EVOLUTIONARY BIOLOGY

Multiple independent origins of the female W chromosome in moths and butterflies

Min-Jin Han[1,2]†, Chaorui Luo[1]†, Hai Hu[1,2], Meixing Lin[1], Kunpeng Lu[1], Jianghong Shen[1], Jianyu Ren[1], Yanzhuo Ye[1], Eric Westhof[1,3], Xiaoling Tong[1,2]*, Fangyin Dai[1,2]*

Lepidoptera, the most diverse group of insects, exhibit female heterogamy (Z0 or ZW), which is different from most other insects (male heterogamy, XY). Previous studies suggest a single origin of the Z chromosome. However, the origin of the lepidopteran W chromosome remains poorly understood. Here, we assemble the genome from females down to the chromosome level of a model insect (*Bombyx mori*) and identify a W chromosome of approximately 10.1 megabase using a newly developed tool. In addition, we identify 3593 genes that were not previously annotated in the genomes of *B. mori*. Comparisons of 21 lepidopteran species (including 17 ZW and four Z0 systems) and three trichopteran species (Z0 system) reveal that the formation of Ditrysia W involves multiple mechanisms, including previously proposed canonical and noncanonical models, as well as a newly proposed mechanism called single-Z turnover. We conclude that there are multiple independent origins of the W chromosome in the Ditrysia (most moths and all butterflies) of Lepidoptera.

INTRODUCTION

Amphiesmenoptera, which include the Trichoptera and Lepidoptera, have a different sex chromosome system (Z0/ZW, heterogametic females) to most other insect orders (XY system, heterogametic males) (*1*, *2*). A previous study showed that the Z and X chromosomes of winged insects evolved from independent ancestral chromosomes (*3*). Furthermore, the Z chromosomes of Amphiesmenoptera have a unique origin due to their deep conservation in some distantly related moths (*4*, *5*). However, the origin of the lepidopteran W chromosome, which is mainly present in Ditrysia, is unclear. The lepidopteran W chromosome is absent in early divergent lineages (such as Micropterigidae) of the Lepidoptera and their closely related order, the Trichoptera. Consequently, the Z0/ZZ sex chromosomes are considered the ancestral system of Lepidoptera and Trichoptera (Caddisfly) with the W chromosome acquired secondarily by Lepidoptera (*2*, *6*, *7*). Current views on the origin of W are based mainly on comparisons of Z chromosomes and direct comparisons of the W chromosomes in a limited number of species. For example, comparisons of Z chromosomes in some moths indicate that the W chromosome of *Cameraria ohridella* (family Gracillariidae) is generated by a canonical process (Z-autosome fusion), whereas the W of Ditrysia is thought to result from a noncanonical mechanism (recruitment of a B chromosome, which represents a dispensable chromosome) (*5*, *8*). Furthermore, a comparative genomics study in two butterflies not only supports the hypothesis of a B chromosome origin but also suggests that independent origins of the W chromosome have occurred in butterflies due to the lack of homology between the W sequences of *Danaus iulia* and *Kallima inachus* (*9*). To better understand the origin of the W chromosome in Lepidoptera, particularly in Ditrysia, comparative genomic analyses with a larger number of lepidopteran species are needed. Recently, following the development of sequencing technologies, more and more lepidopteran genomes containing the W chromosome have been published (*10*–*23*). These data allow the exploration of the origins of the W chromosome through genomic comparisons.

The silkworm, *Bombyx mori*, is not only an economically important insect but also a model insect species whose importance is surpassed only by that of *Drosophila melanogaster* (*24*, *25*). Over the past two decades, research on silkworm genomics has made remarkable progress. Since the publication of the draft silkworm genome in 2004 (*26*, *27*), population-scale resequencing data of the silkworm genome (*28*, *29*), the chromosome-level reference genome (*30*, *31*), and the high-resolution pan-genome (*30*) have been published successively. However, the silkworm W sequences have not been described in these genomic studies due to their highly repetitive nature, which poses a great challenge during identification and assembly (*31*). To date, knowledge of the *B. mori* W chromosome has come mainly from a small amount of bacterial artificial chromosome sequences (479 fragments with a total size of 1.39 Mb) (*32*) and a few short reads (3679 reads with a size of 1.23 Mb) (*33*). The identification of additional or complete W chromosome sequences is crucial for understanding the mechanism of sex determination in silkworms. A good understanding of this mechanism could eventually promote sericulture by enabling the rearing of males only by female-male inversion (rearing male silkworms has many economic advantages) (*34*), which also has implication for sex-specific control of lepidopteran pests (*35*–*37*).

In this study, we therefore set out to develop methods for obtaining high-quality W chromosome sequences in *B. mori* and to investigate the origin of the W chromosome in Lepidoptera, particularly in Ditrysia. We developed a pipeline to identify W-linked sequences based on the chromosome quotient (CQ). We also assembled the genome of females of *B. mori* at the chromosome level by combining PacBio long reads and Hi-C data. Last, we performed comparative genomic analyses on 24 species, including 21 lepidopterans (including 17 ZW/ZZ systems and 4 Z0/ZZ systems) and three trichopteran samples (Z0/ZZ system), to investigate the origin and evolution of the W. Our study provides a useful tool for identifying highly

[1]State Key Laboratory of Resource Insects, Institute of Sericulture and Systems Biology, Southwest University, Chongqing 400715, China. [2]Key Laboratory of Sericultural Biology and Genetic Breeding, Ministry of Agriculture and Rural Affairs, College of Sericulture, Textile and Biomass Sciences, Southwest University, Chongqing 400715, China. [3]Architecture et Réactivité de l'ARN, Institut de Biologie Moléculaire et Cellulaire, UPR9002 CNRS, Université de Strasbourg, Strasbourg 67084, France.
*Corresponding author. Email: fydai@swu.edu.cn (F.D.); xltong@swu.edu.cn (X.T.)
†These authors contributed equally to this work.

repetitive sex chromosome sequences in other species. In addition, we characterize the W chromosome and many newly identified genes, which will promote functional genomics studies and help to understand the mechanism of sex determination in silkworms. Last, our results provide new information on the origin and evolution of the W chromosome.

RESULTS

Genome assembly and identification of the W chromosome

To assemble a high-quality genome of females of *B. mori*, we generated a total of 41.20-Gb-long reads (~88-fold coverage) using the PacBio Sequel platform. The first assembled genome was refined using our previously released Illumina reads (*38*). We obtained a contig-level assembly with a genome size of ~467.7 Mb, consisting of 366 contigs with an N50 length of 13.6 Mb and one longest contig length of 20.3 Mb (table S1). This indicates that the assembled genome is highly continuous. The size of the assembled genome of females is larger than that of the previously released *B. mori* genome derived from males (~460.3 Mb in SilkDB3.0 and ~454.7 Mb in SilkBase), which is expected with the addition of the W chromosome in the assembly of the genome from females. We then anchored the assembled contigs along the 29 chromosomes ($n = 27 + Z + W$) using 63.70 Gb (~136-fold coverage) Hi-C reads. The results anchored 176 contigs (constituting 48% of all contigs and ~98% of the whole-genome nucleotide bases) on the 29 chromosomes (fig. S1 and table S2). Next, we assessed the quality of the assembled genome at chromosome-level using the coverage rate of BUSCO genes (lepidoptera_odb10), the mapping rate of next-generation sequencing (NGS) reads of the female genome, and the synteny analysis between the female genome assembled here and previously released genomes derived from males. For the BUSCO assessment, 98.5% of complete BUSCOs, including 97.9% single-copy and 0.6% duplicated, were detected in the assembled genome (table S3). Besides, 99.79% of the NGS reads were successfully mapped to the chromosome-level genome. Last, 28 chromosomes (27 autosomes and 1 Z chromosome) display conserved synteny between the three versions of the genome (Fig. 1A). The results of these assessments indicate that the assembled genome has high quality and completeness.

To identify the W chromosome from the assembled genome, we designed a pipeline to calculate CQ and identify the sex chromosome (see Materials and Methods for details). The CQ value represents the ratio of male (ZZ) to female (ZW) read coverage in a genome region (*39*). Theoretically, the CQ values obtained for the Z and W chromosomes should be close to two and zero, respectively. With our method, the CQ values of chromosome 1 (known Z chromosome) and chromosome 29 were approximately two and zero, respectively, while the CQ values of the other chromosomes were all consistently around one, indicating that chromosome 29 is the candidate W chromosome with a size of ~10.1 Mb (Fig. 1B and table S2). To test whether chromosome 29 is the W chromosome of *B. mori*, 15 specific fragments of chromosome 29 were selected for polymerase chain reaction (PCR) verification in six individuals representing three male and three female samples from the same strain (P50). All 15 fragments were detected only in the females but absent in the males (fig. S2). In addition, the Fem [a Piwi-interacting RNA (piRNA)] precursor sequence, which had been identified on the W chromosome as a primary signal for *B. mori* sex determination (*40*), was used as query in BLASTN search against chromosome 29. This detected 137 copies of the Fem precursor with most copies occurring in tandem repeats (Fig. 1C and table S4), close to the Fem copy number estimated previously by quantitative PCR (*41*). These data suggest that chromosome 29 is indeed the W chromosome of *B. mori*. To validate the broad applicability of the pipeline in identifying sex chromosomes across different species, we tested the method on five additional species, including three moths and two birds. Encouragingly, we successfully identified the sex chromosomes in all these species (fig. S3).

Genome annotation and description of the W chromosome

We used RepeatModeler2 and RepeatMasker to annotate the transposable elements (TEs). We estimated that TEs account for ~57% of the whole genome of females of *B. mori* (fig. S4A and table S5), which is higher than previously reported in genomes from males (*42*, *43*). This suggests that the W chromosome is rich in repetitive sequences. TEs constitute ~87% of the W chromosome nucleotides, a value much higher than found in autosomes and Z chromosome (Fig. 2A, fig. S4B, and table S6). Among TEs on the W chromosome, the most abundant transposons (~54% of the W) are long interspersed nuclear elements (LINEs), followed by the long terminal repeat (LTR), DNA transposon, Rolling-circles/Helitron, and short interspersed nuclear elements (fig. S4A and table S6). LTR transposons form ~20% of the W chromosome, a value ~10-fold higher value than on the autosomal (~3%) and Z (~2%) chromosomes (Fig. 2B). These trends are similar in other lepidopteran species where the W chromosome has been released (fig. S5).

To annotate protein-coding genes, we used evidence from ab initio predictions, homology-based searches and alignment of full-length transcripts from entire individuals at multiple developmental stages. A total of 15,950 genes were identified in the *B. mori* genome derived from females (Fig. 2C and table S7). This is unexpected, since this value is lower than that of the genome from males (16,069 predicted genes in SilkDB3.0 and 16,880 genes in SilkBase). Compared to the previous two gene sets from the male reference genomes (details in Materials and Methods), we were surprised to find differences in the predicted gene sets of the same strain genome (Fig. 2C). For instance, in our predicted gene set, a total of 7681 previously predicted genes were missed, including 3380 unique genes from SilkBase, 2738 unique genes from the SilkDB3.0, and 1563 genes that were found specifically in both SilkBase and SilkDB3.0. These discrepancies could be due to the absence or incomplete filtering of transposon-encoded genes in previous analyses. Transposon-encoded genes are often highly repetitive and typically mobile in the genome, which can greatly affect gene prediction results. Approximately 52% (3976 of 7681) of the genes that appear in SilkDB3.0 and SilkBase but are absent in our gene set are transposon-encoded genes (Fig. 2C). Furthermore, the process of whole-genome gene prediction relies heavily on transcriptome evidence, which can also lead to different gene sets. In the genome from females, we used full-length transcriptome data from entire individuals at multiple developmental stages for gene prediction. Whereas the previous gene predictions in two genomes derived from males were based on short-read transcriptome data of different tissues and developmental stages. Among the remaining 3705 non-transposon–encoded genes specific to SilkDB3.0 and SilkBase, approximately 83% (3048) genes lack evidence of full-length transcripts (Fig. 2C), which can be the result of low expression levels (fig. S6A). Consequently, the main reasons why genes are missing from our predicted gene set,

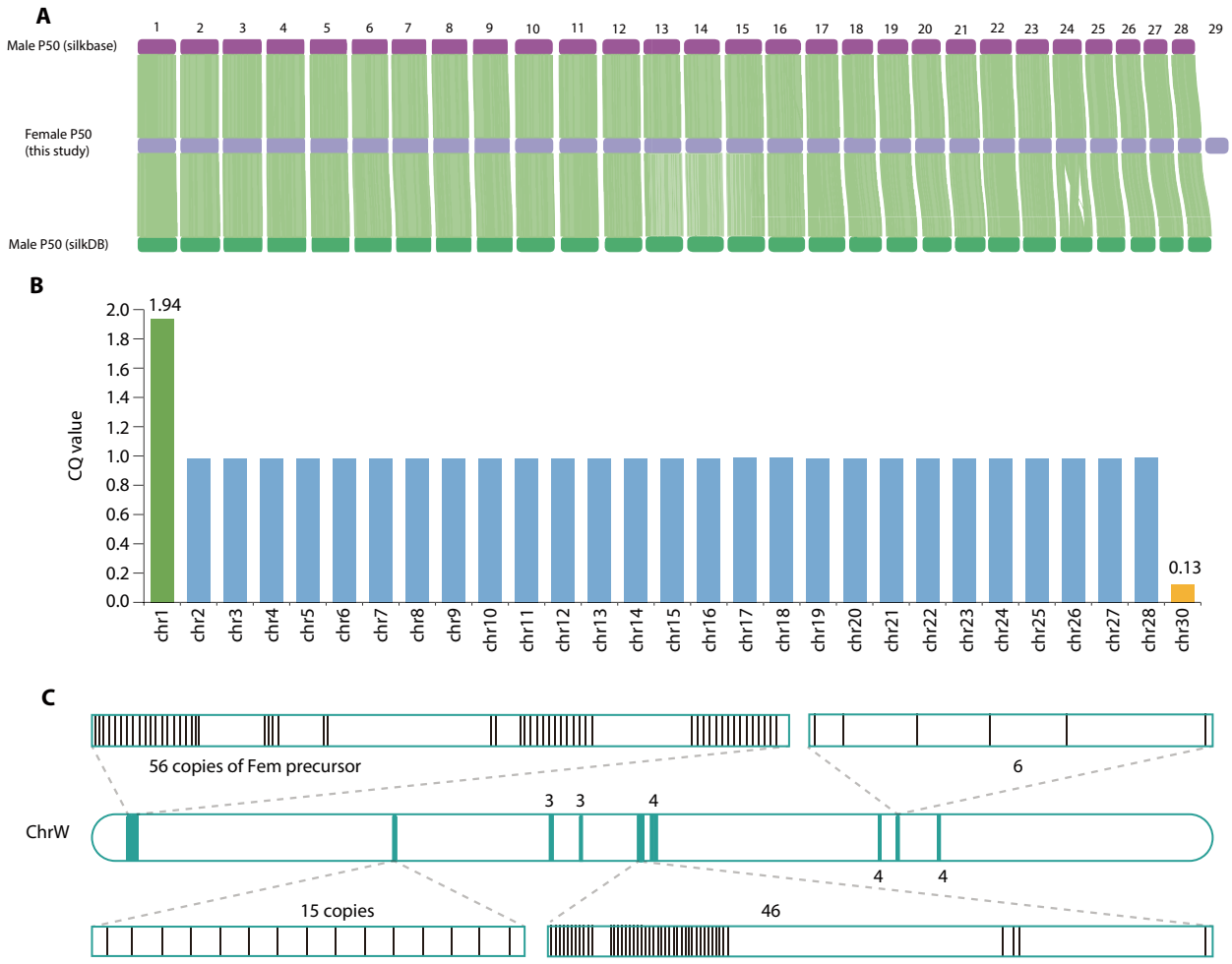


Fig. 1. The assembled chromosomes and identification of the W chromosome in *B. mori*. (**A**) The syntenic relationships between chromosomes of three genomes (light green), including female P50 genome assembled in this study (dilute purple) and two male P50 genomes released in the database of SilkDB (https://silkdb.bioinfotoolkits.net/) (green) and Silkbase (https://silkbase.ab.a.u-tokyo.ac.jp/) (purple). Chromosome 29 has no homolog chromosome in the male genomes. (**B**) The identification of *B. mori* sex chromosomes using the CQ method. The CQ value represents the ratio of male (ZZ) to female (ZW) coverage reads on a chromosome. Theoretically, the CQ values of Z and W should be equal to two and zero, respectively. (**C**) The distribution of Fem (a Piwi-interacting RNA) precursor on the W chromosome. Each black vertical line indicates a Fem precursor, and the numbers show the copy number of Fem.

compared to published data, are incomplete filtering of transposases and bias in transcriptomes resulting from the methodology used for obtaining the previous data.

Of note, we have newly identified 3593 non-transposon–encoding genes in the silkworm genome (Fig. 2C and table S7), all with full-length transcript or evidence of expression, with over 82% (2930) of them having either gene ontology (GO) terms or homologous genes in the National Center for Biotechnology Information (NCBI) nonredundant (NR) protein database (table S7) and with ~73% (2604) of them having homologous genes in other insects (table S8). This suggests that these newly identified genes are actually present in the silkworm genome. We also found that the expression levels of these genes are significantly lower than those of genes present in the three versions of the gene sets (fig. S6A). This could explain the absence of these newly identified genes in the previously published two gene sets (*42*, *43*). In addition, 3517 of the newly identified genes are found on the 27 autosomes and the Z chromosome, suggesting that these genes were overlooked in the previous annotation of genomes from males (fig. S6B). We found 76 genes on the W chromosome (fig. S6B and table S7). The gene density (~8 genes/Mb in average) on the W is much lower than on the autosomes and Z chromosome (Fig. 2A and fig. S4B), supporting the view of a gene-poor sex-specific chromosome (*2*, *31*). Among the 76 protein-coding genes identified on the W chromosome, 23 proteins corresponded to no known proteins in NCBI NR protein databases, 18 proteins corresponded to proteins whose functions are described as uncharacterized or hypothetical, and 35 corresponded to proteins whose function is precisely described (table S7). To determine whether these genes are specific to the W chromosome, we used the W chromosome genes as a query to perform BLAST searches on autosomes and Z chromosome genes. We found that 65 genes (86%) have very similar homologous genes (with identity >80% and coverage >80%) on the autosomes or Z chromosome (fig. S7 and table S9), indicating that these genes could have been acquired through translocation from other parts of the genome.

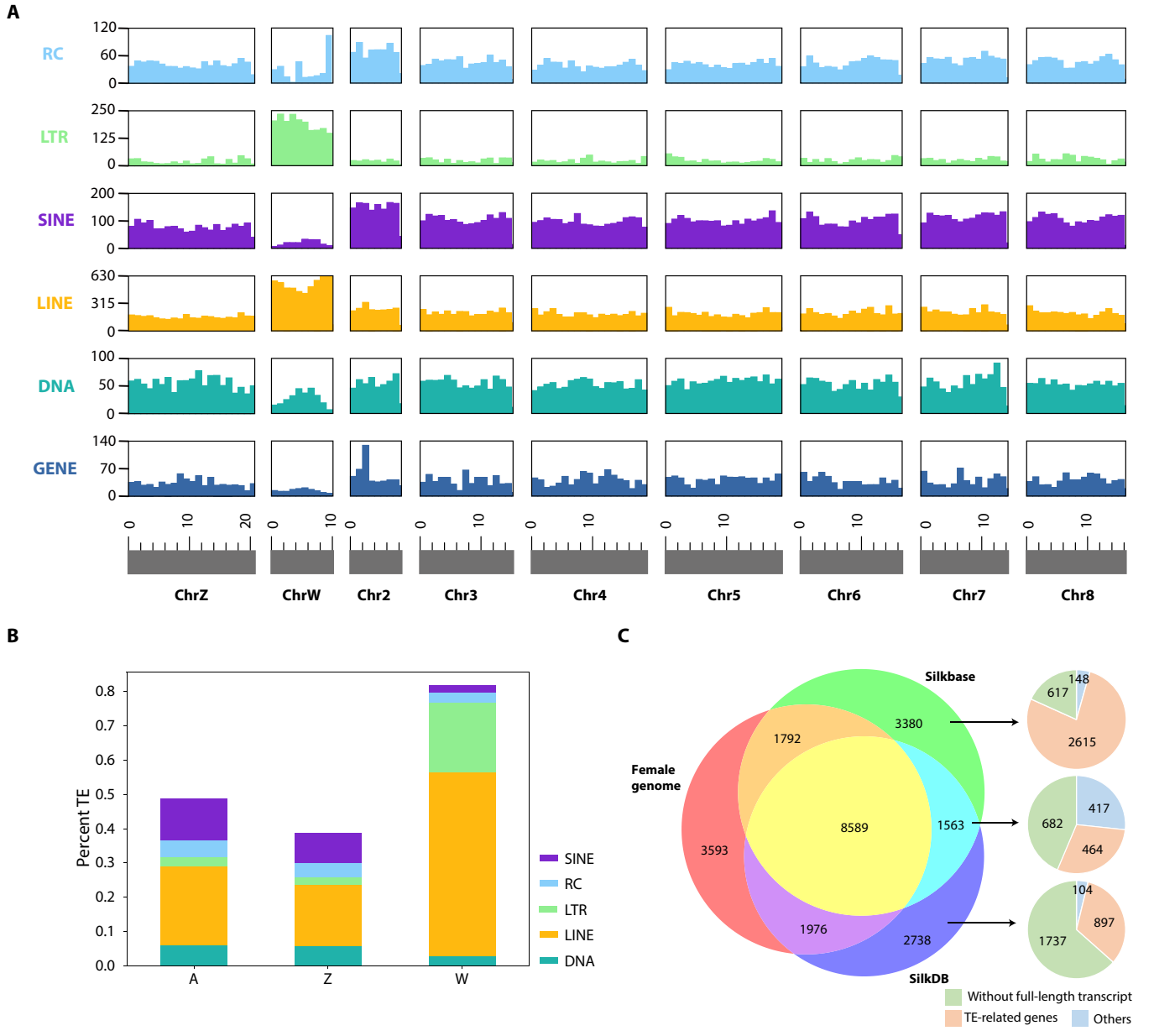


Fig. 2. TEs and protein-coding genes in the silkworm. (**A**) The distribution of protein-coding genes and the different types of TEs on different chromosomes of female *B. mori* (in 1-Mb windows). The vertical coordinate of the gene distribution is the number of genes. The vertical coordinate of the TEs distribution is the number of bases in kilobase (kb). (**B**) Stacked bar charts with the proportion of different TEs on the autosomal (A), Z, and W chromosomes. LINEs and LTRs content on the W chromosome are much higher than that on autosomal and Z chromosomes. (**C**) The Venn diagram at the right represents the comparison results of three different predicted gene sets from the genomes of the same strain. The red circle represents the female silkworm gene set identified in this study, while the green and blue circles represent the previously predicted male silkworm gene sets in SilkDB and Silkbase. The top and bottom pie charts in the right panel represent the proportions of transposase-encoding genes and genes without evidence of full-length transcripts in specific genes from Silkbase and SilkDB, respectively. The middle pie chart represents the proportions of transposase-encoding genes and genes without evidence of full-length transcripts in genes that are simultaneously present in both Silkbase and SilkDB but are missing in the predicted gene set of the female silkworm genome.

Next, we analyzed the expression characteristics of genes on the W and found that the expression level of these genes was significantly lower than that of genes on autosomes and Z chromosome (fig. S8A and table S7). Among the 76 genes, 54 genes virtually showed no expression [transcripts per million (TPM) values <2]. Of the remaining 22 genes showing relatively higher expression levels, 14 genes show higher levels of expression in eggs within 24 hours of laying. However, the fact that 10 of these 14 genes also show relatively higher expression levels in male eggs (fig. S8B and table S7) is perplexing. One possible explanation is that the RNA from these 10 genes may be of maternal origin, given that their expression occurs during early embryonic development. Another explanation could

be that the RNA detected in male eggs may come from genes paralogous to the 10 genes on the Z and autosomal chromosomes. Of the 22 genes, 4 are primarily expressed in female eggs or ovaries. Of these four genes, the functions of three of them (*DFBM015880*, *DFBM015927*, and *DFBM015926*) remain unknown, while one gene (*DFBM015912*) has been annotated as a gene close to the *angel homolog 2*–like gene, but its biological function remains uncertain in insects (table S7). Therefore, on the basis of the above information concerning the W chromosome genes, we cannot currently determine their biological functions in female silkworms. However, it will be interesting to determine their function in the context of female-specific biology in the future.

The origin and evolution of W chromosome in Lepidoptera

Among Lepidoptera, W chromosomes are found mainly in Ditrysia, which includes most moths and all butterflies, with a few occurrences in non-Ditrysia, such as Tischeria ekebladella (family Tischeriidae) (Fig. 3A). An earlier study suggested that the independent origins of the W chromosomes in Ditrysia and Tischeriidae, as well as the W of Ditrysia, reflect secondary origins early in Ditrysia evolution (*8*). At present, three models for the origin of the W chromosome have been proposed (Fig. 3B). These are two canonical models (Z-autosome fusion and sex chromosome turnover) and one noncanonical (recruitment of a B chromosome). In the canonical models, the Z and W chromosomes contain the same pair of ancestral autosomes (*2*, *44*). In the noncanonical model, W is derived from a B chromosome, which represents a dispensable chromosome that arises mainly through rearrangements or duplications from standard chromosomes (*45*, *46*). At present, most studies support the view that W in Ditrysia was generated by the noncanonical model (*5*, *8*, *9*). However, it remains unclear whether the W chromosomes in various species of Ditrysia all derive from a common ancestor or reflect multiple independent origins.

To investigate the origin of the W chromosome in Ditrysia, we first performed synteny analyses between the Z chromosomes of 24 species, including 21 species (17 ZW systems and 4 Z0 systems) of Lepidoptera and 3 samples (Z0 systems) of Trichoptera using the locations of orthologous genes. The genes linked to the Z chromosome in 13 species among the 17 ZW systems show synteny across the entire Z chromosome (Fig. 3C). Strong synteny between the Z chromosomes of Lepidoptera and Trichoptera is also observed, supporting the notion the that Z chromosome of Amphiesmenoptera has a unique origin (*4*, *5*). The Z chromosomes in four species of Ditrysia, *Boloria selene*, *Melinaea menophilus*, *Pieris brassicae*, and *Cydia pomonella* appeared as a result of Z-autosome fusions (Fig. 3C), indicating that these Z chromosomes acquired a neo-Z and suggesting that these four species could generate a neo-W by Z-autosome fusion. Regarding the neo-Z chromosomes in these four species, our results are consistent with previous findings (*11*, *47*, *48*). However, these earlier studies did not provide evidence for the presence of neo-W in these four species.

To determine whether the W chromosomes of species with neo-Z chromosomes were generated by Z-autosome fusion, we performed synteny analysis to examine sequence homology between the W and other chromosomes (Z and autosomes) in each species of the 17 ZW systems. Considering the gene-poor nature and high repeat sequence content in the W chromosome, the repeat-masked W chromosome sequence was used as query to perform a homology search against the repeat-masked Z and autosome chromosomes and to identify synteny blocks between the W and other chromosomes using Satsuma2synteny with a minimum alignment length of the default value (0), 100, and 500 base pairs (bp), respectively. Under the Z-autosome fusion hypothesis, W chromosome should be more similar to the Z chromosome than to the autosomes in the same genome (Fig. 3B). Among the four species with neo-Z chromosome, namely *B. selene*, *M. menophilus*, *P. brassicae*, and *C. pomonella*, we found that, in the genome of *C. pomonella*, the number of synteny blocks and the total length of synteny blocks between the W and Z chromosomes were substantially higher than those between the W chromosome and autosomes (Fig. 3D and fig. S9). This provides further evidence that *C. pomonella* generated a neo-W chromosome by Z-autosome fusion. In *P. brassicae*, we found that the W chromosome also exhibits a higher level of homology with its Z chromosome, although it is not as prominent (fig. S9). In the other two species (*B. selene* and *M. menophilus*) with a neo-Z chromosome, we found no evidence of a neo-W chromosome. The W chromosomes of these two species show the greatest number and the longest total length of synteny blocks with the autosomes rather than with the Z chromosome (fig. S9). We speculate that in *B. selene* and *M. menophilus*, one chromosome of an autosomal pair fused with the Z chromosome to form the neo-Z, while the second chromosome of the autosomal pair may have been lost without fusing with the W chromosome to form a neo-W. This resulted in two species retaining only the ancestral W chromosome. Alternatively, we cannot formally exclude the possibility that the neo-W could persist but diverged sufficiently to lose detectable homology to the neo-Z.

Unexpectedly, in the three species *Maniola jurtina*, *Anthocharis cardamines*, and *Phlogophora meticulosa*, we found that the number and total length of synteny blocks are highest between the W and Z chromosomes (Fig. 3D and figs. S10 to S12). According to the three possible hypotheses proposed for the origin of the W chromosome (Fig. 3B), the W chromosome will exhibit the highest homology with the Z chromosome when generated by canonical models, Z-autosome fusion and sex chromosome turnover. Under the Z-autosome fusion hypothesis, we would expect to observe Z-autosome fusion events, but no such events were observed in these three species (Fig. 3C), rejecting the possibility that the W chromosomes in these three species are generated through Z-autosome fusion. Furthermore, according to the sex chromosome turnover hypothesis, the Z chromosome in a species would transform into an autosome, and a pair of autosomes would then transform into Z and W chromosomes, respectively (Fig. 3B). If this were the case, the Z chromosome generated through sex chromosome turnover would show no homology with the Z chromosomes generated through non-sex chromosome turnover in other species. However, we observed a high degree of homology among the Z chromosomes of the entire order Lepidoptera (Fig. 3C), which also rejects the possibility that these three species generate their W chromosomes via sex chromosome turnover. Therefore, the origin of the W chromosome in these three species may be due to a new mechanism that we name single-Z turnover, in which one of the Z chromosomes in a pair acquires a female-determining factor, creating a new W chromosome that shares homology only with the Z chromosome (Fig. 4).

It is probable that the W chromosomes in these three species formed recently because some of their closely related species have ancestral W chromosomes that share a common origin. It is further possible that the ancestral W chromosomes of these three species have been lost. For example, *B. selene*, *D. iulia*, *K. inachus*, *Vanessa*

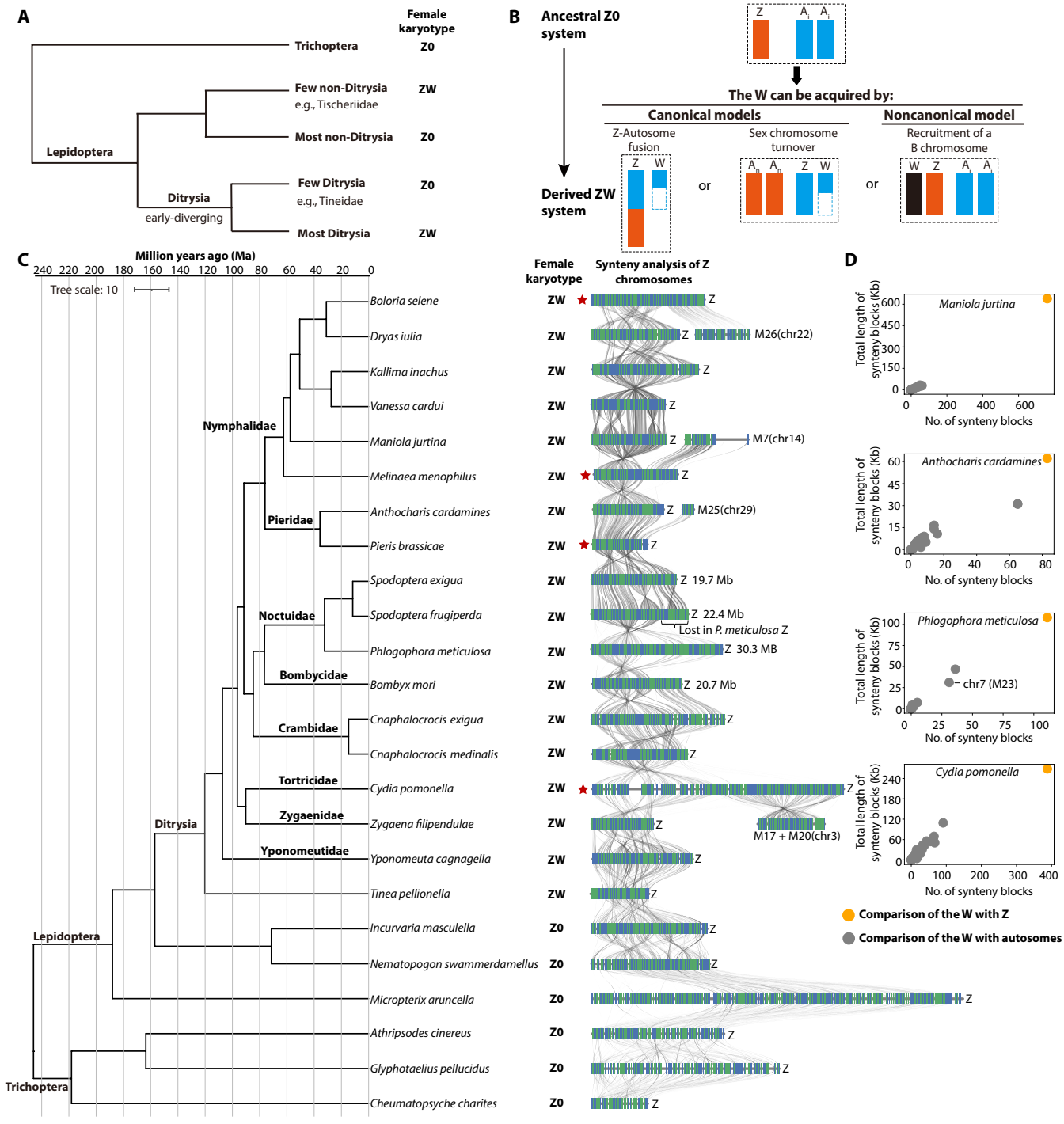


Fig. 3. The deduced origins of the W chromosomes are based both on the gene synteny within the Z chromosomes and on the sequence homology between Z and W chromosomes in the same species. (**A**) The species distribution of sex chromosome systems, containing female Z0 and ZW systems in Lepidoptera and Trichoptera based on prior studies. The W chromosome is mainly found in Ditrysia (Lepidoptera). (**B**) The three possible mechanisms of W chromosome formation. The ancestral female Z0 system has a Z chromosome (in red) and pairs of autosomes (A_i, in blue). For the canonical models, the W and Z chromosomes are descended from a single autosome pair and should be more similar than the W and autosomes in a ZW system. In the noncanonical model, the W is derived from a B chromosome that arose by rearrangements or duplications from one or more autosomes. The white regions with blue dotted lines represent the further degradation of the neo-W chromosome. (**C**) The tree on the left represents the phylogenetic relationships of the species. The topological structure of the tree and the divergence times between the species were estimated by maximum likelihood (ML) based on single-copy genes of lepidopteran BUSCO set. On the right, the female karyotype of each female is presented after the corresponding species name. In the middle (in green) are shown the results of the gene synteny for the Z chromosomes. The red five-pointed stars indicate the species with Z-autosome fusion. (**D**) The scatter plots represent the number (*x* axis) and total length (*y* axis, in unit of kilobase) of synteny blocks between the W (masked repetitive sequence) and autosomes (in gray) and Z (in yellow), respectively, in four ZW systems.

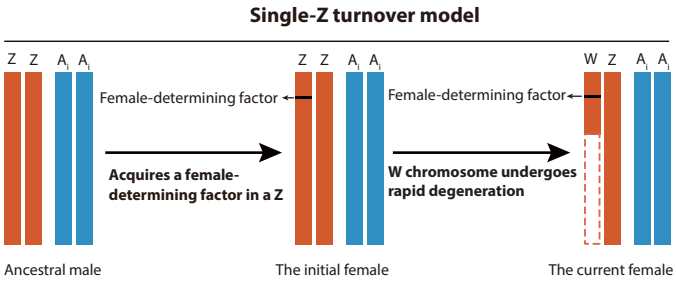


Fig. 4. The single-Z turnover model, a newly proposed mechanism of W origin. Three conditions would be required for the occurrence of this mechanism. First, the W chromosome of a species has high homology with its own Z chromosome. Second, no Z-autosome fusion has occurred. Third, there is no homology between the W of this species and the W chromosomes generated through other mechanisms in closely related species. In such a situation, we suggest that the W chromosome of this species has been generated through a single-Z turnover mechanism.

cardui, and *M. jurtina* all belong to the Nymphalidae family. The W chromosomes in the first four species share a common origin. Only *M. jurtina* has a W chromosome formed through single-Z turnover, and it does not show any homology with the W chromosomes of the other species of Nymphalidae. Therefore, we speculate that the ancestral W chromosome in *M. jurtina* has been lost.

In the case of the abovementioned species with a neo-W or in which occurred single-Z turnover, we expect to observe that the W chromosomes may have no (or fewer) homologous sequences with the W produced by the nonclassical model (recruitment of a B chromosome) in closely related species. On the other hand, W chromosomes derived from a common ancestor may have greater homology between each other. We therefore studied the homology (indicated by the number and total length of synteny blocks) between the repeat-masked W of each species and the repeat-masked genome of its closely related species using reciprocal comparisons (details in Materials and Methods). Among the species of Nymphalidae, we observed maximum bidirectional homology between the W chromosomes of three pairwise comparisons, *K. inachus* versus *V. cardui*, *B. selene* versus *V. cardui*, and *B. selene* versus *K. inachus*, irrespective of the minimum alignment length (0), 100, or 500 bp (Fig. 5A and figs. S13 and S14). For instance, when the W of *V. cardui* was used as query for homology search in each chromosome of *K. inachus*, the number and total length of synteny blocks between the W chromosomes were the highest. In the reverse comparison, when we used the W of *K. inachus* as the query to search for homology sequences in each chromosome of *V. cardui*, the W chromosomes of both species always showed the highest level of homology, suggesting that the W chromosomes of these two species originated from a common ancestor. In contrast, in two closely related species (*B. selene* and *D. iulia*), no homology was detected between the W chromosomes, which is consistent with the results of earlier work (*9*). In addition, it is worth noting the occasional absence of reciprocity in some cases. For example, when the minimum alignment length was set to the default value (0) or 100 bp, using *D. iulia* W as the query for homology against each chromosome of *K. inachus*, we found the highest homology between the W chromosomes of the two species (Fig. 5A and fig. S13). However, in the reverse comparison, using *K. inachus* W for homology search against the *D. iulia* genome, we found the highest homology between *K. inachus* W and two autosomes (M25 and M10) of *D. iulia* (Fig. 5A). This inconsistency in the reciprocal comparisons should be traced to the differences in the length of W chromosome sequences between the different species. For example, the W chromosome in *K. inachus* is ~9.4 Mb in length and, in *D. iulia*, it is only about 2.2 Mb (table S10). The relatively longer sequence length of the W in *K. inachus* may have acquired many autosomal sequences during the evolution, since its W chromosome shares more synteny blocks with autosomes (Fig. 5A). Therefore, in a reciprocal comparison, if one direction of comparison shows the highest numbers and the greatest total length of synteny blocks between W chromosomes, we consider those two W chromosomes to be homologous.

In addition, when the minimum alignment length is set to 500 bp, there are a few cases where results differ from those obtained when the minimum alignment length is set to the default (0) or 100 bp, although most results are consistent (Fig. 5 and figs. S13 and S14). For example, when the minimum alignment length is set to 500 and *D. iulia* W is used as a query for homology search against the *K. inachus* genome, we found the highest homology between *D. iulia* W and an autosome (M5) of *K. inachus*. This is not consistent with the description above where the minimum alignment length was set to the default parameter or 100 bp (Fig. 5A and fig. S13). However, we found that when the minimum alignment length is set to 500, there are only two synteny blocks between *D. iulia* W and the autosome of *K. inachus* (fig. S14), which is not sufficient to determine whether the two chromosomes are homologous. Consequently, in reciprocal comparisons between *D. iulia* and *K. inachus*, we relied on the results obtained with the minimum alignment length set to the default value and 100 bp.

On the evidence of the above, the W chromosomes of four species in the Nymphalidae family, *B. selene*, *D. iulia*, *K. inachus*, and *V. cardui*, probably derive from a common ancestral chromosome. Although there is no evidence of homology between *B. selene* W and *D. iulia* W, they both show homology with *V. cardui* W (Fig. 5A). This could be due to rapid degeneration in *B. selene* W and *D. iulia* W, which results, for each W, in the retention of homologous sequences from different regions of *V. cardui* W (fig. S15). Here, we would like to emphasize that *M. jurtina* W shows the highest homology with Z chromosomes of other closely related species (Fig. 5A and figs. S13 and S14), which is consistent with the finding that *M. jurtina* W shows the greatest homology with its own Z chromosome (Fig. 3D). This confirms the hypothesis that *M. jurtina* W was generated through single-Z turnover. Last, we found no homology between *M. menophilus* W and the W chromosomes of the other five species in the Nymphalidae. However, on the basis of the present results, we cannot determine whether *M. menophilus* W has an independent origin or whether the loss of homology is due to a longer divergence time (>60 Ma) between *M. menophilus* and the other five Nymphalidae species (Fig. 5A).

In the two species of Pieridae, *A. cardamines* and *P. brassicae*, we found that the W chromosomes have little or no homology (fig. S16). For instance, when we set the minimum alignment length to 100 bp and used the *A. cardamines* W as a query to search for synteny blocks in the entire genome of *P. brassicae*, the total length of synteny blocks between the two W chromosomes was only around 4 kb, and the total number of synteny blocks was only 10. Furthermore, by analyzing the possible of generation of the two W chromosomes, we found that the *A. cardamines* W chromosome showed the highest homology with its own Z chromosome, but no

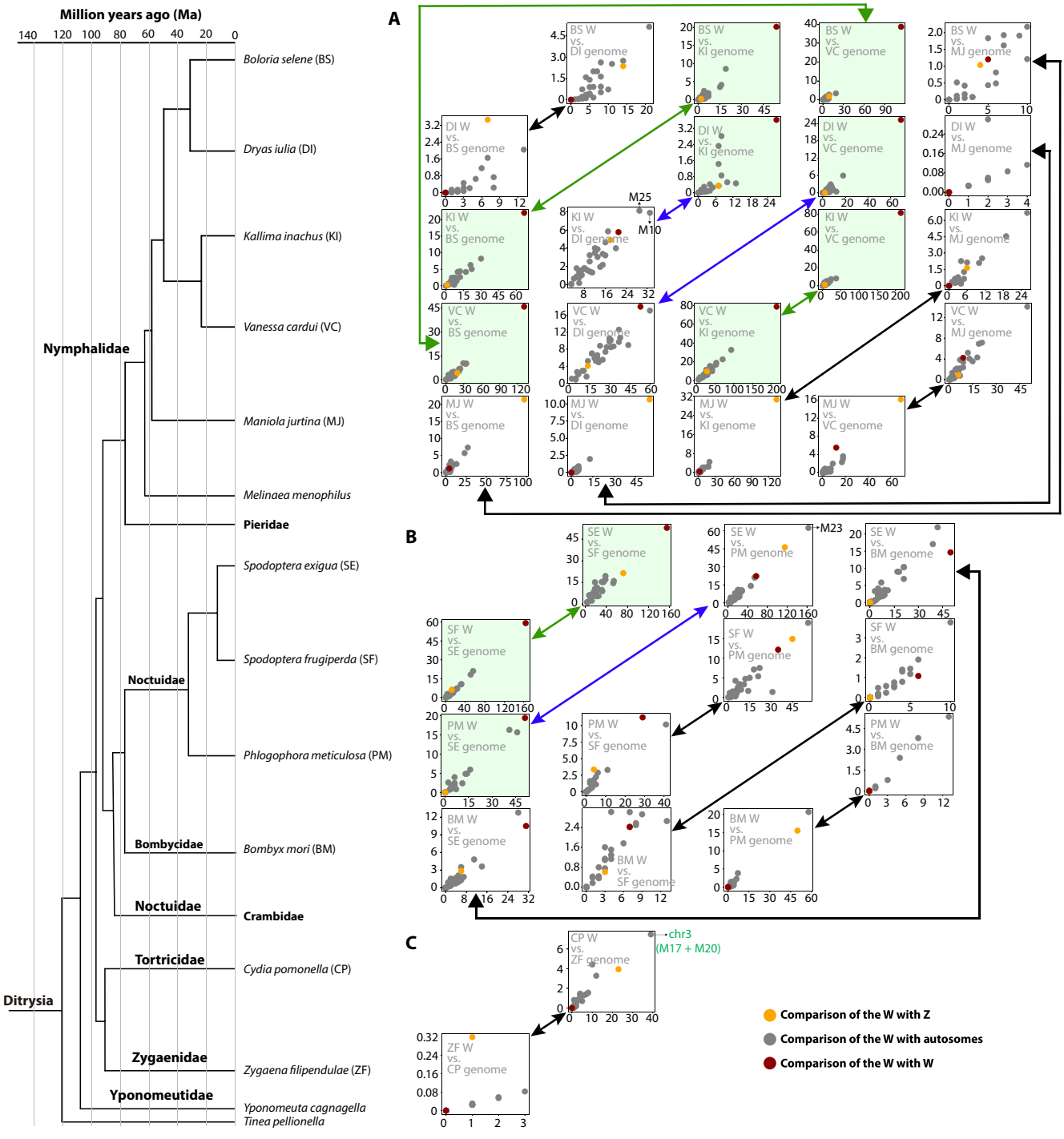


Fig. 5. The homology between the W of each species and each chromosome of a closely related species. The synteny blocks between chromosomes serve as indicators of chromosome homology. They were estimated using Satsuma2synteny with the minimum alignment length set to the default value (0). The homology between the W of a species and each chromosome of its closely related species in (**A**) Nymphalidae, (**B**) Noctuidae and Bombycidae, and (**C**) two species including one Tortricidae and one Zygaenidae. In each scatter plot, both the W query and target genome are displayed with the vertical axis representing the total length of synteny blocks and the horizontal axis the number of synteny blocks. Each dot represents the number and total length of synteny blocks for a comparison between the indicated W chromosome with an autosome (gray dot), Z (yellow point), and W (red point) of the closely related species. The double-headed arrow lines connect the reciprocal comparisons between two species. For example, in the comparison between *B. selene* and *D. iulia*, one direction of comparison involves using the *B. selene* W as the query to search against each chromosome of *D. iulia*, while the other direction involves using the *D. iulia* W as the query to search against each chromosome of *B. selene*. The color code for the double-headed arrow lines represent bidirectional highest homology (green), unidirectional highest homology (blue), and bidirectional lack of homology between two W chromosomes in reciprocal comparisons (black). The scatter plots with a light green background represent the W chromosomes between the two species with the highest number and total length of synteny blocks. The tree on the left is from Fig. 3C.

Z-autosome fusion occurred in the *A. cardamines* genome (Fig. 3, C and D). This rejects the possibility that the *A. cardamines* W chromosome is generated by the Z-autosome fusion model. Therefore, we assume that the *A. cardamines* W can also be generated through the single-Z turnover mechanism. In *P. brassicae*, we observed a relatively higher homology between the W and Z chromosomes (fig. S9) and a neo-Z chromosome formation (Fig. 3D). The synteny blocks between the W and Z of *P. brassicae* are mainly located on the neo-Z chromosome of *P. brassicae* (fig. S17). However, there is only one specific region of synteny between the neo-Z and W chromosome (fig. S17). This pattern is definitely more consistent with a recent segmental duplication from the neo-Z onto the W. If homology between neo-Z and W reflect shared ancestry as chromosomal homologs, then one would expect a broad, interspersed pattern of homology detected along the length of the neoZ and W. To investigate whether the synteny blocks between *P. brassicae* W and its neo-Z are due to a recent segmental duplication from the neo-Z onto the W, we analyzed the sequence identity of these synteny blocks. If these were caused by a recent segmental duplication, we would expect these synteny blocks to exhibit high sequence identity. The results indeed showed that these synteny blocks have a high sequence identity (with an average sequence identity of about 93%) (fig. S17), suggesting that the synteny blocks between *P. brassicae* W and its neo-Z are indeed due to a recent segmental duplication from the neo-Z onto the W. We considered the recent finding that the closely related species of *P. brassicae*, *Pieris mannii*, and *Pieris rapae*, generated neo-W (*49*) and explored whether *P. brassicae* W and its neo-Z do not show other synteny blocks besides the segmental duplication because of high divergence between *P. brassicae* W and its neo-Z leading to the loss of synteny blocks. We analyzed the distribution of synteny blocks between *P. brassicae* W and *P. brassicae* neo-Z, respectively, with the chromosome M25 of the closely related species *A. cardamines*. This analysis was prompted by the observation that the protein-coding gene synteny revealed a Z-M25 chromosome fusion event in the *P. brassicae* (Fig. 3C). If *P. brassicae* generated a neo-W from M25, one would expect to observe numerous synteny blocks still retained between *P. brassicae* W and *A. cardamines* M25. The results revealed that synteny blocks between *P. brassicae* W and *A. cardamines* M25 are scattered throughout the entire *P. brassicae* W chromosome, indicating that *P. brassicae* W is indeed a neo-W formed from the M25 chromosome (fig. S17). Moreover, we found that the synteny blocks between *P. brassicae* W and *A. cardamines* M25, as well as the synteny blocks between *P. brassicae* neo-Z and *A. cardamines* M25, are located in different regions of the *A. cardamines* M25 chromosome (fig. S17), suggesting that *P. brassicae* neo-W and *P. brassicae* neo-Z have retained different regions of the M25 chromosome. Moreover, considering the lack of notable homology between the W chromosomes of *A. cardamines* and *P. brassicae* and their generation through different mechanisms, we presume that the W chromosomes of these two species have independently originated.

Among the three species of the Noctuidae, the W chromosomes of *Spodoptera exigua* and *Spodoptera frugiperda* show the highest homology. Whether using the W chromosome of *S. exigua* or *S. frugiperda* as the query sequence, the number and total length of synteny blocks between the W chromosomes of these two species are the highest (Fig. 5B and figs. S13 and S14). In addition, in the single-direction comparison (*P. meticulosa* W versus *S. exigua* genome), the W of *P. meticulosa* also exhibits the highest homology with the W of *S. exigua* (Fig. 5B and fig. S13). The above results indicate that the W chromosomes of *S. exigua*, *S. frugiperda*, and *P. meticulosa* may have originated from a common ancestor. Neither Z-autosome fusion event nor homology between the Z and W chromosomes was observed in the genomes of *S. exigua* and *S. frugiperda* (Fig. 3C and fig. S9), which suggests that the W chromosomes of these two species may originate from a common ancestral B chromosome. However, the W of *P. meticulosa* shows the highest homology with its own Z chromosome (Fig. 3D), and the W chromosomes of two Spodoptera species show the highest homology with an autosome (M23) and the Z of *P. meticulosa* (Fig. 5B and figs. S13 and S14). We speculate that these results are due to the integration of varying levels of W chromosome sequences into the M23 autosome and Z chromosome in the *P. meticulosa* genome. This is supported by our observation that the Z of *P. meticulosa* has undergone significant variation compared to Z chromosomes of the two Spodoptera species (Fig. 3C). For example, from the analysis of gene collinearity between Z chromosomes, it is evident that the Z chromosome of *P. meticulosa* exhibits a large deletion at the left end compared to Z chromosomes of the two Spodoptera species. Furthermore, the total length of the *P. meticulosa* Z (~30 Mb) is obviously higher than that of *S. frugiperda* (~22 Mb) and *S. exigua* (~20 Mb) Z chromosomes, which indicates that the Z chromosome in *P. meticulosa* has undergone extensive gain and loss events.

To check whether the Z and autosomal M23 of *P. meticulosa* have acquired W chromosome sequences, we investigated first the distribution of homologous blocks between Spodopteran W and *P. meticulosa* Z on the Z chromosome of *P. meticulosa* and, second, the distribution of homologous blocks between Spodoptera Z and *P. meticulosa* Z on the Z of *P. meticulosa*. We then compared the differences between these two distributions. Since the Z chromosomes of Spodoptera and *P. meticulosa* are homologous, if the *P. meticulosa* Z chromosome has acquired a portion of its own W sequence, we would expect to observe that the locations of homologous blocks between Spodoptera Z versus *P. meticulosa* Z and Spodoptera W versus *P. meticulosa* Z on the Z chromosome of *P. meticulosa* are nonoverlapping. We found that the homologous blocks between Spodoptera W and *P. meticulosa* Z were primarily located outside the homologous blocks between the Spodoptera Z and *P. meticulosa* Z in the Z of *P. meticulosa* (fig. S18A). Similarly, we also observed that the homologous blocks between Spodoptera W and *P. meticulosa* M23 were mainly located outside the homologous blocks between Spodoptera M23 and *P. meticulosa* M23 in the M23 of *P. meticulosa* (fig. S18B). These results also suggest that the M23 and Z of *P. meticulosa* has integrated some W sequences. Therefore, the highest homology observed between the W and Z of *P. meticulosa* is likely not related to the formation of the *P. meticulosa* W but rather due to the integration of some W sequences in the Z chromosome. Considering the relatively high homology between the W chromosomes of the three Noctuidae species, we propose that the W chromosomes of these three species have a common origin.

A question remains: Why does the W of *P. meticulosa* show the highest homology with its Z, while the W chromosomes of the two Spodopteran species exhibit the highest homology with *P. meticulosa* M23? A possible explanation is that the W chromosomes of different species have lost varying degrees of blocks with homology with *P. meticulosa* M23. For instance, compared to the homologous blocks between the Spodoptera W and the *P. meticulosa* M23, the number of homologous blocks between the *P. meticulosa* W and its

own M23 chromosome is relatively low (fig. S18C). This could be due to the loss of a larger number of blocks in *P. meticulosa* W than the homologous blocks in its own M23 chromosome during the evolutionary process. In addition, no homology was observed between the W chromosome of *B. mori*, a sister branch of the family Noctuidae in the phylogenetic tree, and the W chromosomes of Noctuidae (Fig. 5B). This does not mean that the W chromosome of the silkworm originated independently because the divergence time between the silkworm and these other species is relatively long (>70 Ma). Thus, the lack of homology between the W chromosomes of the silkworm and other species is probably due to the rapid degeneration of its W chromosome.

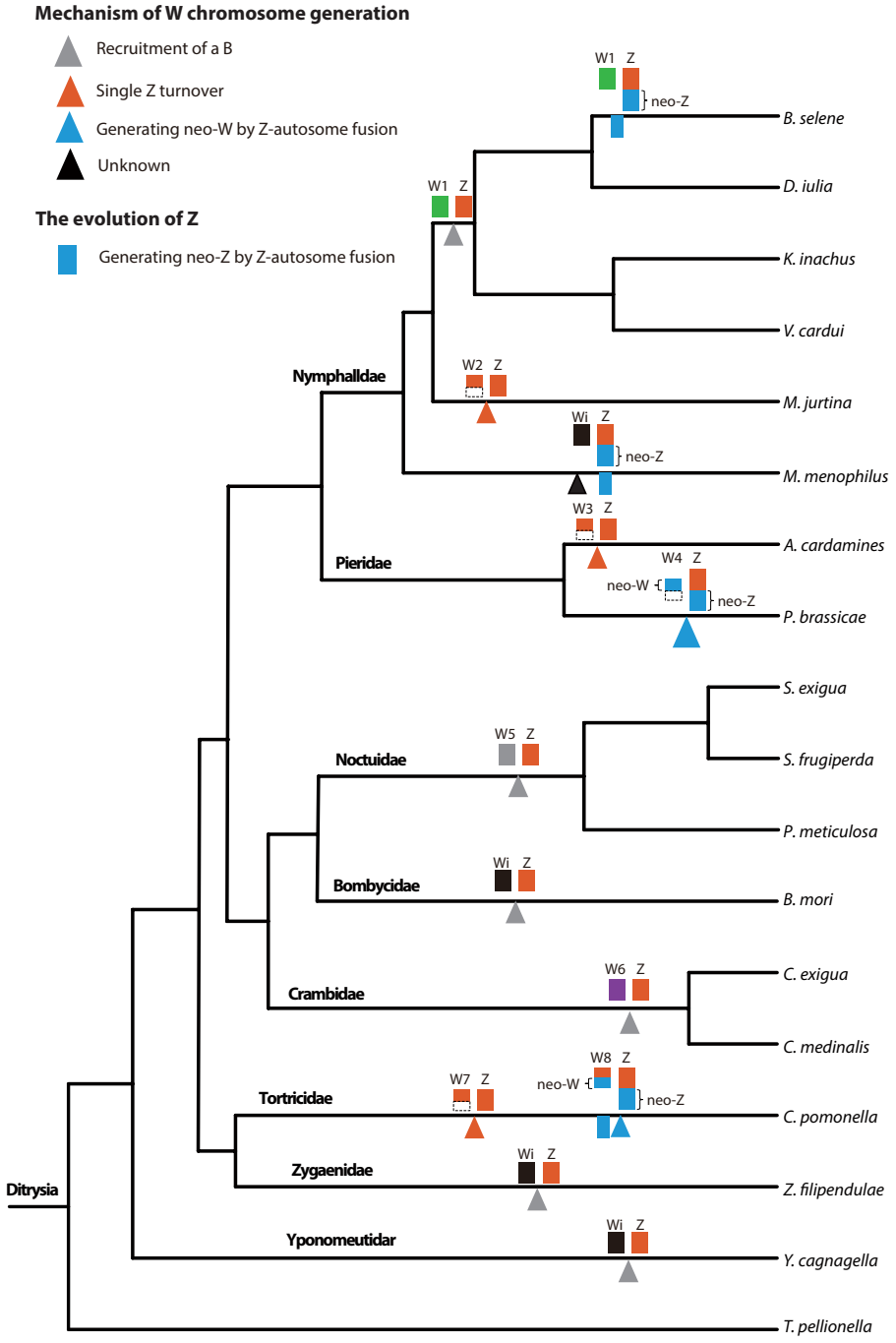
In the two species of Crambidae, the W chromosomes show a very high homology (fig. S19). In addition, neither Z-autosome fusion nor any homology between the W and Z chromosomes within the genome was observed (Fig. 3C and fig. S9), which suggests that the W chromosomes in these two species may have originated from a common ancestral B chromosome. In the remaining three species, *C. pomonella*, *Zygaena filipendulae*, and *Yponomeuta cagnagella*, no homology was observed between their W chromosomes and between those of other species. This is probably due to a longer divergence time (>90 Ma) leading to a complete loss of homologous sequences. We found that the W of *C. pomonella* shows a very high homology with the Chr3 chromosome of *Z. filipendulae* (Fig. 5C and fig. S13) and that the neo-Z chromosome of *C. pomonella* also shows homology with the Chr3 of *Z. filipendulae* (Fig. 3C). These results provide strong evidence that *C. pomonella* acquired a neo-W through Z-autosome fusion. Further, if *C. pomonella* is entirely composed of a neo-W chromosome, one would expect to observe that the homologous blocks between *C. pomonella* W and *C. pomonella* Z are primarily distributed on the neo-Z chromosome and that there would be almost no homology between the ancestral Z chromosome and the neo-W chromosome. However, we observe an almost equal distribution of synteny blocks in the W along ancestral and neo-Z (fig. S20). We speculate that the W chromosome of *C. pomonella* may also contain both the ancestral W and the neo-W, like the Z chromosome of *C. pomonella*, after a fusion event between the ancestral W and the neo-W. The ancestral W chromosome of *C. pomonella* may have originated from the ancestral Z through the single-Z turnover mechanism, since the number of homologous blocks occurring between the ancestral W and ancestral Z chromosomes is higher than the number of homologous blocks between *C. pomonella* W and *C. pomonella* autosomes (fig. S21).

DISCUSSION

Here, we describe comparative genomic analyses of 24 insects, containing 7 Z0/ZZ (female/male) systems and 17 ZW/ZZ (female/male) of lepidopteran species that throw light on the origin and evolution of the W chromosome of Lepidoptera. First, the study reveals the multiple independent origins of the W chromosome in Ditrysia (Fig. 6). This conclusion is contrary to the findings of a recently published work where the authors used the number of collinear genes of W chromosomes from different species to infer their origins (*50*). The use of collinear genes to determine the homology between autosomes or Z chromosomes is a generally accepted method. However, it may not be appropriate for the W chromosomes because they are rich in repetitive sequences and gene-poor (*2*, *31*). Unexpectedly, many genes on the W chromosomes of different species were found in the previous study, with an average of approximately 1287 genes per W chromosome (86,241 genes on 67 Ws) (*50*). The measured gene counts are much higher than the numbers of genes on autosomes (an average of ~667 genes) and Z chromosomes (an average of ~915 genes) (fig. S22). Such large numbers of genes may indicate that the whole genome predicted gene sets used in the previous study may not have been filtered out from genes derived from transposons. According to our analysis of 17 lepidopteran ZW systems, the W chromosomes are all enriched with many LTR and LINE transposons (fig. S5). Autonomous LTR and LINE retrotransposons usually encode genes such as *gag* and *pol*. The repeated occurrence of these protein-coding genes derived from LTR and LINE retrotransposons on W chromosomes can lead to results with many false-positive collinearity signals between different W chromosomes. To verify this, we annotated the predicted genes of the W chromosomes from 16 ZW systems of the present analysis (excluding the silkworm) to determine how many protein-coding genes on the W chromosomes belong to transposon-related genes. The results show that most predicted genes on the W chromosomes (an average of approximately 68%) belong to transposon-related genes (fig. S23A and table S11) and that these genes are mainly composed of gag and pol proteins from LINE and LTR (fig. S23, B and C, and table S12). To further verify whether the large number of collinear genes detected between W chromosomes was due to the incomplete filtering of transposon-related genes, we used the same method as that used in the previous study (*50*) to identify candidate collinear genes between W chromosomes before and after filtering transposon-related genes. The results show that before filtering transposon-related genes, there were indeed some collinear genes between some W chromosomes (fig. S24). However, after filtering transposon-related genes, no collinear genes were detected between the W chromosomes (fig. S24). Further annotation of the proteins detected in the collinear genes before filtering transposon-related genes revealed that approximately 97% of these collinear genes belong to transposon-related genes, with most being pol and gag (fig. S24 and table S13). The above results suggest that it is not appropriate to use the number of collinear genes between W chromosomes, especially the number of collinear genes without filtering transposon-related proteins, to infer the homology of W chromosomes.

In addition, previous works have suggested that the W chromosome in Ditrysia is generated through noncanonical mechanisms (*5*, *9*). Here, we present a new mechanism for the generation of the W chromosome, which we named single-Z turnover. For example, in *M. jurtina*, we found a high homology between the W and Z chromosome but no neo-Z chromosome formation, which excludes the possibility that *M. jurtina* W occurred through one of the three previously proposed models (Fig. 3, C and D). We also found a lack of homology between *M. jurtina* W and chromosomes W of other closely related species (Fig. 5A), which suggests that the W of *M. jurtina* has probably originated independently. We thus propose that the W in *M. jurtina* may have appeared after one of the two Z chromosomes acquired a female-determining factor that leads to the direct transformation of that Z chromosome into the W chromosome (Fig. 4). Among the 17 ZW/ZZ systems, three species appear to have evolved their W chromosomes through this mechanism (Fig. 6).

We show also that the formation of the W chromosome in *C. pomonella* and *P. brassicae* involves a neo-W generated by Z-autosome fusion. This is supported by the high sequence homology



Female Karyotype (*n*)	Z-autosome fusion	Homology between Z and W
29 + ZW	Yes	No
30 + ZW	No	No
30 + ZW	No	No
30 + ZW	No	No
28 + ZW	No	YES
20 + ZW	Yes	No
29 + ZW	No	YES
14 + ZW	Yes	YES
30 + ZW	No	No
30 + ZW	No	No
30 + ZW	No	YES
28 + ZW	No	No
30 + ZW	No	No
29 + ZW	No	No
27 + ZW	Yes	Yes
29 + ZW	No	No
30 + ZW	No	No
29 + Z0	No	No

Fig. 6. Summary of the inferred origins of W chromosomes in species of Ditrysia. The color code for the triangles under each branch of the phylogenetic tree represent the mechanism by which the W chromosome is generated. The blue rectangles under each branch represent the generation of neo-Z through Z-autosome fusion events. The composition of the sex chromosomes for each species is displayed above each branch. The W of *C. pomonella* is derived from Z-autosome fusion and single Z turnover. The W2 of *M. jurtina* and the W3 of *A. cardamines* were formed by single Z turnover. The W4 of *P. brassicae* were generated through Z-autosome fusion mechanism. The Ws of the other 13 ZW systems are likely the result of recruitment of a B chromosome or an unclear mechanism. The W chromosomes of *B. selene*, *D. iulia*, *K. inachus*, and *V. cardui* are named W1 (green rectangle), a single origin for W chromosomes of the species. Similarly, the W chromosomes named W5 (gray rectangle) in three species of Noctuidae evolved from a single ancestor, and the W chromosomes, named W6 (purple rectangle), from the two species of Crambidae are also derived from a common ancestor. The W chromosomes of the remaining four species are named Wi (black rectangle) to indicate that we are unsure whether their W chromosomes are derived from a common ancestor due to their lack of homology. In addition, we are uncertain whether the W chromosomes, including W1, W5, W6, and Wi, generated by the noncanonical model are derived from a single ancestor because of lack of homology between them. The right side of the phylogenetic tree displays the sex chromosome composition of each species, whether Z-autosome fusion has occurred, and whether the W and Z chromosomes within the genome are homologous. The phylogenetic tree is from Fig. 3C.

observed between the W and Z chromosomes, occurred Z-autosome fusion, as well as higher number of synteny blocks between neo-W and neo-Z in *C. pomonella* and *P. brassicae*. A previous study attempted to detect sequence homology between the W and Z chromosomes in *C. pomonella* but did not find evidence of homology between these two chromosomes (*11*). Some possible reasons for the lack of evidence for homology between the W and Z chromosomes are the following. First, repetitive sequences present in the W chromosome were not masked before the synteny analysis, which results in the identification of predominantly short repetitive sequences as the homologous sequences between the W and Z chromosomes in the previous study (*11*). Transposons, the major components of repeat sequences of a genome, are mobile elements in the genome and constitute a major portion of the W chromosome, and their transposition and amplification between chromosomes can affect the determination of homology between the W chromosome and other chromosomes. Second, the authors relied on dot plot visualization to assess the collinearity between the W and Z chromosomes. However, dot plot graphs are typically used to depict collinearity relationships between highly conserved homologous chromosomes within or between closely related species. Last, the previous study only reported whether there is collinearity relationship between the W and Z chromosomes in *C. pomonella* but did not compare the degree of homology between the W chromosome and the other chromosomes. Nevertheless, our study conducted synteny analysis between the W chromosome excluding of repetitive sequences and other repeat-masked chromosomes using the number and total length of synteny blocks as indicators of homology between the W chromosome and other chromosomes. This ultimately revealed that *C. pomonella* W and Z chromosomes exhibit the highest homology compared to the homology between W and autosomes.

Furthermore, the W chromosome in Lepidoptera undergoes rapid evolution or degeneration, which leads to a lack of sequence homology between W chromosomes of distantly related species. Among the 17 ZW systems, we only observed obvious homology in three pairs of W chromosomes (*K. inachus* versus *V. cardui*, *S. exigua* versus *S. frugiperda*, and *Cnaphalocrocis exigua* versus *Cnaphalocrocis medinalis*), with each pair having a divergence time of less than 30 million years (fig. S25). Therefore, it is almost impossible to determine the mechanism of the origin of early W chromosomes in Ditrysia. Given the rapid evolution of W chromosomes, the observed single-Z turnover or neo-W likely represents recent events. Nonetheless, our results support that the existing W chromosomes in Ditrysia species have multiple independent origins through multiple mechanisms.

Last, we emphasize the importance of determining the origin of lepidopteran W based on an analysis of homology and potential generating mechanisms among W chromosomes of different species in the phylogenetic context rather than relying solely on comparisons between a few species or on the homology between W and Z chromosomes. For example, when comparing the W chromosomes of *D. iulia* and *B. selene*, we found no homology between them. However, within the Nymphalidae family, we found that the W chromosomes of *D. iulia* and *B. selene* share a high homology with the W of *V. cardui*. We then can conclude that the W chromosomes of *D. iulia* and *B. selene* originated from a common ancestor but have lost their homologous blocks during rapid evolution. Another example is observed in the Noctuidae family where the *P. meticulosa* W exhibits a high homology with its own Z chromosome. However, we have demonstrated that this is unrelated to the formation of the *P. meticulosa* W. Instead, the *P. meticulosa* W chromosome displays a high homology with two Spodoptera species, pointing to a possible shared origin. The primary reason for the high homology between *P. meticulosa* W and its own Z chromosome is likely due to the acquisition of W chromosome fragments by its Z chromosome during evolution.

MATERIALS AND METHODS

Sample source and DNA extraction

The *B. mori* strain Dazao/P50 used in this study was derived by inbreeding and obtained from the Silkworm Gene Bank, Southwest University, China. Previously, the male of this strain was used to generate the reference genome (*26*, *42*). The sample was grown at 25°C under 12 hour-light/12 hour-dark photoperiods. The whole genomic DNA of the female pupa was obtained by phenol-chloroform extraction.

DNA library construction and genome sequencing

DNA library and genome sequencing were performed by Bio-Marker Technologies Corporation, China. PacBio RS platform was used to obtain long-read sequence. A 20-kb DNA library was constructed and seven single-molecule real-time cells were used for PacBio sequencing. For Hi-C sequencing, a Hi-C library was constructed using the previous protocol (*51*). Paired-end (2 × 150) Hi-C reads were obtained using the Illumina HiSeq X platform.

Contig-level genome assembly

The genome was assembled by the following pipeline: (a) The initial contig-level assembly was performed using CANU v2.2 (*52*) with the parameter corOutCoverage = 1000. (b) The contigs were corrected three times with PacBio reads by racon v1.523 (*53*). We used minimap2 v2.24 (*54*) to map PacBio reads to the contigs. (c) NGS data were further used to polish and improve the genome assembly using Burrow-Wheeler Aligner (BWA) v0.7.17 (*55*) and Pilon v1.24 (*56*).

Chromosome-level genome assembly and evaluation

The raw Hi-C paired reads were removed from adapter sequences and low-quality paired-end reads using FASTP v0.23.2 (*57*). To avoid mismatches and nonspecific alignments, Hi-C reads were first aligned with the post-Polish PacBio Contigs using BWA. Only uniquely aligned paired reads were considered for further analysis. The ALLHiC pipeline (*58*) (https://github.com/tangerzhang/ALLHiC) was used to perform duplicate deletion, sorting, and quality assessment. Hi-C reads were then specifically compared to the corrected post-contig, clustered, sorted, and aligned contigs to chromosome-level sequences. The completeness of the genome assembly was assessed using BUSCO (Benchmarking Universal Single-Copy Orthologs) v5.12 with the lepidoptera_odb10 database. RNA sequencing (RNA-seq) and NGS short reads of *B. mori* genome were used to assess the integrity of the assembly quality. We performed whole-genome synteny analyses for the female genome assembled in this study and two male genomes downloaded from silkdb3.0 (https://silkdb.bioinfotoolkits.net/main/species-info/-1) and Silkbase (https://silkbase.ab.a.u-tokyo.ac.jp/cgi-bin/index.cgi). First, nucmer of MUMmer v3.23 package was used to conduct

whole-genome alignments. We then used RectChr v1.34 (https://github.com/BGI-shenzhen/RectChr) to visualize the synteny relationships among chromosomes.

Identification of W chromosome sequences and PCR validation

We developed a specific pipeline to identify sex chromosomes based on CQ value as follows: (a) Illumina short reads (pair-end) from a single male and female of the Dazao strain were mapped to the initial contigs separately using the BWA-MEM. (b) We used samtools v1.9 (*59*) with parameter -bq 1 to process the sam format file generated in step (a) and generated a bam format file that included only uniquely mapping reads. (c) The bam file was transformed to a bed file using bedtools bamtobed v2.26.0 (*60*). Perl scripts were written to scan the bed file to calculate CQ values for each chromosome. Last, the newly pipeline was submitted to GitHub (https://github.com/ruio-7/CQ-package) and Zenodo (https://sandbox.zenodo.org/records/24111).

To validate the *B. mori* candidate W-linked sequences, we assayed 15 candidate W specific fragments in three male and three female individuals of P50 strain using PCR. Because the W chromosome is rich in repetitive sequences, these 15 primer pairs were designed on the nonrepetitive sequences of the W chromosome. The primers used for these assays are listed in table S14. In addition, a previous report (*40*) on W-linked Fem precursor sequences was used as query in BLASTN (*e* value $<1 \times 10^{-5}$, identity >90%) search against the candidate W chromosome.

Identification of repetitive elements and protein-coding genes

RepeatModeler v2.0.2 (*61*) with -LTRStruct parameter was used to generate a de novo library of repetitive sequences. We then used RepeatMasker v4.1.2 and the de novo library to annotate and mask repetitive elements in the genome. We then combined three approaches including ab initio, homology-based, and transcriptome-based predictions to identify protein-coding genes and their arrangements in the female genome of *B. mori*. For the ab initio prediction, we used augustus v3.3 (*62*) with default parameters to predict genes based on the training set obtained from previous full-length transcriptome (*63*). For the homology-based prediction, we used GeMoMa v1.7.1 (*64*) with default parameters to perform homology-based prediction based on protein sequence of male *B. mori*, *D. melanogaster*, *Apis mellifera*, and *Danaus plexippus*. Protein sequences of male *B. mori* were downloaded from SilkDB3.0 (https://silkdb.bioinfotoolkits.net/main/species-info/-1) and SilkBase (https://silkbase.ab.a.u-tokyo.ac.jp/cgi-bin/index.cgi). Protein sequences of *D. melanogaster*, *A. mellifera*, and *D. plexippus* were downloaded from the NCBI. For the transcriptome-based prediction, our previously released full-length NR transcripts obtained from almost all developmental stages of Dazao strain were aligned to the *B. mori* female genome using GMAP v1.0.0 (*65*) (https://github.com/EagleGenomics-cookbooks/GMAP/releases/tag/1.0.0) with --cross-species --allow-close-indels 0 parameters. The cDNA_Cupcake (https://github.com/Magdoll/cDNA_Cupcake/wiki) was used to remove redundancy and filter sequences with less than 0.9 identity and less than 0.85 coverage. Open reading frames in the transcripts were predicted using PASA v2.4.1 (*66*). Last, gene structures from the above three approaches were merged to generate a comprehensive NR gene set using EVidenceModeler (EVM) v1.1.1 (*67*) with the ab initio prediction weight of "1," the homology-based prediction weight of "5," and the transcriptome-based prediction weight of "10."

Filtering transposon-coding genes

To filter transposon-coding genes, we first used hmmscan with *e* value $<1 \times 10^{-5}$ to search against the hidden Karkov model (HMM) profiles associated with each known transposon family (class-I and class-II RNA transposons) in the Pfam database and filtered the matched proteins. For either class, the detailed HMM models of transposon were listed in table S15. For both classes, we used the known transposon-coding genes from the repbase database (v2018) as query to perform a BlastP search with *E* value threshold of 1×10^{-5} against the predicted proteins of the female *B. mori* genome and discarded the hits. Last, we took all predicted proteins as query to conduct Diamond blast search with *E* value $<1 \times 10^{-5}$ against the NCBI NR proteins database and filtered transposon-coding genes.

Comparative analysis of the predicted genes from different genomes of the same strain

To determine the correspondence between genes in the different sets, we first used minimap2 with parameters -ax splice --secondary = no to align the coding DNA sequences (CDSs) of the prior male genomes with our assembled female genome. This allowed us to determine the locus of each gene from the three predicted gene sets. Subsequently, we used cd-hit with the parameters of -c 0.9 -G 0 -aS 0.5 to cluster the predicted proteins from the three gene sets. Last, we established the correlation between predicted genes by comparing their genomic loci and the results of clustering. Briefly, we define genes as the same if they are located at the same genomic locus and belong to the same cluster. If the genomic locus of our predicted gene excluded any of the previously predicted genes, we defined our predicted gene as a newly identified one. In addition, if genes predicted from different genomes were located at the same genomic locus or overlap but belong to the different clusters, we classify these genes as different.

Gene function annotation and expression

For gene function annotation, we aligned all protein-coding sequences to the NCBI NR proteins database using Diamond blastp with *E* value $<1 \times 10^{-5}$. For GO and pathway annotations (Kyoto Encyclopedia of Genes and Genomes), we used Diamond with *E* value $<1 \times 10^{-5}$ to align eggNOG database v5.0. To predict conserved domains of each gene, we used hmmscan with *E* value $<1 \times 10^{-5}$ and Pfam-A models (as of 11 August 2022) to search all protein sequences.

To characterize gene expression, we downloaded previous released RNA-seq data of *B. mori* from NCBI sequence read archive (accession no. DRR013821.1-DRR030438.1, SRR10035581-SRR10035833, and SRR5602461.1-SRR5602472.1) and used STAR v2.7.10 (*68*) to align RNA-seq reads to the *Dazao* female genome. We then used Stringtie v1.3.0 (*69*) to calculate expression level (TPM) of each gene. We considered genes to be transcribed if the expression level more than 0.5 TPM in at least one sample.

Phylogenetic analysis

In addition to *B. mori*, the genomes and proteins of 23 species, including 3 trichopteran and 20 lepidopteran species, were downloaded from NCBI and InsectBase database (table S10). The single-copy BUSCOs of each species were identified using BUSCO v5.12 with the insect_odb10. A total of 833 single-copy BUSCOs shared

by the detected 24 species were aligned using MAFFT v7.487 with default parameters. We then used IQ-TREE v2.0.3 (*70*) with the parameters of alrt 1000 -bb 1000 to construct the maximum likelihood (ML) tree. Last, the ML tree was used to estimate the divergence times among 24 genomes by r8s v1.81. The phylogenetic tree was visualized in ITOL (https://itol.embl.de/).

Synteny analysis among Z chromosomes and autosomes

We analyzed the synteny relationships among Z chromosomes and autosomes of the 24 samples based on orthologous genes. Ten of the 24 species did not have genome-wide gene annotations. Thus, we used homology-based methods to predict protein-coding genes in these genomes. We used GeMoMa with the default parameters to predict genes based on protein sequences of four species with high BUSCO scores, including *M. jurtina*, *P. brassicae*, *S. frugiperda*, and *B. mori*. We then used OrthoFinder v2.5.4 (*71*) with default parameters to identify orthologous genes in all 24 species. The jcvi package v1.2.7 (https://github.com/tanghaibao/jcvi) was used to display the synteny relationships among the Z chromosomes and autosomes based on the locations of orthologous genes. In addition, the original chromosome numbering of each species used in this study does not have a correspondence between species. To facilitate comparative analysis of chromosomes in subsequent analyses, we adopted a unified nomenclature for the chromosomes of these species based on recently reported Merian elements (*48*). The original chromosome numbering for each species and their corresponding Merian elements are presented in table S16.

Comparisons between W chromosomes and other chromosomes

We used Satsuma2synteny to identify synteny blocks between chromosomes. Because of the scarcity of genes in the W chromosome, we conducted synteny analysis based on the nucleotide sequence comparisons between the W chromosome and other chromosomes. In addition, the mobility of TEs may have an impact on the detection of sequence homology and arrangement between chromosomes, particularly between W chromosomes that are rich in transposons. To eliminate the influence of TEs on the results of synteny analyses, we identified synteny blocks between chromosomes after filtering out repetitive sequences. Considering the variation in the length of W chromosomes among different species, ranging from 1.87 to 15.17 Mb (table S10), we conducted reciprocal comparisons for identification of synteny blocks between W chromosomes. For instance, in the comparison between species A and species B, one direction of comparison involved using the W chromosome sequence of species A as a query sequence to search for homologous sequences in each chromosome of species B. The other direction of comparison involved using the W of species B as a query to search for homologous regions in each chromosome of species A. The synteny blocks between chromosomes were identified using Satsuma2synteny (https://github.com/bioinfologics/satsuma2) with the minimum alignment length set to the default value (0), 100, and 500 bp, respectively. Considering that when the minimum alignment length is set to the default parameter, there are many very short synteny (<50 bp) blocks between different chromosomes (fig. S26), this study also analyzed the number and total length of homology blocks between different chromosomes when the minimum alignment length was set to 100 and 500 bp.

To investigate the effectiveness of Satsuma2synteny in detecting homology between chromosomes of different species, we also used this tool to identify homology between autosomes and between Z chromosomes of different species as positive control. We found that Satsuma2Synteny was effective in identifying homology between Z chromosomes or autosomes of different species, regardless of whether the minimum alignment length was set to the default value or 1000 bp (figs. S27 to S30). Last, the LINKVIEW v2 (https://yangjianshun.github.io/LINKVIEW2/) was used to display the synteny relationships between chromosomes based on the locations of synteny blocks.

Supplementary Materials

This PDF file includes:

Figs. S1 to S30
Tables S1 to S6, S11 and S14
Legends for tables S7 to S10, S12, S13, S15 and S16

Other Supplementary Material for this manuscript includes the following:

Tables S7 to S10, S12, S13, S15 and S16

REFERENCES AND NOTES

1. W. Traut, K. Sahara, F. Marec, Sex chromosomes and sex determination in Lepidoptera. *Sex. Dev.* **1**, 332–346 (2008).
2. K. Sahara, A. Yoshido, W. Traut, Sex chromosome evolution in moths and butterflies. *Chromosome Res.* **20**, 83–94 (2012).
3. J. B. Pease, M. W. Hahn, Sex chromosomes evolved from independent ancestral linkage groups in winged insects. *Mol. Biol. Evol.* **29**, 1645–1653 (2012).
4. Y. Yasukochi, M. T. Okuyama, F. Shibata, A. Yoshido, F. Marec, C. Wu, H. Zhang, M. R. Goldsmith, K. Sahara, Extensive conserved synteny of genes between the karyotypes of *Manduca sexta* and *Bombyx mori* revealed by BAC-FISH mapping. *PLOS ONE* **4**, e7465 (2009).
5. C. Fraisse, 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).
6. W. Traut, F. Marec, Sex chromatin in lepidoptera. *Q. Rev. Biol.* **71**, 239–256 (1996).
7. V. G. Kuznetsova, S. Nokkala, A. Maryanska-Nadachowska, Karyotypes, sex chromosome systems, and male meiosis in *Finnish psyllids* (Homoptera : Psylloidea). *Folia Biol-Krakow* **45**, 143–152 (1997).
8. M. Dalíková, M. Zrzavá, I. Hladová, P. Nguyen, I. Šonský, M. Flegrová, S. Kubícková, A. Voleníková, A. Y. Kawahara, R. S. Peters, F. Marec, New Insights into the evolution of the W chromosome in Lepidoptera. *J. Hered.* **108**, 709–719 (2017).
9. J. J. Lewis, F. Cicconardi, S. H. Martin, R. D. Reed, C. G. Danko, S. H. Montgomery, The Dryas iulia genome supports multiple gains of a W chromosome from a B chromosome in butterflies. *Genome Biol. Evol.* **13**, evab128 (2021).
10. Y. Fu, Y. Yang, H. Zhang, G. Farley, J. Wang, K. A. Quarles, Z. Weng, P. D. Zamore, The genome of the Hi5 germ cell line from *Trichoplusia ni*, an agricultural pest and novel model for small RNA biology. *eLife* **7**, e31628 (2018).
11. F. Wan, C. Yin, R. Tang, M. Chen, Q. Wu, C. Huang, W. Qian, O. Rota-Stabelli, N. Yang, S. Wang, G. Wang, G. Zhang, J. Guo, L. A. Gu, L. Chen, L. Xing, Y. Xi, F. Liu, K. Lin, M. Guo, W. Liu, K. He, R. Tian, E. Jacquin-Joly, P. Franck, M. Siegwart, L. Ometto, G. Anfora, M. Blaxter, C. Meslin, P. Nguyen, M. Dalíková, F. Marec, J. Olivares, S. Maugin, J. Shen, J. Liu, J. Guo, J. Luo, B. Liu, W. Fan, L. Feng, X. Zhao, X. Peng, K. Wang, L. Liu, H. Zhan, W. Liu, G. Shi, C. Jiang, J. Jin, X. Xian, S. Lu, M. Ye, M. Li, M. Yang, R. Xiong, J. R. Walters, F. Li, A chromosome-level genome assembly of *Cydia pomonella* provides insights into chemical ecology and insecticide resistance. *Nat. Commun.* **10**, 4237 (2019).
12. K. S. Singh, D. J. Hosken, N. Wedell, R. Ffrench-Constant, C. Bass, S. Baxter, K. Paszkiewicz, M. D. Sharma, De novo genome assembly of the meadow brown butterfly, *Maniola jurtina*. *G3* **10**, 1477–1484 (2020).
13. J. Yang, W. Wan, M. Xie, J. Mao, Z. Dong, S. Lu, J. He, F. Xie, G. Liu, X. Dai, Z. Chang, R. Zhao, R. Zhang, S. Wang, Y. Zhang, W. Zhang, W. Wang, X. Li, Chromosome-level reference genome assembly and gene editing of the dead-leaf butterfly *Kallima inachus*. *Mol. Ecol. Resour.* **20**, 1080–1092 (2020).
14. B. Zhang, B. Liu, C. Huang, L. Xing, Z. Li, C. Liu, H. Zhou, G. Zheng, J. Li, J. Han, Q. Yu, C. Yang, W. Qian, F. Wan, C. Li, A chromosome-level genome assembly for the beet armyworm (Spodoptera exigua) using PacBio and Hi-C sequencing. bioRxiv 2019.12.26.889121 [Preprint] (2020). https://doi.org/10.1101/2019.12.26.889121.
15. K. Lohse, A. Mackintosh; Darwin Tree of Life Barcoding collective; Wellcome Sanger Institute Tree of Life programme; Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective; Tree of Life Core Informatics collective; Darwin Tree of Life

Consortium, The genome sequence of the large white, *Pieris brassicae* (Linnaeus, 1758). *Wellcome Open Res.* **6**, 262 (2021).
16. K. Lohse, D. Setter; Darwin Tree of Life Barcoding collective; Wellcome Sanger Institute Tree of Life programme; Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective; Tree of Life Core Informatics collective; Darwin Tree of Life Consortium, The genome sequence of the small pearl-bordered fritillary butterfly, *Boloria selene* (Schiffermuller, 1775). *Wellcome Open Res.* **7**, 76 (2022).
17. K. Lohse, J. Weir; Darwin Tree of Life Barcoding collective; Wellcome Sanger Institute Tree of Life programme; Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective; Tree of Life Core Informatics collective; Darwin Tree of Life Consortium, The genome sequence of the meadow brown, *Maniola jurtina* (Linnaeus, 1758). *Wellcome Open Res.* **6**, 296 (2021).
18. L. Zhang, R. A. Steward, C. W. Wheat, R. D. Reed, High-quality genome assembly and comprehensive transcriptome of the painted lady butterfly *Vanessa cardui*. *Genome Biol. Evol.* **13**, evab145 (2021).
19. X. Zhao, H. Xu, K. He, Z. Sh, X. Chen, X. Ye, Y. Mei, Y. Yang, M. Li, L. Gao, L. Xu, H. Xiao, Y. Liu, Z. Lu, F. Li, A chromosome-level genome assembly of rice leaffolder, Cnaphalocrocis medinalis. *Mol. Ecol. Resour.* **21**, 561–572 (2021).
20. D. Boyes, P. W. H. Holland; University of Oxford and Wytham Woods Genome Acquisition Lab; Darwin Tree of Life Barcoding collective; Wellcome Sanger Institute Tree of Life programme; Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective; Tree of Life Core Informatics collective; Darwin Tree of Life Consortium, The genome sequence of the angle shades moth, *Phlogophora meticulosa* (Linnaeus, 1758). *Wellcome Open Res.* **7**, 89 (2022).
21. S. Ebdon, G. Bisschop, K. Lohse, I. Saccheri, J. Davies; Wellcome Sanger Institute Tree of Life programme; Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective; Tree of Life Core Informatics collective; Darwin Tree of Life Consortium, The genome sequence of the orange-tip butterfly, *Anthocharis cardamines* (Linnaeus, 1758). *Wellcome Open Res.* **7**, 260 (2022).
22. H. Xu, X. Zhao, Y. Yang, X. Chen, Y. Me, K. He, L. Xu, X. Ye, Y. Liu, F. Li, Z. Lu, Chromosome-level genome assembly of an agricultural pest, the rice leaffolder *Cnaphalocrocis exigua* (Crambidae, Lepidoptera). *Mol. Ecol. Resour.* **22**, 307–318 (2022).
23. J. Gauthier, J. Meier, F. Legeai, M. McClure, A. Whibley, A. Bretaudeau, H. Boulain, H. Parrinello, S. T. Mugford, R. Durbin, C. Zhou, S. McCarthy, C. W. Wheat, F. Piron-Prunier, C. Monsempes, M.-C. François, P. Jay, C. Noûs, E. Persyn, E. Jacquin-Joly, C. Meslin, N. Montagné, C. Lemaitre, M. Elias, First chromosome scale genomes of ithomiine butterflies (Nymphalidae: Ithomiini): Comparative models for mimicry genetic studies. *Mol. Ecol. Resour.* **23**, 872–885 (2023).
24. Y. Tazima, *The Genetics of the Silkworm* (Academic Press, 1964).
25. M. R. Goldsmith, T. Shimada, H. Abe, The genetics and genomics of the silkworm, bombyx mori. *Annu. Rev. Entomol.* **50**, 71–100 (2005).
26. K. Mita, M. Kasahara, S. Sasaki, Y. Nagayasu, T. Yamada, H. Kanamori, N. Namiki, M. Kitagawa, H. Yamashita, Y. Yasukochi, K. Kadono-Okuda, K. Yamamoto, M. Ajimura, G. Ravikumar, M. Shimomura, Y. Nagamura, T. Shin-I, H. Abe, T. Shimada, S. Morishita, T. Sasaki, The genome sequence of silkworm, Bombyx mori. *DNA Res.* **11**, 27–35 (2004).
27. Q. Xia, Z. Zhou, C. Lu, D. Cheng, F. Dai, B. Li, P. Zhao, X. Zha, T. Cheng, C. Chai, G. Pan, J. Xu, C. Liu, Y. Lin, J. Qian, Y. Hou, Z. Wu, G. Li, M. Pan, C. Li, Y. Shen, X. Lan, L. Yuan, T. Li, H. Xu, G. Yang, Y. Wan, Y. Zhu, M. Yu, W. Shen, D. Wu, Z. Xiang, J. Yu, J. Wang, R. Li, J. Shi, H. Li, G. Li, J. Su, X. Wang, G. Li, Z. Zhang, Q. Wu, J. Li, Q. Zhang, N. Wei, J. Xu, H. Sun, L. Dong, D. Liu, S. Zhao, X. Zhao, Q. Meng, F. Lan, X. Huang, Y. Li, L. Fang, C. Li, D. Li, Y. Sun, Z. Zhang, Z. Yang, Y. Huang, Y. Xi, Q. Qi, D. He, H. Huang, X. Zhang, Z. Wang, W. Li, Y. Cao, Y. Yu, H. Yu, J. Li, J. Ye, H. Chen, Y. Zhou, B. Liu, J. Wang, J. Ye, H. Ji, S. Li, P. Ni, J. Zhang, Y. Zhang, H. Zheng, B. Mao, W. Wang, C. Ye, S. Li, J. Wang, G. K.-S. Wong, H. Yang; Biology Analysis Group, A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science* **306**, 1937–1940 (2004).
28. Q. Xia, Y. Guo, Z. Zhang, D. Li, Z. Xuan, Z. Li, F. Dai, Y. Li, D. Cheng, R. Li, T. Cheng, T. Jiang, C. Becquet, X. Xu, C. Liu, X. Zha, W. Fan, Y. Lin, Y. Shen, L. Jiang, J. Jensen, I. Hellmann, S. Tang, P. Zhao, H. Xu, C. Yu, G. Zhang, J. Li, J. Cao, S. Liu, N. He, Y. Zhou, H. Liu, J. Zhao, C. Ye, Z. Du, G. Pan, A. Zhao, H. Shao, W. Zeng, P. Wu, C. Li, M. Pan, J. Li, X. Yin, D. Li, J. Wang, H. Zheng, W. Wang, X. Zhang, S. Li, H. Yang, C. Lu, R. Nielsen, Z. Zhou, J. Wang, Z. Xiang, J. Wang, Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (Bombyx). *Science* **326**, 433–436 (2009).
29. H. Xiang, X. Liu, M. Li, Y. Zhu, L. Wang, Y. Cui, L. Liu, G. Fang, H. Qian, A. Xu, W. Wang, S. Zhan, The evolutionary road from wild moth to domestic silkworm. *Nat. Ecol. Evol.* **2**, 1268–1279 (2018).
30. X. Tong, M.-J. Han, K. Lu, S. Tai, S. Liang, Y. Liu, H. Hu, J. Shen, A. Long, C. Zhan, X. Ding, S. Liu, Q. Gao, B. Zhang, L. Zhou, D. Tan, Y. Yuan, N. Guo, Y.-H. Li, Z. Wu, L. Liu, C. Li, Y. Lu, T. Gai, Y. Zhan, R. Yang, H. Qian, Y. Liu, J. Luo, L. Zheng, J. Lou, Y. Peng, W. Zuo, J. Song, S. He, S. Wu, Y. Zou, L. Zhou, L. Cheng, Y. Tang, G. Cheng, L. Yuan, W. He, J. Xu, T. Fu, Y. Xiao, T. Lei, A. Xu, Y. Yin, J. Wang, A. Monteiro, E. Westhof, C. Lu, Z. Tian, W. Wang, Z. Xiang, F. Dai, High-resolution silkworm pan-genome provides genetic insights into artificial selection and ecological adaptation. *Nat. Commun.* **13**, 5619 (2022).
31. H. Abe, K. Mita, Y. Yasukochi, T. Oshiki, T. Shimada, Retrotransposable elements on the W chromosome of the silkworm, Bombyx mori. *Cytogenet. Genome Res.* **110**, 144–151 (2005).
32. S. Kawaoka, K. Kadota, Y. Arai, Y. Suzuki, T. Fujii, H. Abe, Y. Yasukochi, K. Mita, S. Sugano, K. Shimizu, Y. Tomari, T. Shimada, S. Katsuma, The silkworm W chromosome is a source of female-enriched piRNAs. *RNA* **17**, 2144–2151 (2011).
33. S. Li, M. Ajimura, Z. Chen, J. Liu, E. Chen, H. Guo, V. Tadapatri, C. G. Reddy, J. Zhang, H. Kishino, H. Abe, Q. Xia, K. P. Arunkumar, K. Mita, A new approach for comprehensively describing heterogametic sex chromosomes. *DNA Res.* **25**, 375–382 (2018).
34. H. Sakai, M. Sumitani, Y. Chikami, K. Yahata, K. Uchino, T. Kiuchi, S. Katsuma, F. Aoki, H. Sezutsu, M. G. Suzuki, Transgenic Expression of the piRNA-resistant masculinizer gene induces female-specific lethality and partial female-to-male sex reversal in the Silkworm. *PLOS Genet.* **12**, e1006203 (2016).
35. X. Chen, Y. Cao, S. Zhan, A. Tan, S. R. Palli, Y. Huang, Disruption of sex-specific doublesex exons results in male- and female-specific defects in the black cutworm, Agrotis ipsilon. *Pest Manag. Sci.* **75**, 1697–1706 (2019).
36. X. Li, Q. Liu, H. Liu, H. Bi, Y. Wang, X. Chen, N. Wu, J. Xu, Z. Zhang, Y. Huang, H. Chen, Mutation of doublesex in *Hyphantria cunea* results in sex-specific sterility. *Pest Manag. Sci.* **76**, 1673–1682 (2020).
37. X. Yang, K. Chen, Y. Wang, D. Yang, Y. Huang, The sex determination cascade in the silkworm. *Genes* **12**, 315 (2021).
38. M. Han, J. Ren, H. Guo, X. Tong, H. Hu, K. Lu, Z. Dai, F. Dai, Mutation rate and spectrum of the silkworm in normal and temperature stress conditions. *Genes* **14**, 649 (2023).
39. A. B. Hall, Y. Qi, V. Timoshevskiy, M. V. Sharakhova, I. V. Sharakhov, Z. Tu, Six novel Y chromosome genes in *Anopheles mosquitoes* discovered by independently sequencing males and females. *BMC Genomics* **14**, 273 (2013).
40. T. Kiuchi, H. Koga, M. Kawamoto, K. Shoji, H. Sakai, Y. Arai, G. Ishihara, S. Kawaoka, S. Sugano, T. Shimada, Y. Suzuki, M. G. Suzuki, S. Katsuma, A single female-specific piRNA is the primary determiner of sex in the silkworm. *Nature* **509**, 633–636 (2014).
41. S. Katsuma, T. Kiuchi, M. Kawamoto, T. Fujimoto, K. Sahara, Unique sex determination system in the silkworm, *Bombyx mori*: Current status and beyond. *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci.* **94**, 205–216 (2018).
42. M. Kawamoto, A. Jouraku, A. Toyoda, K. Yokoi, Y. Minakuchi, S. Katsuma, A. Fujiyama, T. Kiuchi, K. Yamamoto, T. Shimada, High-quality genome assembly of the silkworm, Bombyx mori. *Insect. Biochem. Mol. Biol.* **107**, 53–62 (2019).
43. F. Lu, Z. Wei, Y. Luo, H. Guo, G. Zhang, Q. Xia, Y. Wang, SilkDB 3.0: Visualizing and exploring multiple levels of data for silkworm. *Nucleic Acids Res.* **48**, D749–D755 (2020).
44. W. Traut, F. Marec, Sex chromatin and sex chromosome systems in nonditrysian Lepidoptera (Insecta). *J. Zool. Syst. Evol. Res.* **38**, 73–79 (2000).
45. M. J. Pokorná, R. Reifová, Evolution of B chromosomes: From dispensable parasitic chromosomes to essential genomic players. *Front. Genet.* **12**, 727570 (2021).
46. B. Vicoso, D. Bachtrog, Numerous transitions of sex chromosomes in Diptera. *PLOS Biol.* **13**, e1002078 (2015).
47. P. Nguyen, M. Sýkorová, J. Šíchová, V. Kůta, M. Dalíková, R. Č. Frydrychová, L. G. Neven, K. Sahara, F. Marec, Neo-sex chromosomes and adaptive potential in tortricid pests. *Proc. Natl Acad. Sci. U.S.A.* **110**, 6931–6936 (2013).
48. C. J. Wright, L. Stevens, A. Mackintosh, M. Lawniczak, M. Blaxter, Chromosome evolution in Lepidoptera. bioRxiv 2023.05.12.540473 [Preprint] (2023). https://doi.org/10.1101/2023.05.12.540473.
49. D. Berner, S. Ruffener, L. A. Blattner, Chromosome-level assemblies of the *Pieris mannii* butterfly genome suggest Z-origin and rapid evolution of the W chromosome. *Genome Biol. Evol.* **15**, evad111 (2023).
50. X. Chen, Z. Wang, C. Zhang, J. Hu, Y. Lu, H. Zhou, Y. Mei, Y. Cong, F. Guo, Y. Wang, K. He, Y. Liu, F. Li, Unraveling the complex evolutionary history of lepidopteran chromosomes through ancestral chromosome reconstruction and novel chromosome nomenclature. *BMC Biol.* **21**, 265 (2023).
51. J.-M. Belton, R. P. McCord, J. H. Gibcus, N. Naumova, Y. Zhan, J. Dekker, Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).
52. S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, A. M. Phillippy, Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
53. R. Vaser, I. Sovic, N. Nagarajan, M. Sikic, Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
54. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
55. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
56. B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, A. M. Earl, Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE* **9**, e112963 (2014).

57. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
58. X. Zhang, S. Zhang, Q. Zhao, R. Ming, H. Tang, Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).
59. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
60. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
61. J. M. Flynn, R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte, A. F. Smit, RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9451–9457 (2020).
62. S. Nachtweide, M. Stanke, Multi-genome annotation with AUGUSTUS. *Methods Mol. Biol.* **1962**, 139–160 (2019).
63. Z. Dai, J. Ren, X. Tong, H. Hu, K. Lu, F. Dai, M.-J. Han, The landscapes of full-length transcripts and splice isoforms as well as transposons exonization in the lepidopteran model system. *Front. Genet.* **12**, 704162 (2021).
64. J. Keilwagen, F. Hartung, M. Paulini, S. O. Twardziok, J. Grau, Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinformatics* **19**, 189 (2018).
65. T. D. Wu, C. K. Watanabe, GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
66. B. J. Haas, A. L. Delcher, S. M. Mount, J. R. Wortman, R. K. Smith Jr., L. I. Hannick, R. Maiti, C. M. Ronning, D. B. Rusch, C. D. Town, S. L. Salzberg, O. White, Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
67. B. J. Haas, S. L. Salzberg, W. Zhu, M. Pertea, J. E. Allen, J. Orvis, O. White, C. R. Buell, J. R. Wortman, Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
68. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
69. M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, S. L. Salzberg, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
70. L. T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
71. D. M. Emms, S. Kelly, OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).

Acknowledgments: We are grateful to C. Feschotte at Cornell University for helpful discussion and to all members of F.D.'s laboratory for assistance. **Funding:** This study was supported in part by the National Natural Science Foundation of China no. 31830094 (F.D.), National Natural Science Foundation of China no. 32272940 (M.-J.H.), National Natural Science Foundation of China no. U20A2058 (X.T.), Natural Science Foundation of Chongqing, China no. cstc2021jcyj-cxtt0005 (F.D.), China Agriculture Research System of MOF and MARA no. CARS-18-743 ZJ0102 (F.D.), Creative Research Group of the Natural Science Foundation of Chongqing (F.D.), and Natural Science Foundation of Chongqing no. cstc2020jcyj-msxmX0450 (M.-J.H.). **Author contributions:** Conceptualization: F.D., X.T., and M.-J.H. Methodology: M.-J.H., C.L., X.T., H.H., J.R., K.L., and M.L. Investigation: M.-J.H., C.L., K.L., Y.Y., J.R., and J.S. Visualization: M.-J.H. and C.L. Supervision: F.D., X.T., and M.-J.H. Writing—original draft: M.-J.H. and CL. Writing—review and editing: F.D., E.W., X.T., and M.-J.H. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The female *B. mori* genome project has been deposited at the BIG Data Center (http://bigd.big.ac.cn), Beijing Institute of Genomics, Chinese Academy of Sciences, under BioProject: PRJCA017032 (https://ngdc.cncb.ac.cn). Raw data of PacBio sequencing and Hi-c sequencing used in this study are deposited into the Genome Sequence Archive (GSA) under accession code CRA011073 (https://ngdc.cncb.ac.cn). The genome assembly and annotation in this paper have been deposited into Genome Warehouse (GWH) under accession code GWHCBIK00000000.1 (https://ngdc.cncb.ac.cn/search/?dbId=gwh&q=GWHCBIK00000000.1). The tool for the identification of the W chromosome is submitted in Github (https://github.com/ruio-7/CQ-package) and Zenodo (https://sandbox.zenodo.org/records/24111).

Submitted 16 November 2023
Accepted 14 May 2024
Published 19 June 2024
10.1126/sciadv.adm9851