos, 71
282 (79.8%) were heterozygous ZZ males (both wild-type and edited *Masc* bands) and 18 were wild
283 type. Hatching was reduced in the *Masc*-edited cross (per-brood mean $82.5\% \pm 4.1\%$ SEM) versus
284 wild-type control ($98.0\% \pm 0.8\%$ SEM), with brood as the replication unit (binomial generalized
285 linear mixed model, $P = 2.8 \times 10^{-5}$, Wilcoxon rank-sum on per-brood rates, $P = 5.8 \times 10^{-4}$, $n = 7$
286 broods per cross). *** $P < 0.001$.

287 (e) *Csdsx* splicing after dsRNA targeting *Ago3* or *Siwi*, scored female-type / male-type /
288 intermediate. Genetic females: dsEGFP 14 / 0 / 4, dsAgo3 4 / 9 / 7, dsSiwi 2 / 5 / 4. Both differed
289 from control (Fisher's exact test, $P = 1.7 \times 10^{-4}$ and 7.7×10^{-4}). Genetic males: dsEGFP 0 / 11 / 2,
290 dsAgo3 0 / 21 / 4, dsSiwi 0 / 19 / 1. Not significant ($P = 1.0$ and $P = 0.55$).

291 Expression dynamics of *CsFem*, *CsMasc*, and *CsFem*-derived piRNAs across the same time course
292 trace three parallel trajectories. *CsFem* transcript rises from 8 hpo and peaks at 24 hpo in females,
293 with no detectable expression in males (Figure 5b, Additional file 1: Figure S9). *CsMasc* mRNA is
294 initially equivalent in both sexes but falls sharply in females from 16 hpo onward, coincident with
295 the rise in *CsFem*-derived piRNAs (Additional file 1: Figure S8) and the appearance of female-type
296 *Csdsx*. The temporal anticorrelation between *CsFem* piRNA accumulation and *CsMasc* decline
297 suggests a causal link.

298 Because *CsMasc* is Z-linked, we tested its role in sex determination by CRISPR–Cas9 disruption.
299 G0 founders injected with Cas9 ribonucleoproteins targeting *CsMasc* were crossed to wild-type
300 adults, so that progeny inheriting both wild-type and edited alleles are ZZ males. Among the G1
301 progeny, hatching fell from a pooled 98.2% to 84.9% (binomial generalized linear mixed model
302 with brood as a random effect, $P = 2.8 \times 10^{-5}$, $n = 7$ broods, Figure 5d), and 71 of 89 unhatched
303 embryos (Figure 5c) carried both wild-type and edited *CsMasc* alleles, identifying them as
304 heterozygous ZZ males. Editing *CsMasc* thus predominantly killed genetic males, indicating that
305 *CsMasc* is required for male development. Independently, injection of dsRNA targeting either
306 *CsAgo3* or *CsSiwi* (two PIWI-clade proteins) within 1 hpo shifted female *Csdsx* splicing toward the
307 male pattern (Figure 5e, Additional file 1: Figure S10), demonstrating that an intact piRNA pathway
308 is required for robust feminization.

309 To confirm the full structure of *CsFem* and enable precise dsRNA design for functional testing, we
310 performed RACE (rapid amplification of cDNA ends), which recovered the full-length *CsFem*
311 transcript. Recovery of 5' and 3' RACE products is consistent with an mRNA-like structure carrying
312 a 5' cap and a 3' poly(A) tail. dsRNA was designed to target the *CsFem*-specific region, avoiding
313 cross-reactivity with *CsMasc*.

314 Injection of dsCsFem within 2 hpo produced four concordant effects: (i) *CsFem* transcript was
315 reduced by 61% (Additional file 1: Figure S10), (ii) *Csdsx* splicing shifted from the female toward
316 the male or intermediate pattern in 6 of 11 injected females (~55%) (Figure 6a,b), (iii) *CsMasc*
317 mRNA rose 2.2-fold in females (Figure 6c), and (iv) *CsMasc*-targeting antisense piRNAs declined
318 by 36% (Figure 6d,e). That all four readouts shift in the predicted direction upon a single
319 perturbation establishes *CsFem* as a feminizing piRNA precursor whose transcription is required to
320 maintain *CsMasc* repression and female-type *dsx* splicing.

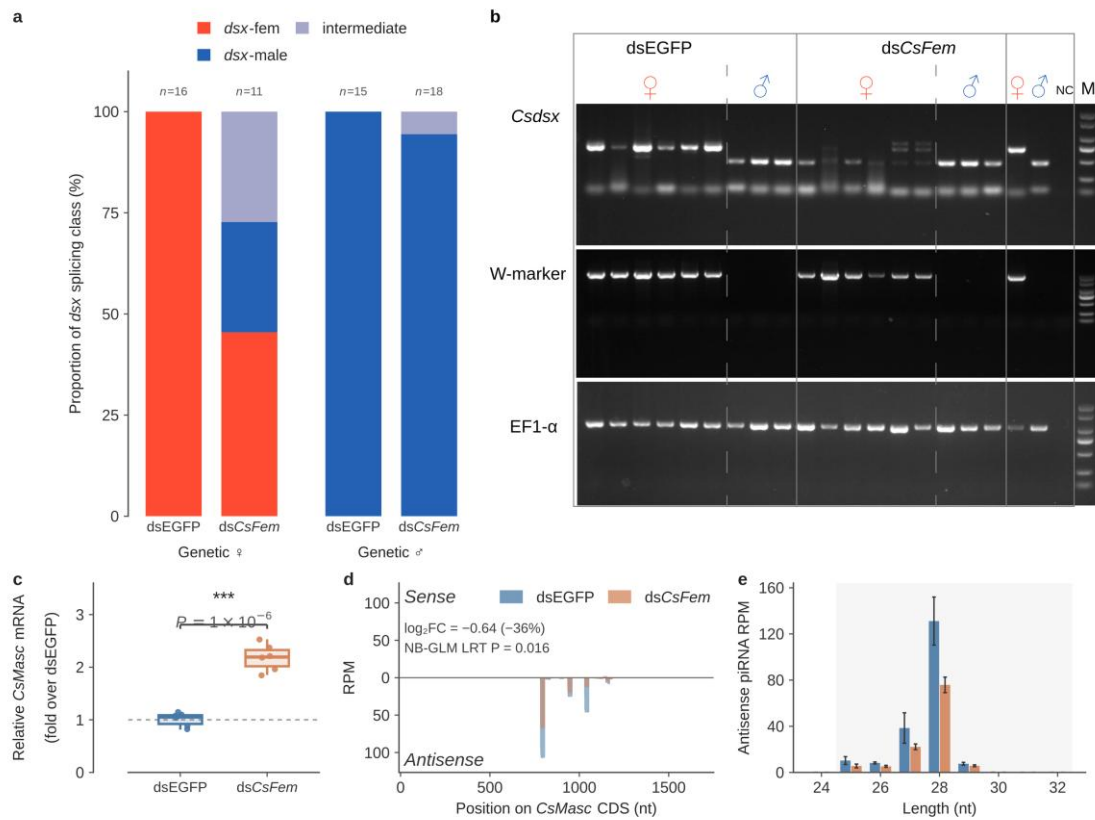


Figure 6. dsCsFem knockdown shifts genetic-female *Csdsc* splicing and derepresses *CsMasc* mRNA in early *C. suppressalis* embryos.

(a) *Csdsc* splicing classes (female-type / male-type / intermediate) after dsCsFem. Genetic females: dsEGFP 16 / 0 / 0, dsCsFem 5 / 3 / 3, shifting toward male-type/intermediate splicing (Fisher's exact test, $P = 0.0016$). Genetic males: dsEGFP 0 / 15 / 0, dsCsFem 0 / 17 / 1. Not significant ($P = 1.0$).

(b) Genotyping and splicing after targeting a non-*Masc*-homologous segment of *CsFem*. The three-track gel reports *Csdsc* splicing, the W-marker and the autosomal control, as in Figure 4b. Gel lanes: ♀ and ♂, uninjected female and male references. NC, no-template control. M, size marker. All lanes of each track are shown in their original order, cropped as a block from the corresponding gel image (Methods).

(c) *CsMasc* transcript in genetic-female embryos 48 h after dsCsFem or dsEGFP injection, by qRT-PCR ($n = 6$ per group). *CsMasc* rose 2.2-fold under dsCsFem (Welch's two-sided t -test on ΔCt , $P = 1.0 \times 10^{-6}$, *** $P < 0.001$). Boxes, median and IQR. Points, individual embryos.

(d) Strand-resolved piRNA coverage along the *CsMasc* CDS after dsEGFP (blue) or dsCsFem (orange) injection. Bars, per-base mean RPM of 25–32 nt reads, sense above and antisense below the zero axis. Antisense $\log_2\text{FC} = -0.64$ (–36%, 95% confidence interval –1.04 to –0.24, negative-binomial GLM likelihood-ratio $P = 0.016$, $n = 3$ unsexed pools).

(e) Read-length distribution of antisense piRNAs on the *CsMasc* CDS (25–32 nt) for dsEGFP and dsCsFem. Mean RPM $\pm$ SEM ($n = 3$).

In Sf9 cells, co-expression of *CsFem* and *CsMasc* reduced *CsMasc* reporter levels (Additional file 1: Figure S11), but the small RNAs produced lacked piRNA hallmarks (size mode, ping-pong signature). These results indicate that Sf9 cells do not recapitulate the embryonic piRNA-processing environment and that *CsFem* transcription *in vivo* likely depends on W-chromosomal chromatin context.

Discussion

Our survey shows that sequences homologous to the masculinizing genes *Masc* and *znf-2* are not incidental features of a few W chromosomes but a widespread component of the lepidopteran W, present in 106 of 266 species across 20 families. This reframes the repeated, independent evolution of feminizing piRNA loci: rather than a series of coincidences requiring a similar rare invention in lineage after lineage, convergence follows from a sequence-level predisposition of the W chromosome itself. Two properties of the W favor this accumulation: its transposon-dense composition provides multiple routes by which gene-derived sequences can be captured, and its

355 suppressed recombination promotes the retention of such sequences once integrated. Crucially, the
356 same transposon density that drives accumulation also creates a piRNA-permissive environment. In
357 *B. mori*, the W chromosome is a major source of female-enriched piRNAs, most of which derive
358 from its resident transposons and function in transposon defense[18]. Gene-derived sequences
359 embedded on such a transposon-rich, piRNA-producing chromosome are situated within a
360 chromatin and processing context that can co-opt them into piRNA production[19], consistent with
361 the non-canonical, dual-strand transcription described for piRNA clusters in *Drosophila*[20–22].
362 The reservoir is thus not only accumulated on the W but positioned within a genomic environment
363 that is inherently permissive to piRNA biogenesis, linking transposon defense and sex
364 determination through a common molecular platform.

365 The structural forms we recover, intronless copies, intron-retaining copies, and short-fragment
366 tandem arrays, are each consistent with a distinct mode of origin. Intronless copies lack parental
367 introns, consistent with reverse transcription of spliced mRNA and reintegration. In some cases,
368 such as *PxyFem*[7], the flanking sequence contains retrotransposon ORF fragments, further
369 supporting this mechanism. Not all intronless copies retain identifiable retrotransposon flanking
370 sequence. Intron-retaining copies preserve parental intron-exon structure with high sequence
371 identity, pointing to DNA-level transfer whose precise mechanism remains unclear. Short-fragment
372 tandem arrays consist of brief parental-homologous units iterated head-to-tail, consistent with local
373 amplification. These categories are not rigidly bounded: *SeFem* in *S. exigua* comprises an intronless
374 multi-exon repeat unit amplified into a tandem array, bridging the first and third forms.

375 The three new precursors we characterize, *CsFem*, *SeFem*, and *SfFem*, illustrate that diverse
376 structural forms can independently give rise to feminizing piRNAs. *CsFem* (~1.9 kb) is an
377 intronless multi-exon copy of *Masc* whose core is flanked on one side by a retroelement-derived

378 ORF. *SeFem* represents a particularly informative case: its repeat unit is an intronless multi-exon
379 segment derived from *znf-2*, yet the locus as a whole is organized as a tandem array of 10 genomic
380 copies in its primary block (of which the precursor transcript spans approximately 3), bridging the
381 intronless and tandem-array categories. *SfFem* (~5.7 kb) takes yet another form: a partial single-
382 exon *Masc*-homologous sequence embedded within a PIF/Harbinger transposon transcript, where
383 the transposase ORF dominates the locus. Despite these structural differences, all three produce
384 antisense piRNAs with ping-pong signatures, demonstrating that the piRNA pathway can process
385 precursors of widely varying architecture. The contrast within *Spodoptera* adds a further dimension:
386 *S. frugiperda* targets *Masc* while its congener *S. exigua* targets *znf-2*, demonstrating that even
387 closely related species can independently recruit different masculinizing genes from the same
388 reservoir after speciation. This multi-path pattern is echoed in previously characterized systems: in
389 *O. furnacalis*, the functional feminizer *OfFem* derives from a tandem array rather than from
390 coexisting pseudogenized intronless copies[8], and in *L. dispar*, two structurally unrelated W-linked
391 loci (*Fet-W* and *LdMasc-W*) coexist and complement each other in suppressing *LdMasc*[12,14].
392 Together, these cases show that the reservoir provides a palette of structural options from which
393 feminizing signals can be assembled through independent, lineage-specific trajectories (Figure 7).

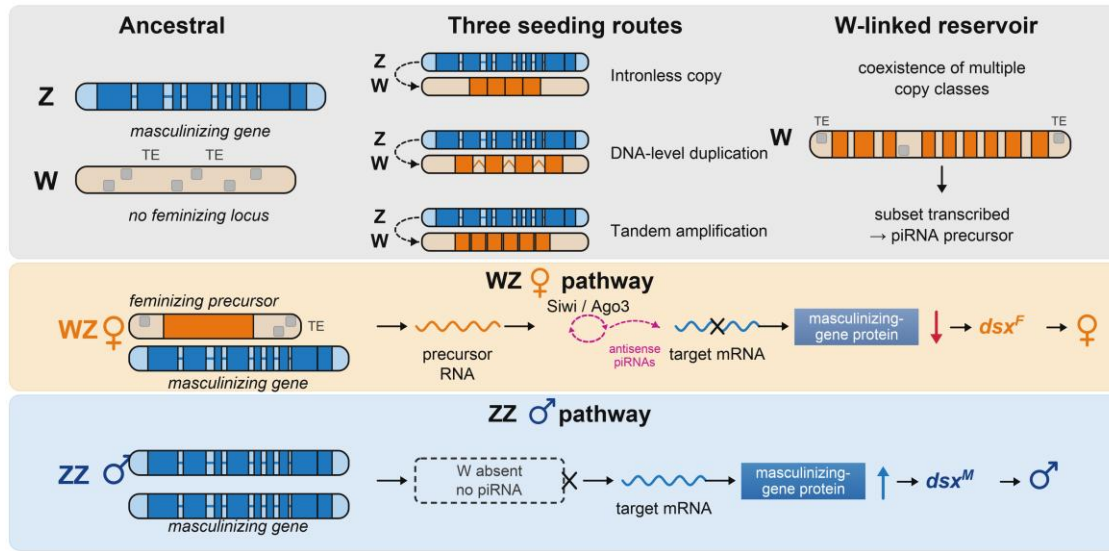


Figure 7. Working model: a W-linked reservoir of masculinizing-gene homologs underlies recurrent feminizing piRNA loci in Lepidoptera.

Ancestrally the W chromosome carries transposable elements but no feminizing locus, while the masculinizing gene (*Masc* or *znf-2*) resides on the Z. Three routes independently seed *Masc*- or *znf-2*-homologous sequence onto the W: retrotransposition of spliced mRNA (intronless copies), DNA-level duplication of genomic segments (intron-retaining copies) and local tandem amplification of exon-derived repeats. These coexisting copy classes form a W-linked reservoir, a subset of which is transcribed into candidate piRNA precursors. In WZ females, a reservoir-derived precursor enters the Siwi/Ago3 ping-pong pathway and generates antisense piRNAs that silence the masculinizing-gene transcript, permitting female-type *dsx* splicing. In ZZ males, the precursor is absent, masculinizing-gene expression is maintained, and male-type *dsx* splicing proceeds. Z, Z chromosome. W, W chromosome. TE, transposable element. *dsx*, *doublesex*. *dsxF* and *dsxM* denote the female- and male-type *doublesex* splice isoforms. Orange, W-linked copy or precursor. Blue, Z-linked masculinizing gene. Beige, W chromosome. Gray, transposable element. Magenta, ping-pong amplification and the antisense piRNAs it produces. X, a blocked step (target silenced in

410 females, piRNA signal absent in males). Downward and upward arrows, reduced and maintained
411 masculinizing-gene protein. ♀ and ♂, female and male.

412 That *Masc* and *znf-2* are the repeatedly co-opted substrates likely reflects their conserved upstream
413 positions in the sex-determination cascade: repression of either gene can shift *dsx* splicing toward
414 the female pattern[3–12], making them high-value targets for any W-borne silencing system. The
415 transposon-chimeric *Fem* of *B. mori*[5,13] demonstrates that sequence complementarity to the
416 masculinizing gene, not derivation from it, is the functional requirement, but duplicating the target
417 gene is the most direct way to acquire such complementarity, which likely explains why target-
418 derived copies predominate. In *Bicyclus anynana*, the W carries pseudogenized *Masc* copies with
419 no detectable transcription, and sex is instead determined by a zygoty-based switch on *BaMasc*:
420 heterozygosity at exons 8–9 drives male development, while hemizygoty (WZ) or homozygosity
421 (ZZ) leads to the female pathway[23]. This mechanism, strikingly convergent with the
422 complementary sex determiner (*csd*) system in the honeybee, shows that the presence of W-linked
423 homologs does not guarantee co-option into feminization. More broadly, all confirmed piRNA-
424 based feminizing systems to date are in moths, and *Pieris rapae*, the one butterfly in our piRNA
425 screen, was negative. Whether butterflies (Papilionoidea) generally rely on alternative mechanisms
426 such as zygoty-based switches or Z:A ratio counting, as demonstrated in *Samia cynthia ricini*[24],
427 rather than W-linked piRNAs, is an open question whose resolution will require broader sampling
428 of butterfly species. Nevertheless, the 106-species reservoir documented here establishes that W-
429 linked masculinizing-gene homologs arise at high frequency across the order, increasing the
430 probability that at least some lineages will co-opt them into piRNA-based feminization.

431 The principle we describe, a target-homologous duplicate on a non-recombining sex-limited
432 chromosome giving rise to small RNAs that silence the parental gene, extends beyond Lepidoptera

433 and beyond piRNAs. In *Diospyros* (persimmon), the Y chromosome carries OGI, a partial duplicate
434 of the feminizing gene *MeGI*, whose transcripts are processed into small RNAs that silence *MeGI*
435 and permit male development[25]. In *Populus* (poplar), Y-linked partial duplicates of *ARR17*
436 produce 24-nt siRNAs that silence the autosomal *ARR17* locus via DNA methylation[26], and in
437 *Salix purpurea* (willow), a ZW system uses W-linked palindromic *ARR17* copies[27], and the W
438 chromosome of the dwarf willow *Salix herbacea* carries PISTILLATA pseudogenes proposed to
439 suppress their parental gene through interfering RNAs[28]. In the persimmon and poplar cases the
440 polarity is inverted (plants silence a feminizing gene), but the logic is identical: a target-derived
441 duplicate on the sex-limited chromosome produces silencing small RNAs. That two kingdoms
442 converge on this architecture through non-homologous small-RNA pathways suggests it is a general
443 feature of how small-RNA-based sex determination originates.

444 Several limitations temper these conclusions. W chromosomes are difficult to assemble and
445 homology-based detection may overlook rapidly diverging copies, so the 106 species recovered
446 here is a lower bound. Functional validation was carried out in *C. suppressalis*. The three piRNA-
447 positive species provide independent correlative support, and extension to additional species is a
448 natural next step. The transcriptome screen relied on public data of uneven tissue and
449 developmental-stage coverage, so the 20% transcription rate is a lower bound whose true value will
450 require systematic profiling. Tandem-array copy-number estimates are sensitive to W-chromosome
451 assembly quality, and the piRNA screen covered only four species, leaving the prevalence of active
452 piRNA production across the reservoir unknown.

453 **Conclusions**

454 Testing whether reservoir density predicts the birth rate of feminizing loci, resolving the origin and
455 turnover of the W across the order[29–34], and determining why some lineages target *znf-2* rather
456 than *Masc* are natural next steps. More broadly, the convergence between Lepidoptera and
457 dioecious plants suggests that the repeat-rich, non-recombining sex-limited chromosome may be a
458 recurrent cradle for the independent origin of small-RNA-based sex determinants across eukaryotes.

459 **Methods**

460 **Genome dataset and W/Z assignment**

461 Chromosome- or scaffold-level assemblies with annotated W chromosomes were retrieved from the
462 International Nucleotide Sequence Database Collaboration (INSDC)/GenBank (256 species),
463 National Center for Biotechnology Information (NCBI) RefSeq (9 species), and the National
464 Genomic Data Center Genome Warehouse (1 species, *Bombyx mori*), giving 266 species across 29
465 families (accessions in Additional file 2). W- and Z-chromosomal scaffold identities were taken
466 from deposited assembly reports and chromosome annotations. Parental coding sequences were
467 extracted per species from the genome annotation with TBtools[35] during genome preparation.

468 **W-linked copy detection and structural classification**

469 For the screen of W chromosomes across 266 genomes, each parental gene was decomposed into
470 per-exon and per-intron query features and aligned with nucleotide Basic Local Alignment Search
471 Tool (BLASTN, BLAST+ v2.16[36], E-value 1×10^{-3} , word size 11) against the W chromosome
472 of the corresponding assembly. High-scoring segment pairs were aggregated per W-linked copy and
473 scored by exon coverage and parental-intron retention. Candidate parental gene models, and contigs

474 for species lacking usable annotation, were resolved beforehand by reciprocal protein BLAST
475 (BLASTP) against previously published reference proteins. For *Plutella xylostella* Masc, the
476 published sex-separated female assembly[7] was screened because the standard assembly lacks an
477 assembled W. Each retained W-linked copy was classified as a retrocopy or a DNA-level
478 duplication by intron-homology testing. The parental gene for each copy was identified by
479 alignment to the reference genome with W scaffolds excluded. Parental coding sequence was then
480 mapped to the W-linked copy by splice-aware alignment (minimap2 v2.28[37]), and sequence
481 between aligned exon blocks was compared to the corresponding parental introns by BLASTN (E-
482 value $\leq 1 \times 10^{-3}$). An exon was scored as present when its query coverage reached 50%, only
483 introns falling between the first and last present exon were evaluated for retention. A copy in which
484 any in-scope intron was retained (coverage $\geq 45\%$) was classified as a DNA-level copy, full if all
485 parental exons were present, partial otherwise. If in-scope intron coverage fell in the 15–45% range,
486 the copy was flagged as ambiguous. A copy carrying two or more present exons in a contiguous run
487 but no retained intron was classified as a retrocopy. Short copies in which no single exon reached
488 the 50% coverage threshold were rescued as retrocopies when the best exon segment showed $\geq 85\%$
489 identity to the parent and maximum intron coverage stayed below 15%, this rescue applied directly
490 to single-exon parents, while multi-exon parents additionally required the best exon to reach 50%
491 coverage. Copy-level calls were aggregated per species and gene: a combination was scored as
492 retrocopy-positive if any copy was classified as a retrocopy, DNA-copy-positive if any was a DNA-
493 level copy, and MIXED if both were present. Combinations in which all copies were ambiguous,
494 single-exon-only, or undetermined were not counted as retrocopy-positive. Full classifier code and
495 parameters are in the deposited repository (GitHub and Zenodo, see the Data Availability
496 Statement).

497 **Species tree and gene-tree phylogenetics**

498 For the 266-taxon species tree (Figure 1b), single-copy orthologs were identified per assembly with
499 BUSCO (lepidoptera_odb12[38]), and the top 500 BUSCO families ranked by species coverage
500 were retained. Per-gene protein alignments were generated with MAFFT v7.525[39] (--auto, 20
501 threads) and concatenated into a supermatrix. A maximum-likelihood tree was inferred with IQ-
502 TREE2[40] under the LG+G4 model, applying a topological constraint that enforced the family-
503 level backbone of Kawahara et al. (2019)[41]. The tree was rooted on *Ectoedemia*
504 *decentella*(Nepticulidae) and *Tischeria ekebladella* (Tischeriidae). Figure 1b shows this tree as a
505 compact circular phylogeny summarizing the per-species screening calls. For the Figure 2 gene
506 trees, one tree was built per gene and structural class. Intronless (retrocopy) trees used the parental
507 coding sequences of the 27 *Masc*-positive and 17 *znf-2*-positive species together with their W-
508 linked retained blocks from the screening output. Intron-retaining copies were first projected onto
509 exon sequence by aligning each copy to its own parental coding sequence with BLASTN and
510 concatenating non-overlapping high-scoring segment pairs in query order. Tandem-array units were
511 extracted with the parameters of the array screen (BLASTN-short, word size 7, E-value 1, dust off,
512 identity $\geq 80\%$, length ≥ 20 bp) and each unit was realigned to the parental coding sequence for
513 verification. Within each species and gene, W-linked sequences were clustered at $\geq 95\%$ nucleotide
514 identity with CD-HIT-EST[42] while parental sequences were left unclustered, because clustering
515 parents together merges the cross-species similarity that the test evaluates. The member with
516 highest identity to the parent was the display tip, and cluster size is the $\times N$ label used in Additional
517 file 1: Figures S2 to S4. Per-gene alignments were generated with MAFFT v7.525 (--localpair --
518 maxiterate 1000, L-INS-i for the intronless sets, --auto for the larger DNA-copy and tandem sets)
519 and trimmed with trimAl (-gappyout). Maximum-likelihood phylogenies were inferred with IQ-

TREE2[40] (v2.2.2.7) under ModelFinder Plus[43] (-m MFP, BIC) with 1,000 ultrafast bootstrap replicates[44] (-B 1000, UFBoot $\geq$ 95 = well-supported). The intron-retaining and tandem-array trees were rooted on an outgroup outside the ingroup of that gene, *Plutella xylostella* for *Masc* and *Argyresthia albistria* for *znf-2*, each confirmed as an ortholog of the intended gene by reciprocal tblastx before use. The two intronless trees were midpoint-rooted. Species assignments under the exclusivity criterion were identical under outgroup and midpoint rooting. A species was scored as exclusive when its parental coding sequence and all of its W-linked copies formed a clade containing no sequence from another species. For Figure 2g, the parental coding-sequence interval covered by each W-linked record was expressed as a relative position along the parent. Figure 1b reports raw W-linked copy counts, whereas Figure 2 uses clustered representatives. Trees were drawn in R v4.4.3 with ape v5.8[45] and ggtree[46].

W-linkage confirmation

For the full positive set, W-linkage was assigned from existing chromosome annotations: each candidate scaffold was matched to the chromosome-assignment map of its source assembly, and copies on a W-designated scaffold were scored as W-linked (Additional file 2). Experimental confirmation by PCR was performed for the three functionally profiled species (*C. suppressalis*, *S. frugiperda*, *S. exigua*): genomic DNA from individually sexed adults gave female-specific amplification of the candidate locus, and RT-PCR on sex-separated cDNA gave female-specific precursor transcription (Figure 4b, primers in Additional file 4). Primers were designed from flanking W-chromosomal sequence with NCBI Primer-BLAST[47] and amplified with 2 $\times$ Taq Master Mix (Vazyme). For *C. suppressalis*, W identity was confirmed independently by the chromosome quotient (CQ) method[48]. Male and female adults were resequenced. Raw reads were quality-filtered with SOAPnuke v2.2.6[49] (filter -n 0.001 -l 10 -q 0.5 -Q 2), aligned to the *C.*

543 *suppressalis* assembly with BWA-MEM v0.7.19 (bwa mem -M -Y), and sorted and duplicate-
544 marked with SAMtools v1.22. CQ was computed per chromosome as the male-to-female ratio of
545 normalized read depth, scaffolds with CQ < 0.5 were designated W-derived (Additional file 1:
546 Figure S5).

547 **Sample collection, RNA extraction and sequencing**

548 Individual *C. suppressalis* eggs were isolated from egg masses by 1 M KOH (12 min) and PBS
549 wash (3 min) adapted from[50], and collected at 0, 8, 16, 24 and 48 hours post-oviposition (hpo) at
550 27 °C. Individual embryos can be molecularly sexed only from 16 hpo, when W-linked genomic
551 DNA is amplifiable. From 16 hpo, the genomic DNA and RNA of each embryo were co-extracted
552 with the MiniBEST Universal RNA Extraction Kit (TaKaRa). The DNA fraction was PCR-
553 genotyped with W-linked primers, using the autosomal *EFl α* primer as the genomic-DNA
554 amplification control, and the RNA was used for expression. Genotype-resolved analyses are
555 therefore reported for 16, 24 and 48 hpo. Single-embryo cDNA synthesis and pre-amplification
556 used the Single Cell Sequence Specific Amplification Kit (Vazyme). Bulk RNA was isolated with
557 TRIzol (Vazyme). All PCR amplifications used 2× Taq Master Mix (Vazyme). Total RNA from
558 pooled eggs (3 biological replicates per time point) at 0, 8, 16, 24 and 48 hpo was used for
559 sequencing. Small-RNA libraries were prepared with the MGIEasy Small RNA Library Prep Kit
560 (polyacrylamide gel electrophoresis (PAGE) size selection) and sequenced single-end 50 bp on
561 DNBSEQ-T7 (BGI Genomics). Paired-end 150 bp mRNA-seq libraries were prepared in parallel
562 with the BGI Optimal mRNA Library Prep Kit (oligo(dT) enrichment). Raw data are deposited in
563 the China National Center for Bioinformation (CNCB) Genome Sequence Archive under BioProject
564 PRJCA058350 (GSA accession CRA038921). Small-RNA reads were quality- and adapter-trimmed
565 with Trim Galore v0.6.10 (Cutadapt[51]) using the BGI DNBSEQ 3' adapter and a 25–32 nt

566 window (-a AGTCGGAGGCCAAGCGGTCTTAGGAAGACAA --length 25 --max_length 32 --
567 quality 20 --stringency 3). This 25–32 nt window defines the piRNA size range used throughout
568 (trimming, ping-pong, per-stage antisense counts, specificity, and the dsCsFem analysis).

569 **Small-RNA mapping and piRNA identification**

570 Trimmed reads were mapped to the *CsMasc* CDS with Burrows–Wheeler Aligner (BWA)
571 v0.7.19[52] (bwa aln at -n 0, -n 1 and -n 2 to test sensitivity to the mismatch threshold, with the
572 formal call for each species using the setting recorded in Additional file 3, which accommodates
573 sequence divergence between W-linked retrocopy-derived piRNAs and the Z-linked parental CDS
574 reference), filtered by mapping-quality score (MAPQ, samtools view -q 20, MAPQ $\geq$ 20), and
575 sorted and indexed with SAMtools v1.22[53]. Reads in the 25–32 nt window were classified as
576 candidate piRNAs. Per-stage, strand-resolved coverage, read-length and 5' motif profiles across the
577 0–48 hpo series are in Additional file 1: Figure S8. Strand-specific coverage was computed in reads
578 per million (RPM), with each library's total clean 25–32 nt read count as the denominator. Position
579 1–15 nucleotide composition was visualized with WebLogo 3[54]. Each species was mapped at
580 zero, one and two mismatches, and Additional file 3 reports the recovered antisense and ping-pong
581 counts at every setting together with the setting carried through to the formal call for that species, so
582 the specificity of target assignment and its sensitivity to the mismatch threshold can be read directly.
583 Polyadenylated mRNA libraries from *C. suppressalis*, *O. furnacalis*, *S. frugiperda*, *S. exigua* and *P.*
584 *rapae* eggs were assembled *de novo* with Trinity v2.15.2[55] (--SS_lib_type RF), with reads
585 trimmed by Trim Galore v0.6.10 as above. piRNA precursors were identified with the W-linked
586 piRNA precursor miner: (1) antisense piRNA-sized reads (25–32 nt) were extracted from BWA-
587 aligned BAM files, (2) ping-pong pairs were detected by matching antisense 5' positions to sense 5'
588 positions at a –10 bp offset with a 10-nt seed-complementarity check[56], (3) piRNAs were aligned

589 to the *de novo* transcriptome with BLASTN-short (BLAST+ v2.16, ungapped, plus strand only, E-
590 value ≤ 1000), (4) hits were filtered to ≤ 2 mismatches and alignment length $\geq$ query length - 2, (5)
591 candidate transcripts were ranked by piRNA coverage ratio (piRNA-mapping positions / transcript
592 length) and those $\geq 10\%$ retained, (6) retained candidates were aligned to the reference genome with
593 BLASTN (pident > 95%, qcovs > 50%) for W-chromosomal support, (7) candidates were classified
594 as *best_hit_on_W*, *W_supported_not_top_hit* or *no_W_support*.

595 RACE

596 The full-length *CsFem* precursor was determined by 5' and 3' RACE (HiScript-TS 5'/3' RACE Kit,
597 Vazyme) from total RNA of female *C. suppressalis* eggs (48 hpo). Products were cloned into
598 pTOPO-Blunt (Zero Background pTOPO-Blunt Simple Cloning Kit, Vazyme) and Sanger-
599 sequenced across independent clones per end (RACE primers in Additional file 4). A 5' cap and
600 poly(A) tail were inferred from recovery of 5' RACE products and 3' poly(T)-primed products,
601 respectively.

602 Transcript quantification

603 For bulk mRNA-seq, pooled embryo libraries (0, 8, 16, 24, 48 hpo, 3 biological replicates each)
604 were assembled with Trinity v2.15.2 (--SS_lib_type RF) and quantified per sample with Salmon
605 v1.11.4[57]. BLAST against the *CsMasc* CDS identified *CsFem* precursor and *CsMasc* target
606 isoforms. Per-sample counts were aggregated across BLAST-confirmed isoforms. Differential
607 expression of the aggregated *CsFem* and *CsMasc* features (each stage versus 0 hpo) was assessed
608 with DESeq2 (median-of-ratios normalization, empirical-Bayes dispersion across the full
609 transcriptome, Wald test), edgeR quasi-likelihood F-tests and a negative-binomial GLM
610 (MASS::glm.nb) with a trimmed mean of M-values (TMM)-normalized library-size offset returned

611 concordant estimates. The compact letter display in Additional file 1: Figure S9 derives from all-
612 pairwise comparisons among stages on the negative-binomial GLM (emmeans, multiplicity-
613 adjusted). For qRT-PCR, total RNA was extracted from individual embryos with TRIzol, each
614 biological replicate being a single embryo from one synchronized clutch. cDNA was synthesized
615 with the Hifair III 1st Strand cDNA Synthesis Kit (gDNA digester plus, YEASEN) and qRT-PCR
616 run with Hieff qPCR SYBR Green Master Mix (No ROX, YEASEN) on a CFX96 system (Bio-
617 Rad), two technical replicates per sample. *CsUbc* and *CsRpl32* were reference genes. Targets were
618 normalized to their geometric mean and expressed as $2^{-\Delta\Delta Ct}$. The *CsFem* qPCR primers Fem-q-
619 F1/Fem-q-R1 amplify a 142 bp region outside the 480 bp *CsMasc*-homologous core of *CsFem*,
620 *CsFem* primer specificity was confirmed by BLASTN against the *CsMasc* CDS (no match ≥ 15 bp,
621 $E \leq 1 \times 10^{-3}$). All qPCR primers are in Additional file 4.

622 **RNA interference (RNAi) and *CsFem*-precursor knockdown**

623 Double-stranded RNAs targeting *CsSiwi*, *CsAgo3* and *EGFP* (negative control) were synthesized by
624 *in vitro* transcription with the MEGAshortscript T7 Transcription Kit (Thermo Fisher), purified by
625 LiCl precipitation, quantified by agarose-gel densitometry against a DNA mass standard (Image
626 Lab, Bio-Rad), and adjusted to 500 ng/ μ L. Freshly laid *C. suppressalis* eggs (0–1 hpo) were
627 microinjected with 1 nL dsRNA per egg using a pneumatic microinjector (PV850, World Precision
628 Instruments), maintained at 27 °C and harvested at 48 hpo. *Csdsx* splicing was assessed by RT-PCR
629 with primers spanning the sex-specific exons. Each embryo was scored as female-type, male-type or
630 intermediate by the presence of the corresponding splice-product bands. Reported proportions are
631 the fraction of embryos in each scoring class per treatment. To target the *CsFem* precursor itself,
632 dsRNA was designed against a non-*Masc*-homologous region of the 1,889 bp *CsFem* transcript (a
633 220 bp amplicon at precursor positions 1,661–1,880), the 480 bp *CsMasc*-homologous core was

634 excluded from the template (primers dsFem-F/dsFem-R, Additional file 4). Absence of *CsMasc*
635 homology was confirmed by BLASTN against the *CsMasc* CDS (no hit at $E \leq 1 \times 10^{-3}$, longest
636 primer match 9 bp). Synthesis, purification, 500 ng/ μ L working concentration and 1 nL/egg
637 microinjection followed the protocol above. Eggs were injected within 2 hpo, held at 27 °C and
638 harvested at 48 hpo. *Csdsx* splicing was scored on individual genetically sexed embryos with the
639 same primers and scoring criteria, with dsEGFP as control. For small-RNA-seq of knockdown
640 embryos, a parallel batch of freshly laid mixed-sex egg pools was injected within 2 hpo with
641 dsCsFem or dsEGFP ($n = 3$ biological replicates per group, ~500 eggs each) and harvested at 48
642 hpo for small-RNA-seq. Total RNA was extracted with TRIzol, small-RNA libraries were
643 sequenced on DNBSEQ (BGI Genomics, six libraries total). Reads were trimmed with Trim Galore
644 v0.6.10 using the same adapter and quality parameters as described above. Reads of 25–32 nt were
645 retained and mapped to the intron-free spliced *CsMasc* CDS (1,743 bp, mRNA-sense) with BWA
646 v0.7.19 (bwa aln -n 2). Alignments were filtered to primary records at MAPQ ≥ 20 (samtools view -
647 bF 256 -q 20, SAMtools v1.22) and split by strand, the antisense pool corresponds to *CsFem*-
648 precursor-derived piRNAs complementary to *CsMasc* (Figure 4c). Per-base coverage of 25–32 nt
649 reads was computed with samtools depth -aa and converted to RPM using each library's total clean
650 25–32 nt read count (the same per-library piRNA-window normalizer used for Figure 4c and
651 Additional file 1: Figure S8). As a specificity control, the 400 nt EGFP dsRNA template was
652 queried by BLASTN against *CsMasc*, *Csznf-2* and the *CsFem* precursor ($E \leq 1$, word size 7) with
653 no significant similarity, and read names of dsRNA-aligning reads showed empty overlap with
654 *CsMasc*-mapped reads across all six libraries.

655 **CRISPR–Cas9 disruption of *CsMasc***

656 Four guide RNAs targeting *CsMasc* were designed with sgRNAcas9[58] (sequences in Additional
657 file 4) and synthesized by *in vitro* transcription with the MEGAscript T7 Transcription Kit
658 (Thermo Fisher). Cas9 protein (TrueCut v2, Thermo Fisher) and sgRNA were assembled into
659 ribonucleoprotein complexes (Cas9 300 ng/μL, sgRNA 150 ng/μL final) and microinjected (~1 nL
660 per egg) into 0–1 hpo eggs (PV850, World Precision Instruments). G0 mosaic founders were
661 crossed to wild type, and seven G0 × wild-type broods (590 eggs total) were analyzed: hatching was
662 scored per brood and unhatched embryos genotyped at the *CsMasc* target site by a PCR-based assay
663 distinguishing mutant from wild-type alleles. Of 590 eggs, 501 hatched (84.9%) and 89 did not.
664 Among the unhatched, 71 of 89 carried a *CsMasc* mutant allele (Figure 5c,d). Hatch rate was
665 compared between the edited and wild-type crosses with brood as the unit of replication, using a
666 binomial generalized linear mixed model with brood as a random effect and a Wilcoxon rank-sum
667 test on per-brood rates.

668 **Sf9 reporter assays**

669 A dual-luciferase reporter system was built in Sf9 cells (*Spodoptera frugiperda* pupal ovary-
670 derived). The *CsMasc* CDS carrying four predicted piRNA target sites was fused to Firefly
671 luciferase in pIB (pIB-Masc-Fluc), alongside a mutant (pIB-MutMasc-Fluc) with point mutations at
672 all four sites. The full-length *CsFem* precursor (RACE, 1,889 bp) was cloned into pLex-His (pLex-
673 Fem), empty pLex-His was the negative control, and Renilla reporter pIB-Rluc was the
674 normalization control. Four conditions were tested ($n = 6$ biological replicates each): wild-type
675 (WT)-Masc-Fluc + pLex-His, WT-Masc-Fluc + pLex-Fem, mutant (MUT)-Masc-Fluc + pLex-His,
676 and MUT-Masc-Fluc + pLex-Fem. Transfection used Lipofectamine 3000 (Thermo Fisher), cells
677 were harvested 48 h post-transfection and Firefly/Renilla ratios measured with the Firefly & Renilla

Luciferase Reporter Assay Kit (Yeasen). For small-RNA profiling, total small RNA was extracted from three transfection conditions: empty vector, WT-Masc-Fluc + pLex-Fem, and MUT-Masc-Fluc + pLex-Fem. Libraries were sequenced on DNBSEQ (BGI Genomics, ~25 million reads per sample), trimmed with Trim Galore v0.6.10 and mapped to the *CsMasc* CDS with BWA v0.7.19 (bwa aln -n 2), read-length distributions and ping-pong signatures were analyzed as above. For promoter activity, truncated *CsFem* upstream regions (2,119, 1,869, 1,669, 1,269, 1,069, 719 and 519 bp) were cloned upstream of Firefly luciferase in pGL4.10 (Promega), co-transfected with a Renilla normalization control into Sf9 cells ($n = 6$ biological replicates each), and Firefly/Renilla ratios measured 48 h post-transfection against an empty pGL4.10 control.

Gel imaging and image processing

Agarose gels were stained with nucleic-acid dye and imaged on a gel documentation system. The panels in Figures 4b, 5a and 6b show lanes cropped from the corresponding gel images, with each track (splicing assay, W marker and autosomal control) taken from its own gel or exposure and assembled for presentation. Brightness and contrast were adjusted linearly across each whole image, and no local adjustment, gamma correction or lane rearrangement was applied. Uncropped images of every gel shown are provided in Additional file 1: Figures S12 to S14.

Statistical analysis

All analyses were performed in R v4.4.3. Data are mean $\pm$ standard error of the mean (SEM) unless noted. All tests were two-sided. The test used (Welch's t -test, Fisher's exact test, or, for hatch rate, a binomial generalized linear mixed model with brood as a random effect and a Wilcoxon rank-sum test on per-brood rates, as appropriate), exact P value, sample size (n) and relevant summary statistics for each comparison are reported in the corresponding figure legend. Quantitative PCR

700 comparisons were made on ΔCt ($\log_2$ -scale) values. The dsCsFem effect on *CsMasc*-targeting
701 antisense piRNAs was assessed by a single-target negative-binomial GLM (MASS::glm.nb) on raw
702 antisense read counts with a TMM-normalized library-size offset, taking the likelihood-ratio P as
703 the primary statistic.

704 **Use of generative artificial intelligence**

705 Generative artificial intelligence was used to check language and formatting of the manuscript text.
706 It was not used to generate, analyze or interpret data, and the authors take full responsibility for the
707 content.

708 **Abbreviations**

709 dsx, doublesex

710 Masc, Masculinizer

711 piRNA, PIWI-interacting RNA

712 RACE, rapid amplification of complementary DNA ends

713 RPM, reads per million

714 TE, transposable element

715 znf-2, zinc finger protein 2

716 **Declarations**

717 **Ethics approval and consent to participate**

718 This work used insects (*Chilo suppressalis*, *Spodoptera exigua*, *Spodoptera frugiperda* and *Pieris*
719 *rapae*), the *S. frugiperda*-derived Sf9 cell line, and public sequence data. No vertebrate or human
720 material was involved, and research on these invertebrates requires no ethical approval under the
721 applicable national and institutional regulations. All laboratory work followed the biosafety rules of
722 Hunan Agricultural University, and insect cultures were contained throughout.

723 **Consent for publication**

724 Not applicable.

725 **Availability of data and materials**

726 Raw small-RNA and mRNA data generated in this study for *C. suppressalis*, *S. exigua*, *S.*
727 *frugiperda* and *P. rapae*, together with male and female genomic resequencing data for *C.*
728 *suppressalis*, are deposited in the CNCB Genome Sequence Archive under BioProject
729 PRJCA058350 (accession CRA038921). Public datasets reanalyzed here are NCBI BioProject
730 PRJDB19674 [8] for *O. furnacalis* small-RNA sequencing and PRJDB14599 for *O. furnacalis* egg
731 RNA-seq. The 266 genome assemblies screened in this study are listed by accession in Additional
732 file 2 and were obtained from INSDC/GenBank, NCBI RefSeq or the National Genomic Data
733 Center Genome Warehouse as indicated. Additional files 2 to 5 provide the genome-screening calls,
734 W-linked copy coordinates and W-assembly metrics (Additional file 2), the target-specificity and
735 mismatch-sensitivity summary (Additional file 3), the primer, guide and dsRNA sequences
736 (Additional file 4), and the complete 61-species W-copy transcription screen, with per-species calls
737 and the public sequencing accessions used (Additional file 5). The parental Masc and znf-2 coding

sequences and translated proteins used for the W-linked copy screen are included in the Zenodo deposit (see below). Numerical data underlying the quantitative panels are provided in the Zenodo deposit alongside the analysis code. Custom scripts used for W-linked copy screening and precursor mining are archived on Zenodo (DOI: 10.5281/zenodo.19646981). The record is under controlled access during peer review, a reviewer access link is provided with this submission, and the record will be released publicly upon publication. Plasmid constructs generated in this study (pIB-Masc-Fluc, pIB-MutMasc-Fluc, pLex-Fem, and the CsFem promoter-truncation series in pGL4.10) are available from the corresponding authors upon request.

Competing interests

The authors declare that they have no competing interests.

Funding

National Key R&D Program of China grant 2024YFD1400900 (L.Q.)

National Natural Science Foundation of China grant 32472557 (L.Q.)

Science and Technology Innovation Program of Hunan Province grant YLS-2025-ZY04018 (L.Q.)

Teaching and Research Project of Hunan Agricultural University grant 25KJ031 (Yo.L.)

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

Conceptualization: S.K., L.Q. Methodology: S.K., W.D., H.H. Investigation: S.K., Yu.L., Y.Z., Y.G., Y.Q., Z.Y., D.J. Formal analysis: S.K., Z.Y., Yu.L., Y.Z. Funding acquisition: L.Q., Yo.L.

758 Writing, original draft: S.K., W.D., H.H., H.G., Q.G., Yo.L., L.Q. Writing, review and editing: all
759 authors. Supervision: L.Q., Yo.L. All authors read and approved the final manuscript.

760 **Acknowledgements**

761 We thank Jichong Zhuo (Ningbo University), Wei Peng (Hunan Normal University), and František
762 Marec (Czech Academy of Sciences) for helpful discussions. This study made use of publicly
763 available genome assemblies deposited in INSDC/GenBank, NCBI RefSeq, and the National
764 Genomic Data Center Genome Warehouse.

765 **References**

- 766 1. Wilkins AS. Moving up the hierarchy: a hypothesis on the evolution of a genetic sex
767 determination pathway. *BioEssays*. 1995;17:71–77.
- 768 2. Verhulst EC, van de Zande L. Double nexus: Doublesex is the connecting element in sex
769 determination. *Brief Funct Genomics*. 2015;14:396–406.
- 770 3. Marec F, Yoshido A, van't Hof AE. Molecular mechanisms of sex determination in Lepidoptera:
771 current status and perspectives. *Insect Sci*. 2026;33:599–617.
- 772 4. van't Hof AE, Yoshido A, Marec F. Sex determination in moths and butterflies: Masculinizer as
773 key player. *Curr Opin Insect Sci*. 2025;70:101375.
- 774 5. Kiuchi T, Koga H, Kawamoto M, Shoji K, Sakai H, Arai Y, Ishihara G, Kawaoka S, Sugano S,
775 Shimada T, Suzuki Y, Suzuki MG, Katsuma S. A single female-specific piRNA is the primary
776 determiner of sex in the silkworm. *Nature*. 2014;509:633–636.
- 777 6. Kaneda T, Matsuda-Imai N, Kosako H, Shoji K, Suzuki MG, Suzuki Y, Kiuchi T, Katsuma S.
778 The Masc–PSI complex directly induces male-type doublesex splicing in silkworms. *Commun Biol*.
779 2025;8:927.

780 7. Harvey-Samuel T, Xu X, Anderson MAE, Carabajal Paladino LZ, Purusothaman D, Norman VC,
781 Reitmayer CM, You M, Alphey L. Silencing RNAs expressed from W-linked PxyMasc “retrocopies”
782 target that gene during female sex determination in *Plutella xylostella*. *Proc Natl Acad Sci USA*.
783 2022;119:e2206025119.

784 8. Fukui T, Muro T, Matsuda-Imai N, Kaneda T, Kosako H, Hiraki H, Shoji K, Fujii T, Suzuki Y,
785 Toyoda A, Itoh T, Kiuchi T, Katsuma S. Complete transition from chromosomal to cytoplasmic sex
786 determination during prolonged *Wolbachia* symbiosis. *Nat Commun*. 2026;17:104.

787 9. Gopinath G, Arunkumar KP, Mita K, Nagaraju J. Role of Bmznf-2, a *Bombyx mori* CCCH zinc
788 finger gene, in masculinisation and differential splicing of Bmtra-2. *Insect Biochem Mol Biol*.
789 2016;75:32–44.

790 10. Tomihara K, Katsuma S, Kiuchi T. Transcriptome analysis in the silkworm *Bombyx mori*
791 overexpressing piRNA-resistant Masculinizer gene. *Biochem Biophys Res Commun*.
792 2022;616:104–109.

793 11. Visser S, Yoshido A, Provazníková I, Dalíková M, Voříšková D, Chung Voleníková A,
794 Verhulst EC, Marec F. A W chromosome-derived feminizing piRNA in pyralid moths demonstrates
795 convergent evolution for primary sex determination signals in *Lepidoptera*. *BMC Biol*. 2025;23:289.

796 12. Moronuki Y, Kasahara R, Naka H, Suzuki MG. Identification and functional analysis of sex-
797 determining genes in the spongy moth, *Lymantria dispar* (*Lepidoptera*: *Erebidae*). *Insect Biochem*
798 *Mol Biol*. 2025;177:104219.

799 13. Lee J, Fujimoto T, Yamaguchi K, Shigenobu S, Sahara K, Toyoda A, Shimada T. W
800 chromosome sequences of two bombycid moths provide an insight into the origin of fem. *Mol Ecol*.
801 2024;33:e17434.

802 14. Shoji K, Ishida K, Kasahara R, Naka H, Suzuki MG. W chromosome-specific paralogs of the
803 male-determining gene *LdMasc* exhibits a female-determining ability in the spongy moth,
804 *Lymantria dispar*. Research Square. 2026. <https://doi.org/10.21203/rs.3.rs-8963102/v1>. Preprint.

805 15. Bachtrog D. Y-chromosome evolution: emerging insights into processes of Y-chromosome
806 degeneration. *Nat Rev Genet*. 2013;14:113–124.

807 16. Zrzavá M, Hladová I, Dalíková M, Šíchová J, Ůnap E, Kubíčková S, Marec F. Sex
808 chromosomes of the iconic moth *Abraxas grossulariata* (Lepidoptera, Geometridae) and its
809 congener *A. sylvata*. *Genes*. 2018;9:279.

810 17. Zhu Y, Kuang S, Yao Z, Sun Y, Gao H, Gao Q, Ding W, He H, Li Y, Qiu L. Doublesex
811 knockout via CRISPR/Cas9 disrupts fertility and sexual dimorphism of wings in the rice stem borer,
812 *Chilo suppressalis*. *Insect Biochem Mol Biol*. 2025;182:104356.

813 18. Kawaoka S, Kadota K, Arai Y, Suzuki Y, Fujii T, Abe H, Yasukochi Y, Mita K, Sugano S,
814 Shimizu K, Tomari Y, Shimada T, Katsuma S. The silkworm W chromosome is a source of female-
815 enriched piRNAs. *RNA*. 2011;17:2144–2151.

816 19. Katsuma S, Kawamoto M, Kiuchi T. Guardian small RNAs and sex determination. *RNA Biol*.
817 2014;11:1238–1242.

818 20. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ. Discrete
819 small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell*.
820 2007;128:1089–1103.

821 21. Mohn F, Sienski G, Handler D, Brennecke J. The Rhino–Deadlock–Cutoff complex licenses
822 noncanonical transcription of dual-strand piRNA clusters in *Drosophila*. *Cell*. 2014;157:1364–1379.

823 22. Andersen PR, Tirian L, Vunjak M, Brennecke J. A heterochromatin-dependent transcription
824 machinery drives piRNA expression. *Nature*. 2017;549:54–59.

825 23. van't Hof AE, Whiteford I, Yung CJ, Yoshido A, Zrzavá M, de Jong MA, Tan K-L, Zhu D,
826 Monteiro A, Brakefield PM, Marec F, Saccheri IJ. Zygosity-based sex determination in a butterfly
827 drives hypervariability of Masculinizer. *Sci Adv.* 2024;10:eadj6979.

828 24. Yoshido A, Marec F. Deviations in the Z:A ratio disrupt sexual development in the eri silkmoth,
829 *Samia cynthia ricini*. *Genetics.* 2023;224:iyad023.

830 25. Akagi T, Henry IM, Tao R, Comai L. A Y-chromosome-encoded small RNA acts as a sex
831 determinant in persimmons. *Science.* 2014;346:646–650.

832 26. Müller NA, Kersten B, Leite Montalvão AP, Mähler N, Bernhardsson C, Bräutigam K,
833 Carracedo Lorenzo Z, Hoenicka H, Kumar V, Mader M, Pakull B, Robinson KM, Sabatti M,
834 Vettori C, Ingvarsson PK, Cronk Q, Street NR, Fladung M. A single gene underlies the dynamic
835 evolution of poplar sex determination. *Nat Plants.* 2020;6:630–637.

836 27. Zhou R, Macaya-Sanz D, Carlson CH, Schmutz J, Jenkins JW, Kudrna D, Sharma A, Sandor L,
837 Shu S, Barry K, Tuskan GA, Ma T, Liu J, Olson M, Smart LB, DiFazio SP. A willow sex
838 chromosome reveals convergent evolution of complex palindromic repeats. *Genome Biol.*
839 2020;21:38.

840 28. Mao X, Rafati N, Tellgren-Roth C, Ingvarsson PK, Karrenberg S. Sex chromosome evolution
841 mediated by a large inversion and a possible switch of the sex determination gene. *Genome Biol.*
842 2026;27:120.

843 29. Fraïsse C, Picard MAL, Vicoso B. The deep conservation of the Lepidoptera Z chromosome
844 suggests a non-canonical origin of the W. *Nat Commun.* 2017;8:1486.

845 30. Lewis JJ, Cicconardi F, Martin SH, Reed RD, Danko CG, Montgomery SH. The *Dryas iulia*
846 genome supports multiple gains of a W chromosome from a B chromosome in butterflies. *Genome*
847 *Biol Evol.* 2021;13:evab128.

848 31. Han M-J, Luo C, Hu H, Lin M, Lu K, Shen J, Ren J, Ye Y, Westhof E, Tong X, Dai F. Multiple
849 independent origins of the female W chromosome in moths and butterflies. *Sci Adv.*
850 2024;10:eadm9851.

851 32. Shipova AA, Belousova IA, Yakimova ME, Kirichenko NI, Ageev AA, Cusson M, Lukhtanov
852 VA, Ershov NI, Martemyanov VV. From Adam's rib: the Siberian moth's female W chromosome is
853 derived from its Z counterpart. *Mol Biol Evol.* 2026;43:msag055.

854 33. Chen X, Wang Z, Zhang C, Hu J, Lu Y, Zhou H, Mei Y, Cong Y, Guo F, Wang Y, He K, Liu Y,
855 Li F. Unraveling the complex evolutionary history of lepidopteran chromosomes through ancestral
856 chromosome reconstruction and novel chromosome nomenclature. *BMC Biol.* 2023;21:265.

857 34. Orteu A, McCarthy SA, Hornett EA, Gemmell MR, Reynolds LA, Warren IA, Gordon IJ, Hurst
858 GDD, Durbin R, Martin SH, Jiggins CD. The *Hypolimnas misippus* genome supports a common
859 origin of the W chromosome in Lepidoptera. *Genome Biol Evol.* 2024;16:evae215.

860 35. Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, Xia R. TBtools: an integrative
861 toolkit developed for interactive analyses of big biological data. *Mol Plant.* 2020;13:1194–1202.

862 36. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+:
863 architecture and applications. *BMC Bioinformatics.* 2009;10:421.

864 37. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34:3094–
865 3100.

866 38. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and
867 streamlined workflows along with broader and deeper phylogenetic coverage for scoring of
868 eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol.* 2021;38:4647–4654.

869 39. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7:
870 improvements in performance and usability. *Mol Biol Evol.* 2013;30:772–780.

871 40. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear
872 R. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era.
873 Mol Biol Evol. 2020;37:1530–1534.

874 41. Kawahara AY, Plotkin D, Espeland M, Meusemann K, Toussaint EFA, Donath A, Gimnich F,
875 Frandsen PB, Zwick A, dos Reis M, Barber JR, Peters RS, Liu S, Zhou X, Mayer C, Podsiadlowski
876 L, Storer C, Yack JE, Misof B, Breinholt JW. Phylogenomics reveals the evolutionary timing and
877 pattern of butterflies and moths. Proc Natl Acad Sci USA. 2019;116:22657–22663.

878 42. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation
879 sequencing data. Bioinformatics. 2012;28:3150–3152.

880 43. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermini LS. ModelFinder: fast
881 model selection for accurate phylogenetic estimates. Nat Methods. 2017;14:587–589.

882 44. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the
883 ultrafast bootstrap approximation. Mol Biol Evol. 2018;35:518–522.

884 45. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary
885 analyses in R. Bioinformatics. 2019;35:526–528.

886 46. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: an R package for visualization and
887 annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol.
888 2017;8:28–36.

889 47. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. Primer-BLAST: a tool to
890 design target-specific primers for polymerase chain reaction. BMC Bioinformatics. 2012;13:134.

891 48. Hall AB, Qi Y, Timoshevskiy V, Sharakhova MV, Sharakhov IV, Tu Z. Six novel Y
892 chromosome genes in Anopheles mosquitoes discovered by independently sequencing males and
893 females. BMC Genomics. 2013;14:273.

894 49. Chen Y, Chen Y, Shi C, Huang Z, Zhang Y, Li S, Li Y, Ye J, Yu C, Li Z, Zhang X, Wang J,
895 Yang H, Fang L, Chen Q. SOAPnuke: a MapReduce acceleration-supported software for integrated
896 quality control and preprocessing of high-throughput sequencing data. *GigaScience*. 2018;7:gix120.

897 50. Fukui T, Shoji K, Kiuchi T, Suzuki Y, Katsuma S. Masculinizer is not post-transcriptionally
898 regulated by female-specific piRNAs during sex determination in the Asian corn borer, *Ostrinia*
899 *furnacalis*. *Insect Biochem Mol Biol*. 2023;156:103946.

900 51. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
901 *EMBnet J*. 2011;17:10–12.

902 52. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform.
903 *Bioinformatics*. 2009;25:1754–1760.

904 53. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T,
905 McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. *GigaScience*.
906 2021;10:giab008.

907 54. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator.
908 *Genome Res*. 2004;14:1188–1190.

909 55. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L,
910 Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren
911 BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. Full-length transcriptome assembly from
912 RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011;29:644–652.

913 56. Czech B, Hannon GJ. One loop to rule them all: the ping-pong cycle and piRNA-guided
914 silencing. *Trends Biochem Sci*. 2016;41:324–337.

915 57. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware
916 quantification of transcript expression. *Nat Methods*. 2017;14:417–419.

917 58. Xie S, Shen B, Zhang C, Huang X, Zhang Y. sgRNAcas9: a software package for designing
918 CRISPR sgRNA and evaluating potential off-target cleavage sites. PLoS ONE. 2014;9:e100448.

919 **Additional files**

920 Additional file 1 (.docx): Supplementary figures. Figures S1 to S14 with legends, and legends for
921 the four supplementary data files.

922 Additional file 2 (.xlsx): Genome-screening calls, W-linked copy coordinates and W-assembly
923 metrics across the 266 assemblies.

924 Additional file 3 (.xlsx): Target-specificity and mismatch-sensitivity analysis of the piRNA-target
925 alignments.

926 Additional file 4 (.xlsx): Primers, guide RNAs and dsRNA-template oligonucleotides.

927 Additional file 5 (.xlsx): W-copy transcription screen across 61 species, with per-species calls and
928 the public sequencing accessions used.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1.docx](#)