Molecular Biology and Evolution, 2026, 43, 1–21
https://doi.org/10.1093/molbev/msag055
Published: 4 March 2026
Discoveries

MBE OXFORD

From Adam's rib: the Siberian moth's female W chromosome is derived from its Z counterpart

Aleksandra A. Shipova [1], Irina A. Belousova [2,3], Maria E. Yakimova [2,3], Natalia I. Kirichenko [4,5], Aleksandr A. Ageev [6], Michel Cusson [7], Vladimir A. Lukhtanov [8], Nikita I. Ershov *[1,2], Vyacheslav V. Martemyanov *[2,3]

[1]Institute of Cytology and Genetics SB RAS, Novosibirsk 630090, Russia
[2]Sirius University of Science and Technology, Sochi 354349, Russia
[3]Institute of Systematics and Ecology of Animals SB RAS, Novosibirsk 630091, Russia
[4]Sukachev Institute of Forest, Siberian Branch of the Russian Academy of Sciences, Federal Research Center "Krasnoyarsk Science Center SB RAS", Akademgorodok 50/28, Krasnoyarsk 660036, Russia
[5]All-Russian Plant Quarantine Center (VNIIKR), Krasnoyarsk Branch, st. Zhelyabova, 6/6, Krasnoyarsk 660020, Russia
[6]Center of Forest Pyrology, All-Russia Research Institute of Silviculture and Mechanization of Forestry, Krupskoy 42, Krasnoyarsk 660062, Russia
[7]Department of Biochemistry, Microbiology and Bioinformatics, Laval University, Quebec City, Quebec, Canada
[8]Department of Karyosystematics, Zoological Institute of Russian Academy of Science, St. Petersburg, Russia
***Corresponding authors:** E-mails: nikotinmail@mail.ru; martemyanov79@yahoo.com.
Associate editor: Amanda Larracuente

Abstract

Most lepidopteran species have a WZ/ZZ sex chromosome system, suggesting a common origin. However, the exceptional variability of W chromosomes supports their multiple independent origins. Here, we investigated the genome of the recently formed species, *Dendrolimus sibiricus*, which is one of the most devastating pests of the boreal forests in Northern Asia. We found that its karyotype consists of 30 pairs of chromosomes and follows the WZ/ZZ sex chromosome system. We produced a chromosome-scale assembly of its female genome and identified the sequences of both sex chromosomes. With the exception of the W chromosome, the *D. sibiricus* genome displays a very high degree of synteny with genomes of other *Dendrolimus* representatives. The non-conserved W chromosome is composed entirely of recently acquired sequences, mainly young genomic repeats and a few unique loci derived from subtelomeric regions of the same species' Z chromosome as well as from the genome of the endosymbiotic *Wolbachia* bacterium. It shares no common ancestral sequences with the W chromosomes of two related species, *Dendrolimus tabulaeformis* and *Lasiocampa quercus*. They, in turn, display a similar organization relative to the W of *D. sibiricus*, including Z-derived loci. The W of *D. tabulaeformis* also contained multiple fragments of the mitochondrial genome, as seen in classical B chromosomes. We conclude that W chromosomes arose from their Z counterparts recently and independently in these three species. We propose that this mechanism may be a common feature of lepidopteran evolution, counteracting the inevitable degeneration of a non-recombining chromosome through its repeated, complete turnover.

Keywords evolution of sex chromosomes, W chromosome turnover, Siberian moth, *Wolbachia*, genome assembly, Lepidoptera

Introduction

One-tenth of all described species on Earth are Lepidoptera (Kristensen et al. 2007; Mora et al. 2011), whose karyotype is typically represented by 31 pairs of chromosomes (Robinson 1971). The genome organization of the currently known butterflies and moths is quite conservative, which made it possible to reconstruct the set of lepidopteran ancestral linkage groups, termed Merian elements: 31 autosomes and the Z sex chromosome (Wright et al. 2024). However, the W sex chromosome falls out of the list of these chromosomes, displaying exceptional variation in length, composition and number, including its complete absence in some species (Traut et al. 2007; Hejníčková et al. 2021).

The W is typically heterochromatic, gene poor, and containing highly repetitive sequences; it is maternally inherited and is a component of the WZ/ZZ (female/male) sex chromosome system prevalent in the Lepidoptera (Traut et al. 2007). Based on a large body of cytogenetic studies, it is believed that this system evolved early in the order from the ancestral Z0/ZZ system characteristic of the basal lepidopteran lineages and sister Trichoptera order, with the W arising independently on at least three occasions in Hepialidae, Tischeriidae, and Ditrysia (Lukhtanov 2000; Voleníková et al. 2023). At the same time,

Received: September 18, 2025. **Revised:** January 11, 2026. **Accepted:** February 20, 2026

sporadic rearrangements, gains and losses of W chromosomes are frequent in the Lepidoptera (Traut et al. 2007). Not surprisingly, the lack of gene conservation among W chromosomes gave rise to the alternative hypothesis that it appeared independently in a variety of different lepidopteran taxa (Lewis et al. 2021; Han et al. 2024). However, authors of the latter hypothesis did not elaborate on the difference between the general origin of the WZ/ZZ system in the Lepidoptera (including the molecular mechanisms of W condensation) and the subsequent rearrangements and turnover of W chromosomes in extant taxa.

No less controversial are the mechanisms underlying the rise and subsequent transformations of W chromosomes. Currently considered are the canonical model of WZ/ZZ origin from a pair of autosomes (Z-autosome fusion or complete turnover of sex chromosomes) and the non-canonical model of supernumerary B chromosome recruitment (Lukhtanov 2000; Fraïsse et al. 2017). More recently, a model of W origin from a copy of Z (single-Z turnover) has also been proposed based on comparative levels of inter-chromosomal homology across a set of lepidopteran genomes (Han et al. 2024). Nevertheless, the observed inter-chromosomal synteny appears to reflect only evolutionary recent events (Han et al. 2024), and the evidence supporting a particular model remains somewhat conjectural in the context of rapid W degeneration.

Existing ideas about the evolution of the lepidopteran W chromosome are not yet connected to its functional impact. As opposed to the Z-counting mechanism in Z0/ZZ species, it is often assumed that the W chromosome in WZ/ZZ species carries a female-determining factor. However, the dominant role of W in determining female sex and its underlying mechanisms have been demonstrated only in a few species, including *Bombyx mori*, *Plutella xylostella*, and *Lymantria dispar* (Kiuchi et al. 2014; Harvey-Samuel et al. 2022; Moronuki et al. 2025). On the other hand, examples have been presented where it does not play any role in sex determination or reproduction, e.g. in *Samia cynthia* ssp. (Yoshido and Marec 2023). At least, W may compensate for the univalency of Z chromosome in females, e.g. to ensure proper meiotic pairing and segregation.

The extremely complex and variable, repetitive structure of lepidopteran W chromosomes not only has complicated comparative cytogenetic studies, but has also, until recently, made their assembly either difficult or impossible. With the advent of highly accurate long-read sequencing, the rate of full-chromosome assembly releases has rapidly increased. Currently, the genomes of over 1,000 lepidopteran species have been sequenced and published in the NCBI database. However, very few of these were obtained from heterogametic females and even fewer provided a partially resolved structure of W. At the present time, the main contributor of high-quality full-chromosome assemblies of Lepidoptera is the Darwin Tree of Life Project (primarily, Wellcome Sanger Institute), whose methodological approach builds on a combination of PacBio HiFi and Arima Hi-C technologies. However, geographic origins of the species included in their study are limited to Britain and Ireland.

The Siberian moth, *Dendrolimus sibiricus*, is one of the 15 species of the Palearctic genus *Dendrolimus* (Jeong et al. 2018), some of which are considered serious coniferous pests causing damage to tree stands over large areas (Rozhkov 1963; Xiao 1992). Indeed, *D. sibiricus* is one of the most important pests of boreal forests in North Asia, causing large-scale outbreaks (Baranchikov and Kondakov 1997). As of 2025, five species of the genus *Dendrolimus* have had their genomes sequenced: *D. kikuchii* (Zhou et al. 2021), *D. punctatus* (Zhang et al. 2020), *D. pini* (Botham and Dainton 2024), the yet unpublished *D. houi* (GCA_039645545.1), and *D. tabulaeformis* (GCA_042997425.1). The structure of W chromosomes has not been obtained for any of these, although female samples were sequenced in the case of *D. punctatus and D. kikuchii*. The genome assembly of the latter contains one fewer chromosome than all other *Dendrolimus* species examined (where $n = 30$), a difference the authors attributed to a chromosome fusion event (Zhou et al. 2021).

Within the genus *Dendrolimus,* speciation of *D. sibiricus* is considered to have occurred relatively recently (Kononov et al. 2016), so even its status as a distinct species or a subspecies of the partially sympatric *D. superans* is still a matter of debate (Mikkola and Ståhls 2008; Zolotuhin 2015; Jeong et al. 2018). In this regard, its genome may still bare clear traces of the recent transformation of its sex chromosome system.

In this study, we provide a chromosome-scale assembly of the *D. sibiricus* genome, including the structure of both Z and W sex chromosomes. In investigating the structural and functional features of the Siberian moth genome, we developed scenarios for the evolutionary turnover of its W chromosome, using both phylogeny- and synteny-based approaches.

Results

Chromosome-scale assembly of the *Dendrolimus sibiricus* genome

The hifiasm *de novo* assembly of PacBio HiFi sequencing reads from a single *D. sibiricus* female, after minor modifications (see M&M section for details), yielded a highly contiguous genome assembly spanning 612 megabases and containing 30 chromosome-length scaffolds, including the Z sex chromosome (Fig. 1). The W (female) sex chromosome, however, was reconstituted separately from several fragments (see below). The assigned BUSCO statistics indicate that the assembly captures nearly 99% of universal lepidopteran single-copy genes, confirming its high level of completeness (Fig. 1, Table S1). Subsequent comparative analyses of synteny and representation of telomeric regions in assemblies of *D. sibiricus* and related species (see below) provided additional confirmation of the high quality of this assembly.

Of the 30 chromosomes, only two gaps were introduced by scaffolding of chromosomes 25 and 29, which were initially split into 2 fragments due to the presence of highly homogeneous extra-long tandem repeats; these turned out to be clusters of rDNA and histone genes, respectively (Fig. 1). The remaining fragmented sequences most likely belonged to the W sex chromosome. As it turned out later, several large segmental duplications displaying a high degree of identity, devoid of protein-coding genes and saturated with transposons, complicated the assembly of this chromosome.

Whenever possible, we eliminated observed errors during manual curation of the assembly. Although we sequenced an additional genomic library of high-precision Illumina short reads for the same individual, we preferred to avoid automatic assembly polishing, since it could introduce more errors than

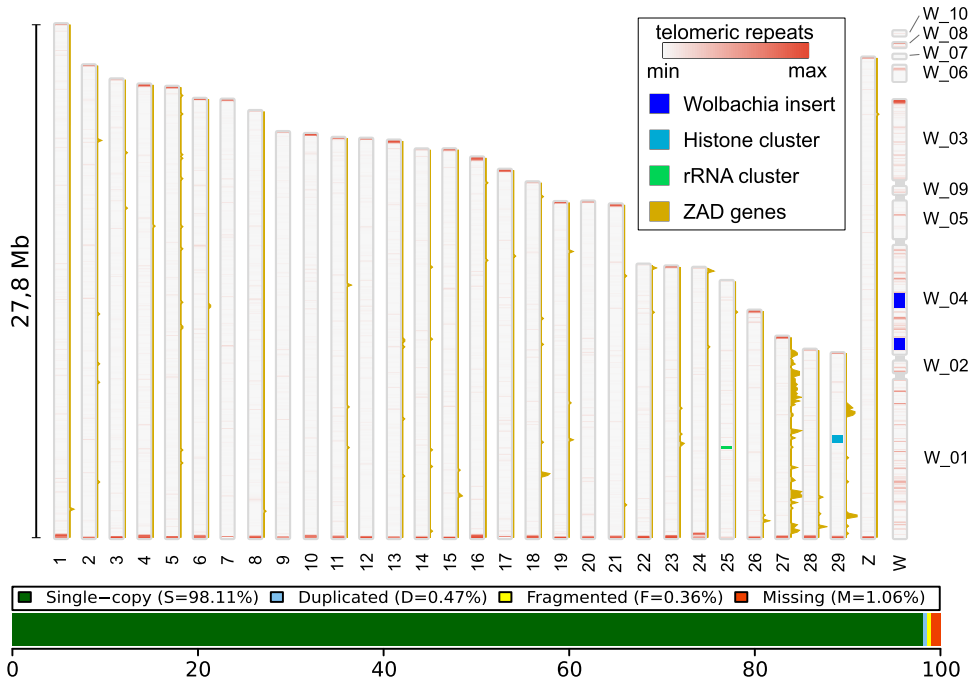

Figure 1 Main features of the *D. sibiricus* genome assembly. Upper panel shows the size distribution of assembled scaffolds, each representing a chromosome (except for the fragmented and partially arranged W chromosome). Red-color gradients provide a visual estimate of the abundance of telomere-associated repeats in 100Kb genomic windows. The loci containing inserts of *Wolbachia* genomic fragments (chrW_04), cluster of rRNA (chr25), and cluster of histone genes (chr29) are marked by separate colors. Lateral histograms (gold color) show the density of ZAD-containing genes, which are highly enriched in chr27. Lower panel shows statistics of BUSCO analysis (Lepidoptera gene set) performed on the full assembly.

improvements in the case of a highly heterozygous diploid organism. Instead, using both HiFi and Illumina data, we identified, verified, and corrected only reliable homozygous calls, as well as those heterozygous variants that apparently disrupted the ORFs compared to the aligned transcripts. In total, 441 such changes were made to the primary assembly, including correction of 17 true deleterious heterozygous SNPs.

We also sequenced polyA-enriched and rRNA-depleted RNA-seq libraries to assemble a representative transcriptome and build a comprehensive gene annotation of the *D. sibiricus* genome, which is described in detail in the Supplementary Notes. Briefly, we annotated 14,461 protein-coding genes characterized by a BUSCO completeness of 97.3%, as well as 6,879 non-coding RNA genes (Table S7).

Karyotype of *Dendrolimus sibiricus*

To assess the consistency of the assembly results with the actual number of chromosomes in the genome of *D. sibiricus*, we performed the karyotyping of both female and male chromosomes.

In 58 meiotic metaphase I (MI) cells studied from 4 males, 30 chromosomal bivalents were observed (Fig. 2a–c). Thus, based on the fact that lepidopteran males typically have a ZZ chromosomal sex determination mechanism (Robinson 1971), we conclude that males have 29 autosomes and 1 (sex) Z-chromosome in the haploid set. In the set, the bivalents form a gradient size range. We observed no deviation from the chromosome number value of $n = 30$.

Sex chromatin was found in the interphase nuclei of somatic cells from the gonads of four females studied. Sex chromatin appeared as a dark-colored rounded body (Fig. 2d). No sex chromatin was found in the interphase nuclei of somatic cells from the gonads of four males studied (Fig. 2e).

Thus, the results of the karyotyping are clearly consistent with the number of chromosome-scale scaffolds in the *D. sibiricus* genome assembly. In addition, the data support the presence of a W chromosome in the karyotype of females, and therefore the presence of WZ/ZZ sex determination system in *D. sibiricus*.

Deciphering and reassembling the sex chromosome sequences

We intentionally sequenced the genome of a heterogametic female of *D. sibiricus* with the aim of tracing W-chromosome-associated sequences, if available. Since the Z chromosome is highly conserved in the Lepidoptera clade (Fraïsse et al. 2017), we easily identified it among 30 chromosome-scale sequences and confirmed its haploid state based on half coverage by HiFi reads. Still, the assembly contained a very significant (~25Mb) pool of fragmented sequences represented in one haplovariant, i.e. potentially related to the W chromosome. We sequenced two additional genomic short-read libraries of female and male *D. sibiricus*, and classified all contigs of the diploid assembly into autosomes, the Z chromosome, and the W chromosome, based on the theoretically expected ratios of coverage by these libraries according to the WZ/ZZ sex determination scheme (Fig. 3).

The W-classified set contained 293 initial unitigs arranged into 33 contigs in the draft genome assembly. The originally assembled Z sequence was corrected as it contained a chimeric

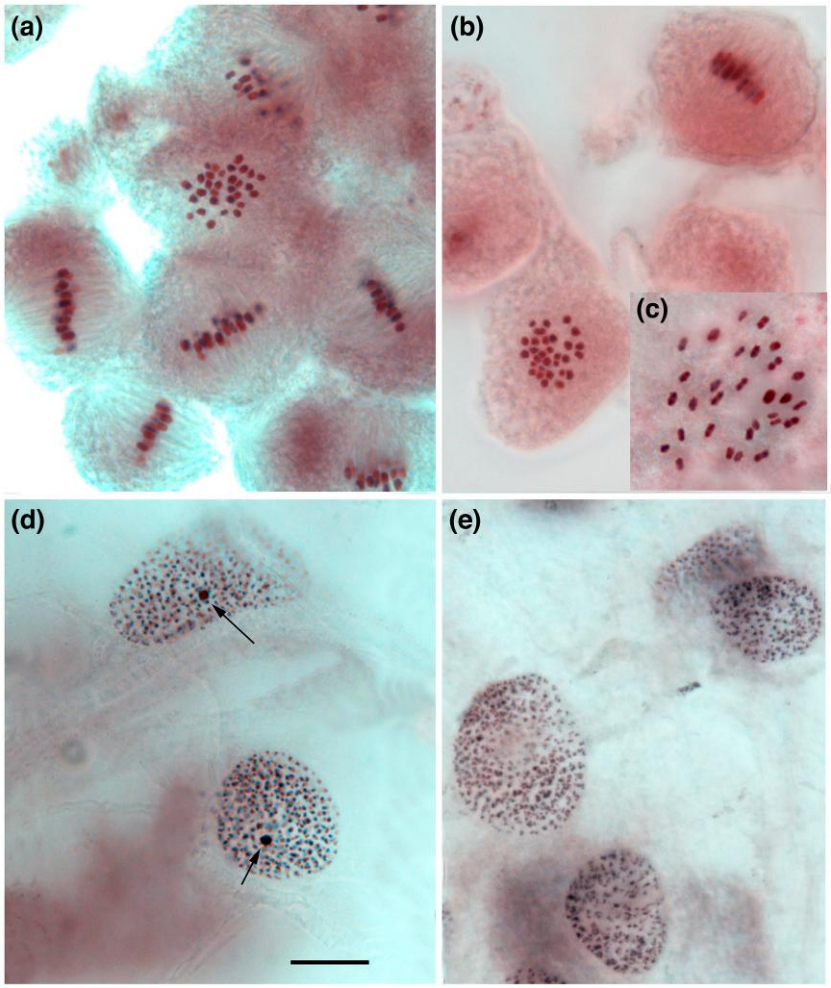

Figure 2 Karyotypes of *Dendrolimus sibiricus*. a)—weakly squashed spermatocyst showing a nearly intact male MI plate (in the center, $n = 30$, polar view) and several MI plates from an equatorial view; b)—polar (left, bottom) and equatorial (right, top) views of male MI nearly intact plates, $n = 30$, c)—MI male plate, squashed preparation, $n = 30$; d)—nuclei of two female interphase cells showing sex chromatin bodies (=W chromosome) (arrows); e)—nuclei of male interphase cells without sex chromatin bodies. The scale is 10 μ for all photographs.

junction with W-classified sequence resulting from incorrect haplotype resolution of unpaired chromosomes by the hifiasm. To avoid such errors, we separately assembled the W chromosome with tuned parameters, taking only HiFi reads mapped to W-classified unitigs.

The resulting assembly graph revealed a highly complex organization of the W chromosome, including a number of highly homologous segmental duplications (two to three tandem copies) (Fig. 3c). We deliberately accounted only 1 copy of such duplications in the final set of 10 contigs with a total length of 22.46 Mbp. However, the actual length of W is about 27.4 Mbp, based on contig coverage by HiFi reads. Unresolved assembly fragmentation of the W chromosome is caused by its exceptional saturation with genomic repeats, together with the fact that some large repeat-rich segments are multiplicated. Nevertheless, we were able to unambiguously arrange “one-copy” contigs into a representative scaffold of the W chromosome using the whole set of HiFi reads (Fig. 3c).

Conserved telomeric structure as a marker of assembly completeness

For various insects, including *Bombyx mori*, it has been shown that telomere maintenance is provided not only by the pentanucleotide repeat $(TTAGG)_n$ synthesized by telomerase, but also by telomere-specific non-LTR retrotransposons from the SART/TRAS family that are specifically integrated into this repeat (Nichuguti and Fujiwara 2020; Lukhtanov and Pazhenkova 2023). We annotated *de novo* genomic repeats for the assemblies of four *Dendrolimus* species and analyzed the overrepresentation of repeat families at the ends of the chromosome-size scaffolds. We found a cluster of repeats highly enriched in the telomeric regions of the chromosomes in most of the studied genomes (Figure S1). Among them, many highly variable retrotransposons were strongly correlated and co-clustered with the telomeric repeat $(TTAGG)_n$, along with several other simple repeats (Fig. 4, Figure S1).

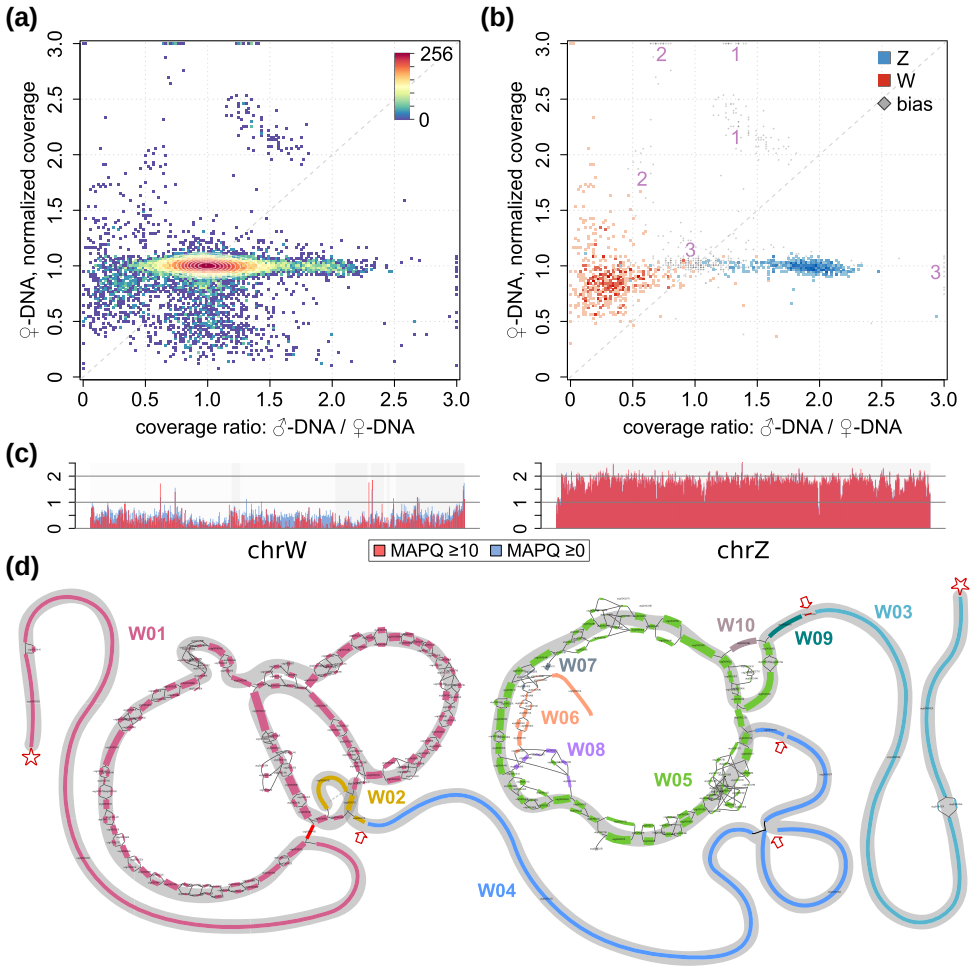

Figure 3 Selection and reassembling of W- and Z-derived genomic regions in the *Dendrolimus sibiricus* diploid assembly. The left heatmap a) shows the distribution of 50 Kb genomic windows based on their coverage by female and male WGS libraries. The color scale represents the number of windows in a bin. Based on the WZ/ZZ model, Z-derived and W-derived windows are expected to appear near (2,1) and (0,1) XY coordinates, respectively. The main peak of density at (1,1) corresponds to the autosome-derived heterozygous windows. In the right heatmap b), identified Z- and W-derived windows are highlighted by blue and red colors. Gray colored diamonds mark concomitant windows with biased copy number that are numbered by their source as follows: (1) rDNA cluster, (2) *Wolbachia* circular genome, (3) cluster of histone genes. Panel c) shows the distribution of the male-to-female coverage ratio along the final contigs of sex chromosomes. Panel d) shows the unitig graph produced by hifiasm during an isolated assembly of the W chromosome. Unitigs are shown as colored blocks and their overlaps as black curves. Block thickness corresponds to depth of coverage. Cyclic paths indicate tandem duplications. The color of the blocks denotes their correspondence to the resulting 10 contigs (single-copy representation). The thick gray outline of the graph indicates the manually curated overlap-based arrangement of the contigs within the W scaffold. Arrows mark manually defined arrangements. Stars mark the inferred telomeric regions.

However, only a few transposons have been found to encode both R1-clade endonuclease ("R1-I-EN" domain) and non-LTR reverse transcriptase ("RT_nLTR_like" domain) (Table S2). The remaining repeats appear to be non-autonomous non-LTR retrotransposons that hijack the retrotransposition machinery of autonomous ones (Fig. 4). Most of them also lack significant similarity to known *B. mori* SART and TRAS elements.

According to the distribution of telomere-associated repeats, 30 major scaffolds in the *D. sibiricus* genome assembly represent the "telomere-to-telomere" sequences of 29 autosomes and the Z chromosome (Fig. 1, Figure S2). It is worth noting that the overwhelming majority of identified repeat motifs clearly delineate telomeric regions in genomes of all studied species except *D. punctatus*. In the latter, these repeats and corresponding telomeric regions of chromosomes were underrepresented (Fig. 4). Thus, enrichment analysis of telomere-associated repeats, as presented here, may serve as an additional measure of assembly completeness of Lepidopteran genomes, and may also reveal the conservation of the mechanisms of telomere length maintenance in Bombycidae, although different families of non-LTR transposons may be involved.

Comparative genomics of four *Dendrolimus* species

Having completed the assembly of the *D. sibiricus* genome, we investigated the conservation of the obtained linkage groups and phylogenetic relationships within the genus *Dendrolimus* at the whole-genome level. At the time of writing, chromosome-scale

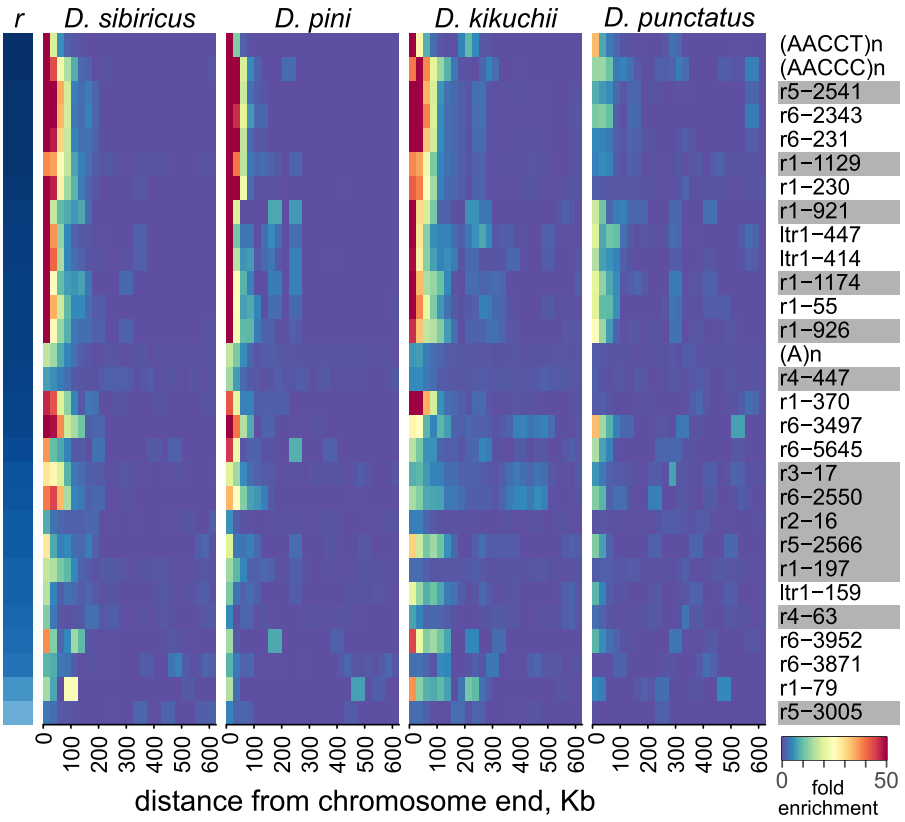

Figure 4 Enrichment of telomere-associated repeats at terminal regions of chromosomes, estimated for the genome assemblies of four *Dendrolimus* species. Left-most bar shows degree of positive correlation between repeat distribution and telomeric $(TTAGG)_n/(AACCT)_n$ repeats. Column on the right: repeat designations; non-LTR transposons are highlighted in gray.

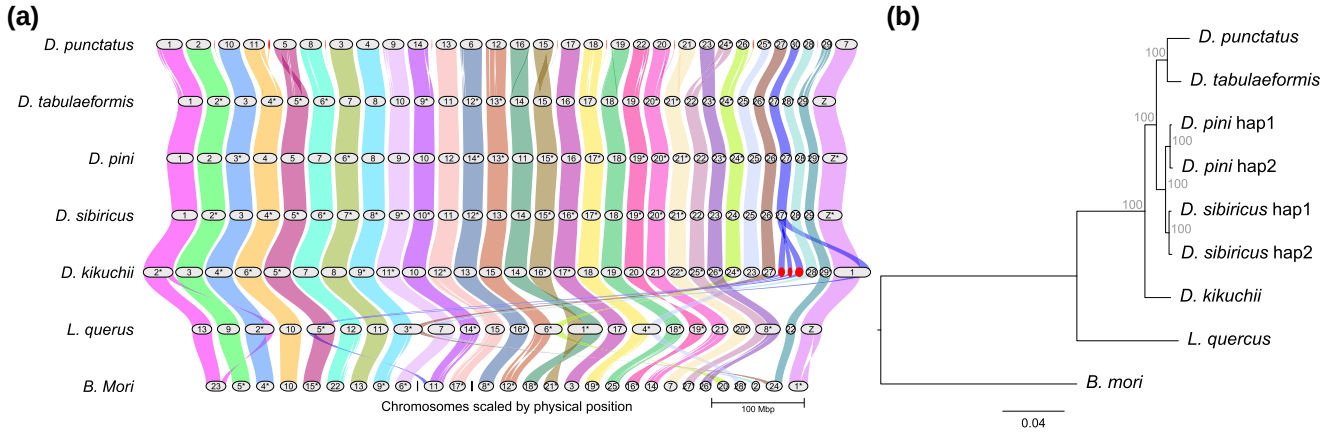

Figure 5 Comparative syntenic and phylogenetic analysis of seven lepidopteran genome assemblies. a) The syntenic riparian plot. Chromosome numbers are those provided by the corresponding NCBI projects. Unplaced scaffolds are marked by red fill. b) The phylogenetic trees based on single-copy orthologous proteins. The dendrogram also includes both partially phased haplotypes from the diploid assemblies of *D. sibiricus* and *D. pini* genomes. The trees are rooted using *B. mori* as outgroup. Numbers show node bootstrap support values. The trees were generated on partitioned data using IQ-TREE maximum likelihood inference with 10,000 bootstraps and WAG substitution model.

assemblies of the female *D. kikuchii*, *D. punctatus*, *D. tabulaeformis*, and male *D. pini* genomes were publicly available, but unfortunately, none of them were annotated. In an effort to circumvent this limitation, we simply transferred our annotation of the *D. sibiricus* genome onto the other three, supplementing it with transcriptome assemblies based on the available RNA-seq data for each species (Table S7).

In addition to the above species, our analysis of genome synteny and collinearity of annotated genes included the moth *Lasiocampa quercus* from the same Lasiocampinae subfamily, and *Bombyx mori* (Bombycidae) as outgroup species. The analysis revealed a high level of identity in the chromosomal organization of *D. sibiricus*, *D. pini*, *D. tabulaeformis*, and *D. kikuchii* (Fig. 5a). For these species, the only substantial change in gene

collinearity was the previously described fusion of sex chromosome Z with the homolog of *D. sibiricus* chromosome 27 in *D. kikuchii* (saturated with *ZAD-ZNF* genes; Supplementary Notes). However, we also observed a second homolog of chromosome 27 in the assembly of the female *D. kikuchii* genome, fragmented into three scaffolds (Fig. 5a). These scaffolds may belong to a neo-W chromosome formed in pair with a fused neo-Z.

An exceptionally large number of intrachromosomal rearrangements were detected in the *D. punctatus* genome relative to the other species. Of the latter, *D. tabulaeformis* is considered a sympatric sister species or even a subspecies of *D. punctatus* group, which is known to have high genetic heterogeneity (Kononov et al. 2016). Thus, it cannot be excluded that some of the observed rearrangements may be the result of imprecise Hi-C scaffolding, especially since some of them are associated with gaps and unplaced contigs (Fig. 5a). Interestingly, the more compact genome of the outgroup *L. quercus* (566 Mb) contains only 22 autosomes, resulting from 7 chromosome fusion events.

Finally, the genomes of *D. sibiricus* and *D. pini* showed an exceptionally high level of sequence similarity, more or less comparable to the identity of the two assembled haplotypes of *D. sibiricus*. That is, at the macro level of chromosomal organization, we did not identify obvious rearrangement of the genomes that could cause the genetic isolation of these two species.

To better understand the level of evolutionary distance between the studied species, we performed a phylogenetic analysis based on 2,805 single-copy orthologous genes, which additionally included information on haplovariants of the diploid genomes of *D. sibiricus* and *D. pini* (Fig. 5b). The resulting tree topology confirmed the evolutionary divergence of *D. sibiricus* and *D. pini*, albeit a relatively recent one. At the same time, the analysis revealed significantly greater distance between the sister species *D. punctatus* and *D. tabulaeformis*, which have been shown to be capable of hybridization (Zhao et al. 1992).

Co-assembled genomic sequences of *Wolbachia* endosymbionts

In addition to the nuclear and mitochondrial (CM102436.1; 15,411 bp) genomes of *D. sibiricus*, the assembly work we conducted also generated the complete sequence of a *Wolbachia* circular genome. *Wolbachia* is the most widespread intracellular bacterial endosymbiont found in arthropods, including butterflies and moths, and it frequently affects the reproduction of its host (Werren et al. 2008). The endosymbiont is vertically transmitted from mother to offspring, but rare horizontal transmission also takes place between species. A recent study pointed to the high prevalence of *Wolbachia* in populations of *D. sibiricus* (= *D. superans* ssp. *sibiricus* (Bykov et al. 2020)), which belong to two related haplotypes, ST-41 and ST-353. The assembled genome belongs to supergroup B of *Wolbachia pipientis* (MLST: ST-353) and is most closely related to the known isolate from *Apotomis turbidana* (OX366384; 99.94% identity and 94% coverage).

Moreover, we discovered two other large fragments of the *Wolbachia pipientis* genome as part of the nuclear genome of *D. sibiricus*, namely within the W chromosome (W_04, Fig. 1). The high degree of identity between the two integrated regions suggests that one of them arose as a consequence of segmental duplication of a first fragment initially integrated into the host nuclear genome. Since the inserted fragments display a high degree of identity to the co-assembled circular endosymbiont genome, including conservation of coding ORFs, we speculate that these insertions occurred relatively recently. As expected from their prokaryotic origin, these insertions are most likely transcriptionally inactive and nonfunctional; in the *D. sibiricus* transcriptomes, there were no reliable mRNA products that mapped selectively to W-linked inserts while not mapping to the circular genome of *Wolbachia* (Figure S3).

Given that evolution of the sex determination system and speciation are two interconnected processes, and *Wolbachia* may act as a cytoplasmic feminization factor, we explored the possibility that *Wolbachia* may have played a role in the speciation of *D. pini* versus *D. sibiricus*, as these two species have overlapping ranges. We searched for circular or integrated *Wolbachia* sequences in genome assemblies of other *Dendrolimus* species but found none. We then reassembled the genomes of *D. pini* and *D. punctatus* from available raw data. Strikingly, we identified two different circular *Wolbachia* genomes in a genome assembly obtained from a single *D. pini* male: one from supergroup B (MLST: ST-374) and one from supergroup A (MLST: ST-92). Interestingly, the ST-41 *Wolbachia* haplotype was detected earlier in two tested *D. pini* samples from the Kaliningrad region (Bykov et al. 2020). Given that *D. sibiricus* and *D. pini* are very closely related and have overlapping habitats (Kononov et al. 2016), we attempted to predict potential cytoplasmic incompatibility between these species arising from differences in *Wolbachia* strains. By comparing three *Wolbachia* genomes, we identified both common and exclusive pairs of *cifA* and *cifB* toxin–antitoxin genes (Table S3) that are thought to mediate cytoplasmic incompatibility (Martinez et al. 2021). Given the overwhelming prevalence of *Wolbachia* in *D. sibiricus* populations (Bykov et al. 2020), the observed differences suggest that *Wolbachia* may be involved in the reproductive isolation or even speciation of *D. pini* and *D. sibiricus*, a subject warranting additional research.

The landscape of repeats differs dramatically between the W and the rest of chromosomes

In addition to the presence of *Wolbachia* genome fragment insertions, a distinguishing feature of the *D. sibiricus* W chromosome is a notable expansion of LTR transposons. The proportion of these repeats was observed to be 36.5%, which is more than three times that found in the whole genome (Figure S4). The total proportion of repeats in the W chromosome was calculated to be ~80%.

Analysis of the interspersed repeat landscape, using Kimura 2D distances as a proxy of transposon age, revealed major differences between the W chromosome and the rest of the nuclear genome (Fig. 6a). The W chromosome was found to be composed primarily of much younger repeats compared to all other chromosomes, with the expansion of the LTR/Gypsy and LTR/Pao transposons being the most recent and extensive. An analysis of the enrichment of individual repeat families revealed a series of repeats that are strongly overrepresented or even specific to

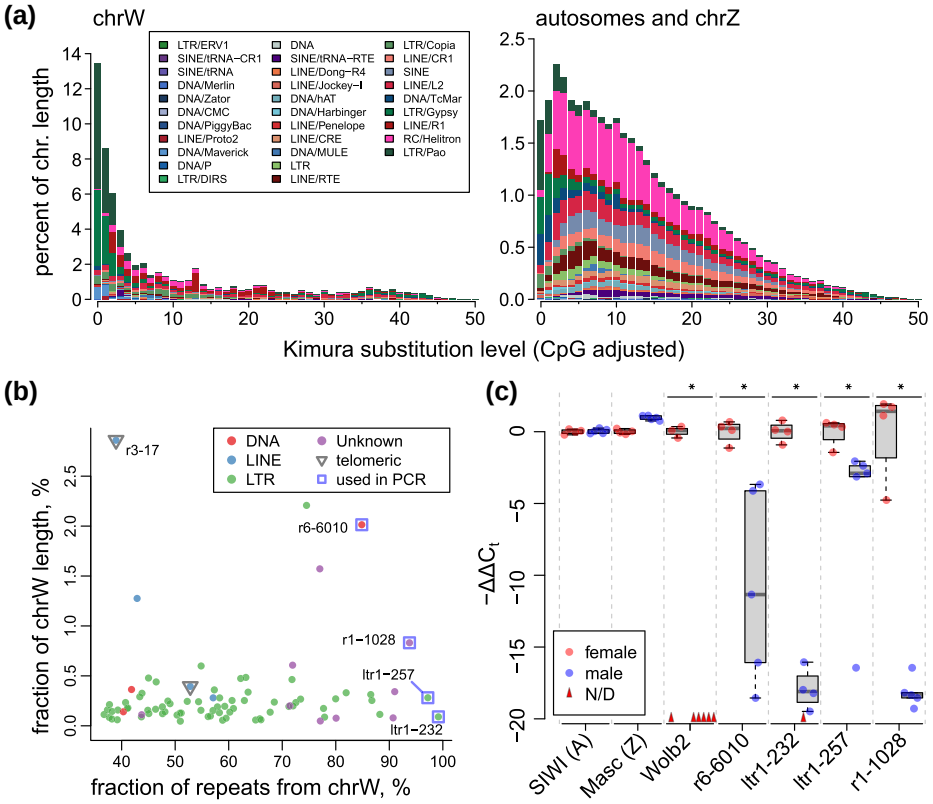

Figure 6 Specific repetitive content of the W-chromosome. a) The landscape of transposable elements in chrW (left) and the rest of the genome (right) of *D. sibiricus* based on Kimura distance-based copy divergence analysis. b) The abundance (Y axis) and chromosome specificity (X axis) of genomic repeats that are enriched on chrW. Only repeats meeting the following criteria are shown: ≥10 repeat units spanning ≥10 Kbp and ≥10-fold overrepresented. c) Assessment of W-specific repeats and *Wolbachia* genome inserts using comparative qPCR analysis of male and female genomic DNA samples. The abundance of target templates is provided as −ΔΔCt values, calculated using autosome-derived SIWI gene (two-copy in males and females) and chrZ-derived *Masc* gene (two-copy in males and single-copy in females) as the reference. The reactions that failed to detect target template (mainly *Wolbachia* inserts in males) are marked by red triangles at the bottom of the boxplot. Asterisks indicate significant sex difference (Mann–Whitney adjusted *P*-value < 0.05).

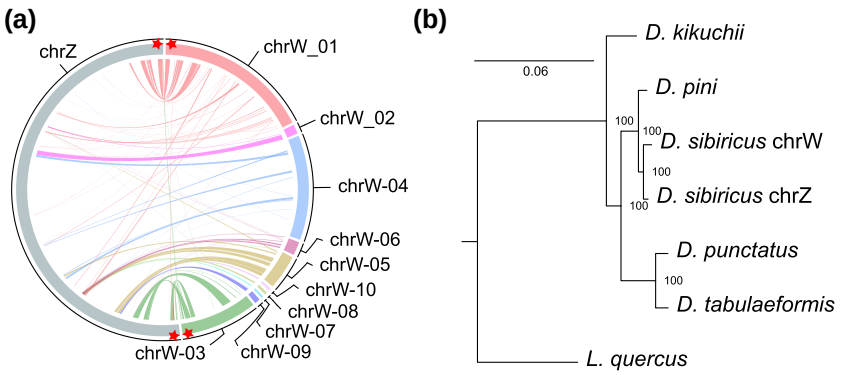

Figure 7 Non-repetitive content of the W chromosome. a) Syntenic map between the Z and fragmented W chromosomes. Telomeric repeats are marked by red stars. Please note the high degree of synteny in the adjacent subtelomeric regions. b) phylogenetic tree built on concatenated sequences of 14 genes from chrW that have valid full-length ORFs, have paralogs on chrZ of *D. sibiricus* and orthologs on chrZ of selected closely related species. Numbers show node bootstrap support values. The tree was generated on non-partitioned data using IQ-TREE maximum likelihood inference with 10,000 bootstraps and JTT + G4 model, and was rooted using *B. mori* as outgroup.

the W chromosome (Fig. 6b, Table S4). It is worth noting that among them was the telomere-specific non-LTR transposon "r3-17" (Fig. 4) scattered throughout the body of the W chromosome and accounting for almost 3% of its sequence.

We tested several of the most overrepresented repeats (highlighted on Fig. 6b), as well as one of the insertions of a fragment of the *Wolbachia* genome as W-chromosome-specific markers using qPCR on DNA samples of *D. sibiricus* females and males

(Fig. 6c). Autosome-specific (*SIWI*) and Z-chromosome-specific (*Masc*) genes were used as internal reference. We expected high within-group variation for abundance estimates of target repeats because the primers were designed to target multiple copies of the variably diverged repeat in the genome (multiply concurring templates). Nevertheless, the PCR results confirmed the specificity of target repeat enrichment in the genomic DNA of females, and consequently in the W chromosome. Thus, features of the W-chromosome organization identified here are confirmed by an analysis that used an independent set of individuals.

Non-repetitive content of the W suggests its origin from the Z

The abundance of young genomic repeats and well-preserved insertions of *Wolbachia* genomic fragments in the W chromosome points to its relatively recent formation, an observation that supports the theory of multiple non-canonical emergences of W chromosomes in the evolution of Lepidoptera, including its formation from B chromosomes (Fraïsse et al. 2017; Lewis et al. 2021).

However, an analysis of synteny among the chromosomes of *D. sibiricus* revealed significant regions of collinearity between the Z and W chromosomes (Fig. 7a). In particular, both distal regions of the W chromosome turned out to be partially preserved distal regions of the Z chromosome, including both telomeric ends.

We then carefully reexamined predicted W-linked coding sequences and defined a final set of 52 genes that lacked evidence of transposable elements (Table S5). Almost all of them are localized precisely in the aforementioned regions of synteny, indicating that they have a close paralog on the Z. A detailed comparison of the Z and W homologs showed that most W chromosome genes are likely pseudogenes with a disrupted or truncated ORF compared to their Z chromosome counterparts, indicating a process of gene decay on the W (Table S5). At the same time, mapped RNA-seq data supported expression of these genes, albeit with a significant proportion of multi-mapped reads. Considering the highly condensed state of the W in interphase nuclei, we suggest that this expression signal is a contamination from highly homologous Z-linked genes. We tested this hypothesis on available RNA-seq data from gonads and gut of male and female *D. sibiricus* larvae. Comparable expression levels of W-linked genes were observed in both female ZW and male ZZ samples, confirming the proposed mapping artifact (Figure S5). It is also worth noting that the corresponding Z-linked paralogs did not show sex-specific differential expression. Thus, the presence of W-linked genes does not affect the dosage compensation efficiency of their Z-linked paralogs, at least at the life stage studied.

For phylogenetic analysis, W-linked genes with a paralog on the Z chromosome and Z-linked orthologs in each of *D. pini*, *D. punctatus*, *D. kikuchii*, *D. tabulaeformis*, *L. quercus* as well as orthologs in *B. mori* were selected. The final phylogeny was inferred from 14 orthogroups for which the precomputed topology of Z-linked orthologs was consistent with the genome-wide phylogeny (Fig. 7b). The results confirmed that the non-repetitive part of the W chromosome originated from a copy of the Z of *D. sibiricus*, at least after the divergence of *D. sibiricus* and *D. pini*. Further gene decay or loss and extensive accumulation of mobile elements probably occurred in conjunction with the restriction of recombination in the Z–W pair.

Similar organization but lack of common ancestry among W chromosomes of related species

To assess the evolutionary novelty of the *D. sibiricus* W chromosome, we examined its conservation with known female genomes of closely related species (genus *Dendrolimus*). In the case of *D. kikuchii*, the female genome assembly lacked a defined W chromosome, as well as raw data to identify its candidates. In contrast, the *D. tabulaeformis* assembly obtained from a larva of undetermined sex contained megabase-size unplaced scaffolds, and the raw Hi-C and HiFi data were available for their analysis. Using the reassembled genome (to obtain a diploid unitig graph), we confirmed the female-specific hemizygous Z state based on the sequencing coverage of the primary assembly (Fig. 8b). The potential W unitigs were also hemizygous but displayed no synteny with related genomes and were separated from telomere-to-telomere sequences of autosomes and Z in either the unitig graph or the Hi-C contact map. We extracted and reassembled the corresponding long reads for these unitigs, yielding 11 final contigs of the *D. tabulaeformis* W chromosome, with a total length of 16.4 Mbp (Fig. 8a).

A striking feature of the *D. tabulaeformis* W chromosome is its high content in horizontally transferred fragments from the insect's own mtDNA (Fig. 8a, orange), echoing the transfer of the *Wolbachia* genome seen in *D. sibiricus*. Together with genomic repeats (Fig. 8a, blue), they occupy almost the entire length of the W. As seen in *D. sibiricus*, the recently amplified Gypsy and Pao LTRs are much more abundant than diverged ones, in contrast with the rest of the genome (Fig. 8e). Similarly, the only chromosome that shows extensive one-to-one homology of gene-containing regions is the *D. tabulaeformis* Z (Fig. 8a, cyan). However, the W sequences derived from Z are more degraded compared to those of *D. sibiricus*. At first glance, the W chromosomes of *D. sibiricus* and *D. tabulaeformis* also display significant local similarity, covering 68.4% and 76.6% of their length, respectively (Fig. 8a, green; referred to as W–W matches). However, as can be seen from the identity in the distribution of repeats and W–W matches, this similarity is predominantly explained by multiple short alignments between individual repeat units that do not form longer syntenic regions, not even collinear chains of repeats.

If the W chromosomes of two species are derived from a single ancestral W chromosome, orthologous regions are expected between them that display little or no identity to the rest of the chromosomes in both genomes. To examine this hypothesis, we aligned each W-W match with the set of autosomes and Z chromosomes of both species and compared the similarity score between the W–W and W–ZA matches. At a 5% difference threshold, only a few short, sparsely distributed W-W alignments had higher identity over their W–ZA counterparts, yet they were almost completely repetitive and still had substantial identity with Zs and autosomes (Fig. 8a, black marks; Table S6). This means that the origin of these two W chromosomes from a common ancestor is unlikely. It is worth noting that the observed high

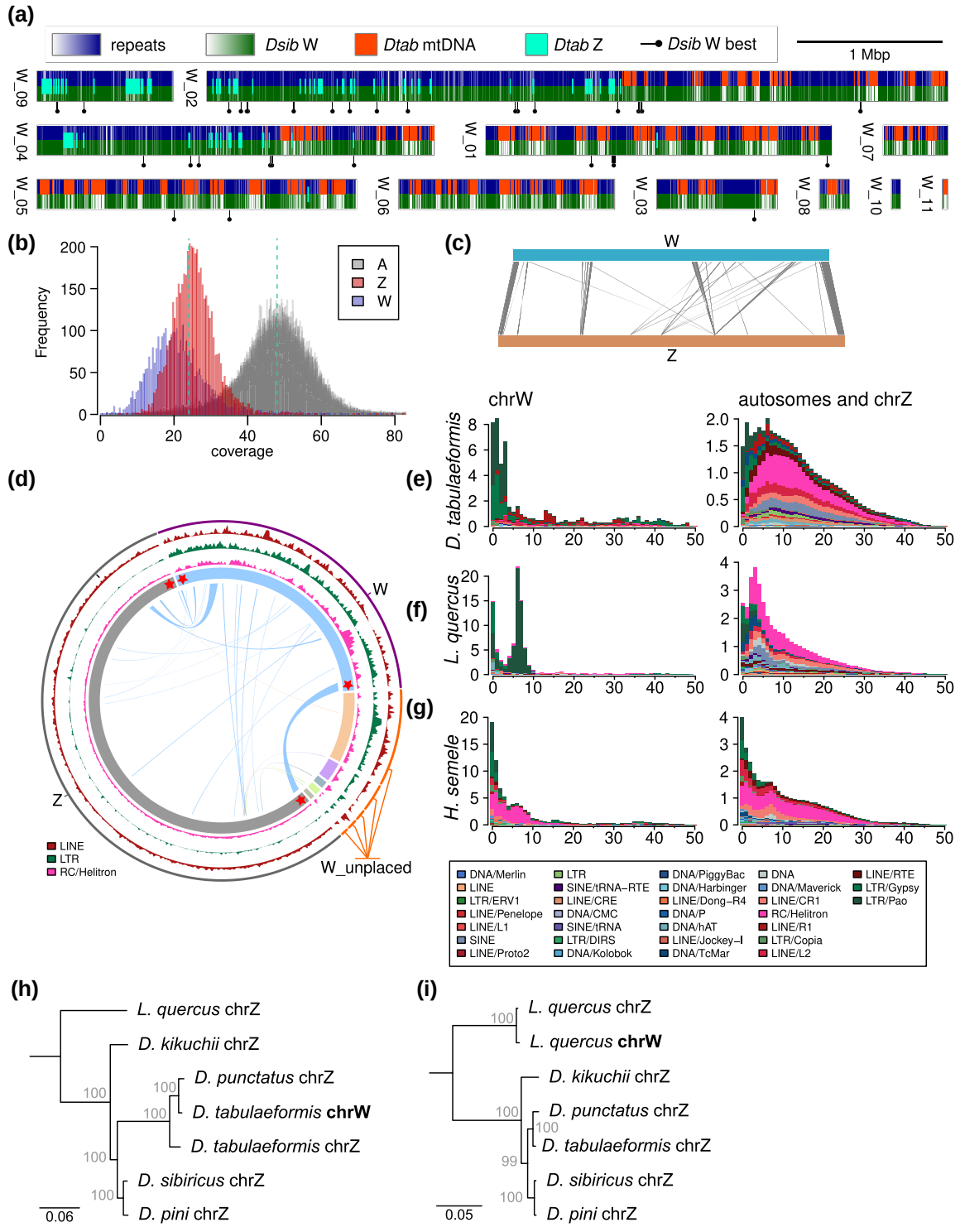

Figure 8 Structural features of the W chromosomes of *D. tabulaeformis*, *Lasiocampa quercus*, and *Hipparchia semele*. a) The size and content of 11 assembled contigs of the *D. tabulaeformis* W chromosome: the density of genomic repeats (repeats) and LASTZ matches with *D. sibiricus* W chromosome (Dsib W), their reciprocal-best set (Dsib W best), minimap2 alignments of its own mitochondrial DNA (Dtab mtDNA) and the Z chromosome (Dtab Z). b) Coverage of genomic HiFi reads in sliding windows of the haploid *D. tabulaeformis* assembly. Each chromosome was processed separately as indicated by the three-color scheme (autosomes, Z, W). c) Syntenic map between the Z and the W of *L. quercus*. d) Syntenic map between the Z and the W (including unplaced W scaffolds) of *H. semele*. Telomeric repeats are marked by red stars. Radial histograms show the density of helitrons, LTRs, and LINEs. e–g) The landscape of transposable elements of *D. tabulaeformis*, *L. quercus*, and *H. semele*, computed for the W chromosome (left) and the rest of the genome (right) using Kimura distance-based copy divergence analysis. h, i) Phylogenetic tree built on W-linked genes of *D. tabulaeformis* (h) and *L. quercus* (i) and their paralogs and orthologs on chrZ of closely related species. Notations and parameters are similar to those in Fig. 7b.

local identity of their repeats may be sufficient to detect significant signal between the two Ws in interspecies FISH experiments.

We also identified seven W-linked protein-coding genes with Z-linked paralogs in *D. tabulaeformis* and performed a phylogenetic analysis (Fig. 8h, 5 genes) similar to that performed for *D. sibiricus* W-linked genes. Unexpectedly, the W chromosome of *D. tabulaeformis* was found to be evolutionarily closest to the Z chromosome of *D. punctatus*. As mentioned above, *D. tabulaeformis* is likely a subspecies of the genetically heterogeneous species *D. punctatus*, and this tree topology likely represents hybridization between these taxa. Interestingly, the divergence of the complete genomes of these species was substantially more pronounced (Fig. 5b).

Thus, the comparative analysis of the W chromosomes of *D. sibiricus* and *D. tabulaeformis* points to their convergent evolution from a copy of their own Z chromosomes rather than divergence from a common ancestral W.

The W chromosome of a more distantly related Lasiocampinae species, *Lasiocampa quercus*, also shares features of the W–Z synteny shown above, involving both subtelomeric regions of the chromosomes (Fig. 8c). In contrast to autosomes and Z, the W chromosome was found to be exceptionally depleted in old, diverged repeats, while the recent expansion of LTRs occurred in two waves, the first of which occurred earlier than in the two *Dendrolimus* species (Fig. 8f). As seen for *D. tabulaeformis*, we found only two short repetitive W–W matches satisfying the above conditions for common ancestry between *L. quercus* and *D. sibiricus* (Table S6). Finally, phylogenetic analysis of 8 of 13 detected pairs of Z- and W-linked genes (Fig. 8i) confirmed the origin of the W chromosome from Z in *L. quercus* and its corresponding evolutionary independence from W of other species.

Finally, we examined a set of other publicly available lepidopteran genomes containing defined W chromosome sequences to assess similarities in their organization to the *D. sibiricus* W chromosome. The W of *Hipparchia semele* (Nymphalidae) was found to best reproduce the clear synteny with the Z and age disproportion of transposons observed in *D. sibiricus*. It is largely composed of genomic repeats but retains clear syntenic blocks with the chromosome Z, again close to the telomeric ends of both chromosomes (Fig. 8d). The distribution of W repeats by divergence was significantly biased toward recently expanded ones, with LTR/Pao, LTR/Gypsy, and RC/Helitrons being the main contributors (Fig. 8g).

To summarize, the W chromosomes of the species examined here (Lasiocampidae and Nymphalidae) most likely arose independently from a copy of their own Z chromosome and evolved through degeneration of its sequences including gene loss and expansion of active genomic repeats. We suggest that this may indicate a common pathway of W chromosome turnover in many lepidopteran species, which in the course of long-term evolution may result in a complete loss of synteny with the progenitor Z chromosome and thereby mimick a B chromosome-like organization.

Discussion

Genome assembly

In this paper, we provide convincing evidence that the female W chromosome of the Siberian moth was derived from its Z counterpart and that a similar Z-derived W inception apparently occurred independently in closely related species. As a first step toward exploring the evolutionary path of the W chromosome in *Dendrolimus* species, we sequenced and assembled the female *D. sibiricus* genome. The assembly we present here comprises 30 telomere-to-telomere sequences corresponding to autosomes and the Z sex chromosome, as well as a reconstructed W sex chromosome scaffold and circular genomes of mitochondria and endosymbiotic *Wolbachia*. Technically, the assembly of high-quality chromosome-scale genomes is becoming a much simpler task thanks to the development of precise long-read technologies, although it does not eliminate the need for further thorough validation and improvement to ensure assembly reliability (Mc Cartney et al. 2022). Since the hifiasm assembler is optimized toward resolving biallelic sequences, hemizygous sex chromosomes, especially the repetitive W chromosome, are more prone to misassemblies. Using the level and the ratio of contig coverage of male versus female genomic libraries, we successfully identified the Z and reassembled the W chromosome. Along with standard metrics, we also validated our assembly with a new alternative measure of Lepidoptera-specific genome assembly completeness, which is based on the known saturation of lepidopteran telomeric regions with specific non-LTR transposons (Lukhtanov and Pazhenkova 2023). It clearly showed the conservation of such saturation in the studied *Dendrolimus* species, and superior completeness of chromosome ends in *D. sibiricus* and *D. kikuchii* assemblies compared with *D. punctatus*.

Importantly, cytological analyses confirmed that the number of chromosomes for *D. sibiricus* is $n = 30$, including a female W chromosome condensed into sex heterochromatin body in the nuclei of interphase cells. For the genus *Dendrolimus*, the number of chromosomes was previously established for *D. pini* ($n = 30$) (Kernewitz 1914; Kernewitz 1915), *D. spectabilis* ($n = 30$) (Makino 1933a), *D. jezoensis* ($n = 30$) (Makino 1933b), *D. punctatus* ($n = 30$) (Zhang et al. 2020), and *D. kikuchii* ($n = 29$) (Zhou et al. 2021). Thus, the number $n = 30$ found in *D. sibiricus* coincides with the modal number for the genus and is one of the most common in the order Lepidoptera (Pazhenkova and Lukhtanov 2023).

With a few exceptions, the organization and composition of both autosomes and the Z chromosome are highly consistent among the genomes of the *Dendrolimus* species examined here, further supporting the correctness of the *D. sibiricus* assembly. Surprisingly, the *D. punctatus* assembly was found to have an excessive number of intrachromosomal inversions, in contrast to its sympatric sister (sub)species, *D. tabulaeformis* (aka *D. punctatus tabulaeformis)*. Moreover, the level of divergence between these two genomes was observed to be significantly higher than that for the *D. sibiricus*—*D. pini* pair (Fig. 5b). Given that hybridization between *D. punctatus* and *D. tabulaeformis* has been shown to occur (Zhao et al. 1992; Kononov et al. 2016), such crosses likely explain our observation that the reconstructed W chromosome of *D. tabulaeformis* is evolutionarily more closely related to the Z chromosome of *D. punctatus* than to its own (Fig. 8h). With respect to the highly conserved Z chromosome, substantial rearrangement was found only in the *D. kikuchii* assembly (Zhou et al. 2021), where the formation of a neo-Z chromosome appears to have occurred through the fusion of an unpaired Z chromosome and an autosome. Such Z-autosome fusions are quite common among different families of Lepidoptera (Šíchová et al. 2016; Yoshido and Marec 2023;

Wright et al. 2024). The assembled genome of the *D. kikuchii* female lacked W chromosome. At the same time, fragments of the second copy of the fused autosome were present in the assembly, which may be components of the neo-W chromosome.

Structure of chromosome W

Like the W chromosomes of most other lepidopteran species, the structure of the *D. sibiricus* W, as reported here, was found to be extremely rich in recently amplified repeats and segmental duplications of their successions, casting doubt on the possibility of making reliable phylogenetic inferences. Altogether, both repetitive and unique structural constituents of the *D. sibiricus* W chromosome point to its evolutionary novelty. The repetitive fraction of W is clearly depleted for highly divergent repeats with reduction of some individual classes, contrasting with the rest of the genome. The horizontally acquired and subsequently duplicated fragment of the symbiotic *Wolbachia* genome shows the most similarity with the circular genome of modern *Wolbachia* strain, including the one sequenced along with the *D. sibiricus* genome.

Finally, the W chromosome of *D. sibiricus* contains significant extended regions of synteny with its own Z chromosome, but not with any autosome. Importantly, highly homologous syntenic blocks occur throughout the entire extent of the Z chromosome, retaining their colinearity, and are most continuous in the distal regions of the chromosome. Thus, it is unlikely that the W chromosome acquired these blocks stepwise and independently. Instead, the Z chromosome likely served as the direct predecessor of the W chromosome, with this transformation occurring after the divergence of *D. sibiricus* and *D. pini*.

Function of chromosome W

As in many other Lepidoptera, the *D. sibiricus* Z chromosome contains one of the key genes of the sex determination system, *Masc*. The protein it encodes is the hierarchically most upstream among the conserved elements of the sex determination cascade, inducing the male-specific splicing of the Doublesex mRNA (Kiuchi et al. 2014; van't Hof et al. 2024; Moronuki et al. 2025). Masc is also required for subsequent dosage compensation of Z-linked genes in males (Kiuchi et al. 2014).

However, the role of the *D. sibiricus* W chromosome in sex determination or female fitness traits remains unknown. The feminization factors known to date, which all act functionally as Masc repressors, are remarkably variable and evolutionarily independent. Among them, W-linked (Kiuchi et al. 2014; Harvey-Samuel et al. 2022; Moronuki et al. 2025), autosomal (Yoshido and Marec 2023), and even cytoplasmic (Fukui et al. 2024) factors have been described. Most known W-linked factors are multicopy and closely related to transposons specific to the W chromosome. The W of *D. sibiricus* is also enriched with repeats that are almost W-specific (Fig. 6b and c). Using a homology-based approach, we searched for potential clusters of anti-Masc RNAs on the W chromosome but did not identify a candidate matching the criteria. Protein-coding W genes derived from the Z chromosome have undergone pronounced pseudogenization. We found no reliable evidence of their expression in somatic cells, which, however, is consistent with the observed condensation of the W chromosome into heterochromatic bodies. Laterally transferred fragments of *Wolbachia* genome lack features indicative of adaptation of their transcription in the eukaryotic system. Furthermore, an analogous lateral transfer event has not been identified in other studied species with similar W chromosome organization.

In this regard, we are inclined to believe that the *D. sibiricus* W chromosome may not carry a sex-determination factor, but rather that its main role is to form a stable bivalent with the Z during meiosis. This is supported by the topology of the preserved homology between Z and W, specifically in both their distal regions, observable across several species of the subfamily *Lasiocampinae* and beyond. Recently, it has been shown for *B. mori* and *Plodia interpunctella* that the initial recognition and pairing of holocentric homologs takes place precisely at the gene-rich subtelomeric distal regions (Hockens et al. 2024). Following this rationale, the fusion of sex chromosomes with a pair of autosomes (so-called neo-Z/neo-W system), common among Lepidoptera, likewise reestablishes the homology of the ZW bivalent.

Therefore, the syntenic blocks inherited from the Z chromosome by the W are structurally and functionally similar to pseudoautosomal regions (PARs) typical of sex chromosomes in many other animals. However, in the case of lepidopteran insects, there is one significant exception: these regions of the W chromosome do not recombine due to achiasmatic meiosis in females, ultimately leading to their degradation.

Evolution of chromosome W

An exceptional variation in the composition of W chromosomes among different lepidopteran species gave rise to the hypothesis of their multiple, independent origins, involving different evolutionary pathways (Fraïsse et al. 2017; Han et al. 2024). At first glance, this hypothesis opposes that of a common early origin of the WZ/ZZ system in ditrysian species. However, what actually matters in the latter hypothesis is not so much the acquisition of an ancestral W chromosome itself, but rather the acquisition of molecular mechanisms responsible for its maintenance and condensation into sex chromatin.

In our study, we showed that the W chromosomes of *D. sibiricus* and two *Lasiocampinae* species, *D. tabulaeformis* and *L. quercus*, are independent derivatives of the respective Z chromosomes of each species. Consequently, despite structural similarities, these W chromosomes did not evolve from a common ancestral chromosome but arose convergently.

Although we observed significant local similarity among the Ws of these species (Fig. 8a), this similarity entirely reflected that of the Z-derived regions and individual genomic repeats, not their arrangements. It is worth noting the advantages of the phylogenetic approach used in the analysis of W chromosome origins compared to a simple homology-based approach. Several Z-derived W-linked genes of *D. sibiricus* are also found in the W chromosomes of *D. tabulaeformis* (two genes) and *L. quercus* (five genes). Technically, they can be considered orthologs inherited from a common ancestral W chromosome, with a level of divergence corresponding to the evolutionary distances between these species. However,

only the phylogenetic analysis involving Z-paralogous genes of all these species allowed us to unequivocally conclude that the duplication of Z-genes onto the W chromosome occurred independently in each species.

At the same time, due to the shared evolutionary mechanism, the organization of the W chromosomes of these species displays some common features. All three W chromosomes show a significant bias toward a predominance of weakly diverged repeats in contrast to the rest of the genome. Moreover, *L. quercus* reproduces the same pattern of synteny of subtelomeric regions of both sex chromosomes as seen in *D. sibiricus*. We also found this pattern in the much more phylogenetically distant genome of *Hipparchia semele* (Nymphalidae) (Macgregor and Saccheri 2023).

In Fig. 9, we have depicted hypothetical scenarios of how a replacement of the ancestral W chromosome could have taken place in the evolution of the *D. sibiricus* W chromosome. Generally, for such a complete replacement to occur, one of the following conditions must be met:

(i) The original W chromosome initially did not carry an obligate female-determining factor, which made it possible for an alternative W chromosome to emerge and coexist with it. Numerous cases have been described where the ditrysian W chromosome plays no role in female sex determination. For example, W or neo-W chromosomes play no role in sex determination or reproduction of *Samia cynthia* subspecies that harbor a varied repertoire of sex chromosomes, including the complete absence of W in *S. cynthia ricini*. Instead, the sex of *S. cynthia* is determined by the Z:A ratio (Yoshido and Marec 2023). Another striking example of W-independent sex determination was recently described in *Bicyclus anynana*, in which the number of alleles or allele heterozygosity in a single Z-linked *BaMasc* gene is the primary sex-determining switch. Again, this resulted in occurrence of reproductively competent females lacking a W chromosome (van't Hof et al. 2024). This condition is more plausible for the observed case of independent turnover of the W in several related species. We propose a simple scenario involving a rearranged copy of Z that has lost the male-determining locus, including the extreme case of a B chromosome formation. The rearranged or mutated product of the Z no longer participates in the determination of the Z dose, leading to the development of a female. Since it is more homologous to the original Z than the ancestral W, it can displace the latter at the level of bivalent formation.

(ii) A feminization factor unlinked to the W chromosome, dominant over the W-linked one, appeared. Here, the first scenario is analogous to the "single-Z turnover" model proposed by Han et al. (2024), in which an unspecified feminizing factor is somehow acquired by a copy of Z, which competitively replaces the ancestral W. It must be acknowledged that this model was initially proposed and justified much earlier—to describe the evolution of Y chromosome from a copy of X in guppies, and the authors referred to its possible involvement in the evolution of lepidopteran WZ/ZZ systems (Charlesworth et al. 2021).

The second and somewhat exotic scenario describes the double involvement of feminizing *Wolbachia* in the evolution of sex chromosomes. *Wolbachia*—host interactions can potentially act both as a factor of the loss of the ancestral W-chromosome (Kageyama et al. 2017) and as a driver of the subsequent reciprocal reorganization of the sex determination system (van't Hof et al. 2024), including rebirth of the W-chromosome. In *Eurema mandarina*, the feminizing *Wolbachia* was shown to disrupt the inheritance of sex chromosomes such that infected females lay only Z0 eggs, thereby resulting in the loss of female-determining W chromosome in infected lineages c.

The prevalence of *Wolbachia* in the modern *D. sibiricus* population reaches 90% (Bykov et al. 2020). It can be assumed that the initial aggressive strategy of feminization or male killing may have been the cause for such a successful long-term spread of the symbiont in the host population, which could further trigger the turnover of sex chromosomes. It can also be suggested that the lateral gene transfer from *Wolbachia* into a copy of the Z could somehow act as a nuclear factor of feminization, resulting in the formation of the W chromosome. A similar model was previously proposed for the isopod *Armadillidium vulgare*, according to which the female-determining function of the reconstituted W was achieved by acquiring a genome fragment from feminizing *Wolbachia* symbiont (Leclercq et al. 2016). As already mentioned, we did not find evidence supporting functionality of the transferred fragment in *D. sibiricus*. Furthermore, this scenario does not apply to other studied species with a similar W chromosome structure.

It should be noted that the observed structure of the *D. sibiricus* and *D. tabulaeformis* W chromosomes largely reproduces the structure of B chromosomes, pointing to a similarity of the mechanisms of their formation and maturation. Most notable among these similarities is the exceptional enrichment of the *D. tabulaeformis* W chromosome with mitochondrial DNA fragments, which reproduces the classic example of saturation of the B-chromosome of rye by mitochondria- and chloroplast-derived DNA (Chen et al. 2024). The horizontal transfer of the *Wolbachia* genome into the W chromosome of *D. sibiricus* can be interpreted in a similar way.

The second similarity is the saturation of the W chromosome with mobile elements, including dispersed distribution of telomere-specific repeats and segmental duplications of stretches of repeats. Third is the presence of fragments of A chromosomes, in this case, the Z chromosome. Perhaps the borrowing of telomeric fragments of Z could have stabilized the initial proto-B chromosome. Its very origin may have been facilitated by genomic instability due to the observed ongoing expansion of LTR repeats. Interestingly, there are known examples of B chromosomes manipulating the sex of insects to achieve their meiotic drive. A supernumerary B chromosome (PSR, paternal sex ratio) in the haplodiploid jewel wasp *Nasonia vitripennis* causes female-to-male conversion by destroying the sperm's hereditary material in young embryos (Benetta et al. 2020).

Irrespective of the actual, historically correct scenario, the rearrangement of the Z chromosome seems to be an advantageous and short-cut process to generate a new version of the W chromosome, a phenomenon often seen in lepidopteran evolution. The Z-derivative can immediately form a reliable bivalent with Z in meiosis. At the same time, it has been shown that the dosage compensation system developed in the Lepidoptera

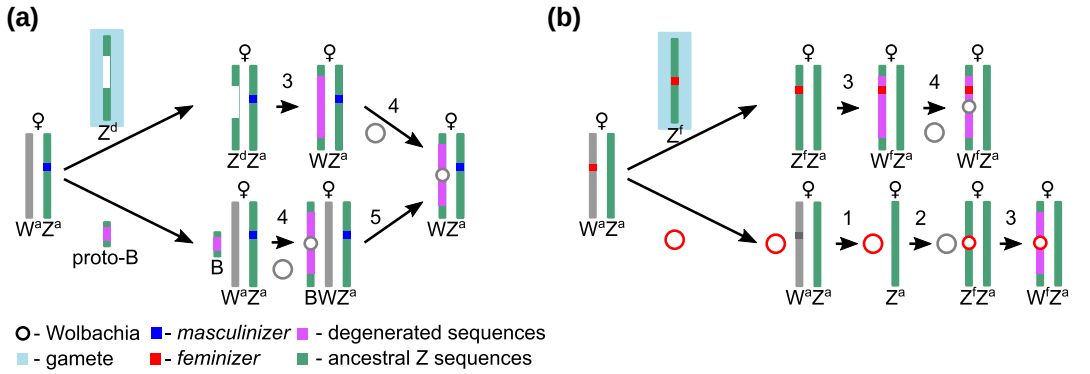

Figure 9 Hypothetical evolutionary trajectories of the *D. sibiricus* WZ/ZZ system, where two functional variants of the ancestral WZ/ZZ system are considered. a) Z-counts sex determination. Here, two scenarios are presented. In the first scenario (top), a Z^d that has lost its *masculinizer* locus completely displaces the ancestral W^a chromosome (3) and acquires a *Wolbachia* genome via HGT (4). In the second scenario (bottom), a maturing Z^a-derived B-chromosome acquires a *Wolbachia* genome via HGT (4) and competitively displaces the W^a (5). b) W-linked feminization. The first scenario is based on the "single-Z turnover" model of Han et al. (2024), in which an unspecified feminizing factor is somehow acquired by a copy of Z, ultimately replacing W (3) and acquiring a *Wolbachia* genome via HGT (4). In the second scenario (top), a feminizing *Wolbachia* strain causes the loss of the ancestral W^a (1), as previously described for *Eurema mandarina* (Kageyama et al. 2017), and its genome is then co-opted by the host via HGT into a copy of the Z^a (i.e. Z^f) (2), as described for *Armadillidium vulgare* (Leclercq et al. 2016). Subsequent degeneration of non-recombining Z^f into the final W^f is provided by the achiasmatic meiosis in females (3).

allows for some flexibility in the altered copy number of Z-linked genes (Yoshido and Marec 2023). It was also found that chromosomal fusions are biased toward the Z sex chromosome in the evolution of Lepidoptera (Wright et al. 2024), suggesting milder consequences of such a rearrangement.

We speculate that the evolutionary trajectory for the Z-to-W conversion proposed here for the four species examined, including *D. sibiricus*, may be common in the Lepidoptera. However, because degeneration is the only possible evolutionary fate for a non-recombining W, gradually erasing traces of homology with the ancestral Z, the scenarios proposed here may eventually become undetectable.

Methods

Samples

The female *D. sibiricus* larva used for genome sequencing was collected by Alexandr Ageev at 55.027991N, 96.007906E in Krasnoyarsk Krai, Russia. Given the high morphological similarity of *D. sibiricus* and *D. superans* species, the species identity of the collected sample is further confirmed by the following. The insect specimen was collected from the Siberian fir tree (Abies sibirica), the common host plant of *D. sibiricus* and in a geographic area where *D. sibiricus* predominates. The differential marker sequence of the ITS2 region from the collected sample has 100% identity to those of *D. sibiricus* samples from the phylogenetic study of Kononov et al. (2016), that were collected in Krasnoyarsk, Sakhalin, and Khakassia (GenBank IDs: KJ007736, KJ007738, KJ007739, KJ007743, KJ007744, KJ007746-KJ007752). The sex of the larva was initially determined based on the morphology of developing gonads, with additional control by qPCR testing (Belousova et al. 2019).

Siberian moths used for RNA-sequencing studies (adults, eggs, and pupae) were collected in the same location as the aforementioned female used for genome sequencing. For cytogenetic analysis, *D. sibiricus* larvae of different instars (three to five) were collected by Y. Tokarev in Russia, Altaisky Krai, Shipunovsky District, 52.181384N, 81.898631E, 18-19 May 2023. For qPCR study, insects were collected in taiga forest of Krasnoyarsk Krai, Russia, 53.965235N, 105.966797E.

DNA extraction, WGS library construction, and sequencing

The gut was dissected from a single female last instar larva of *D. sibiricus*, washed from gut content, and frozen in liquid nitrogen. High-quality genomic DNA was extracted using MagAttract HMW DNAKit 48 (Qiagen, USA) with added RNase A according to the manufacturer's protocol. The extracted DNA that met quality and quantity standards was split into two parts, which were used to construct a PacBio HiFi WGS library and a short fragment WGS library.

The HiFi SMRTbell library was constructed using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, CA, USA) following the manufacturer's instructions. The library was sequenced on the PacBio Sequel II system (Pacific Biosciences), using 1 SMRT Cell 8M, Sequel II binding Kit 2.0, and Sequel II Sequencing Kit 2.0 (Pacific Biosciences). Consensus HiFi reads were generated using CCS v.6.4.0 (Pacific Biosciences) software with enabled "–hifi-kinetics" parameter. The short fragment library with an insert size of 350 bp was sequenced in paired-end mode (2 × 150 bp) on Illumina NovaSeq 6000 platform (Illumina, USA). Both libraries were prepared and sequenced by Novogene company (Hong Kong, China).

The gut of one male individual of *D. sibiricus* was homogenized and treated with proteinase K and RNase A followed by phenol −chloroform extraction of genomic DNA. The 100 ng gDNA was incubated with dsDNA Fragmentase (New England Biolabs, MA, USA) for 35 min at 37 °C and purified with Agencourt AMPure XP beads (Beckman Coulter). A library was prepared using

NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs) according to the manufacturer's instructions and sequenced in paired-end mode (2 × 150 bp) on Illumina NovaSeq 6000 platform.

RNA extraction, cDNA library construction, and sequencing

In order to annotate a wider range of expressed genes in the *D. sibiricus* reference genome, an equimolar pool of total RNA samples from various tissues and life stages of both sexes was sequenced using both polyA-mRNA enrichment and rRNA depletion approaches.

RNA was extracted from fresh or frozen samples using the PureZOL (Bio-Rad, USA) according to the manufacturer's protocol. Samples were purified from DNA and other impurities using DNase I (New England Biolabs, USA) treatment followed by purification on Agencourt RNAClean XP (Bekman Coulter) beads. The quality and amount of isolated RNA was assessed using a NanoDrop 2000 spectrophotometer.

The scheme of pooling is described on Figure S6. Various body parts (head, thorax, abdomen) were used to isolate RNA from adult female and adult male. RNA was mixed equimolarly separately for each individual. Next, the RNA of the adult male and adult female was mixed in a ratio of 1:1. For late instar larvae, the head and cuticle (i.e. hypodermis and muscles) were used separately, and the total RNA was mixed in a 1:1 ratio. Also from larvae, we dissected two sets of organs/tissues (1: salivary glands, trachea, fat body; 2: Malpighian tubules, intestines, gonads), and the total RNA was mixed in a 1:1 ratio. The isolated RNA was further mixed equimolarly separately for each larva. Separation into different tissues was not applied for the female pupa and eggs. The final pooling was 1:2:1:2 (egg:larva:pupa:adult).

The rRNA-depleted cDNA library was prepared using the Zymo-Seq RiboFree Total RNA Library Prep Kit (Zymo Research) according to the manufacturer's protocol. The polyA-cDNA library was prepared using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (NEB) according to the manufacturer's protocol. Fragment size selection was performed using Agencourt AMPure XP Beads (Beckman Coulter), after which several cycles of library amplification were performed. The size and quantity of the resulting libraries were determined using an Agilent 2100 Bioanalyzer. The resulting libraries were sequenced in paired-end mode (2 × 250 bp) on Illumina NovaSeq 6000 platform by Novogene (Hong Kong, China).

Validation of W-chromosome-specific sequences by qPCR

Genomic DNA was extracted from individual midgut samples using QIAamp DNA Mini Kit (Qiagen, USA) according to the manufacturer's protocol. The targeted regions of autosomal *SIWI* gene and Z-linked *Masc* gene (single-copy in ZW females and double-copy in ZZ males) included intronic sequences and served as internal copy number controls. A pair of primers was designed to specifically target the insertion site of *Wolbachia* genome in the W chromosome, as opposed to the circular genome of *Wolbachia*. A special primer design was performed for the repetitive sequences that showed the most specific enrichment in the W chromosome compared to the rest genome, namely: "6f6010", "l1f232", "l1f257", and "r1f1028". For each repeat, a multiple sequence alignment of its copies was prepared using mafft aligner and screened for W-specific 20-mers that do not occur in the rest genome. The corresponding regions with at least two W-specific mismatches were used for primer selection. The list of primers is provided in Table S8.

Real-time PCR reactions were prepared using the BioMaster HS-qPCR SYBR Blue kit (Biolabmix, Russia) and performed on a BioRad CFX96 thermal cycler with cycling parameters: 95 °C for 5 min and 40 cycles of: 95 °C for 10 s, 60 °C for 1 min, plate read. The following melting curve analysis was performed from 60 °C to 95 °C with 0.5 °C increment each 5 s.

Analysis of *D. sibiricus* karyotype

For cytogenetic analysis, four males and four females at the last larval stage were used. Testes and ovaries were removed from the last instar larvae and immediately placed in the 3:1 fixative (three parts of ethanol and one part of glacial acetic acid) for 12 h. Then gonads were stained with 2% acetic orcein for 36 −48 h at 20 °C. We used a two-phase method of male chromosome analysis as described in Lukhtanov et al. (2006). In the first phase, the stained testes were placed into a drop of 40% lactic acid on a slide, the gonad membranes were torn apart using fine needles, and several drops of 45% acetic acid were added. Then the preparation was covered with a cover slip (without pressure on the coverslip), and intact or weekly squashed cells were studied and photographed. The first phase was most useful for counting the number of chromosome bivalents at metaphase I (MI) of male meiosis. In the second phase, different stages of chromosome spreading were studied using a slight, gradually growing pressure on the coverslip. The second phase was most useful for studying the chromosome structure. Both male and female somatic interphase nuclei were used to analyze the presence/absence of sex chromatin (=W chromosome). Leica DM2500 light microscope equipped with HC PL APO 100×/1,44 Oil CORR CS lens and S1/1.4 oil condenser head was used for bright-field microscopy analysis.

Genome assembly of *D. sibiricus*

The draft partially phased genome assembly of *D. sibiricus* was built from PacBio HiFi data using hifiasm v0.15.4-r343 (Cheng et al. 2021). Assembly quality assessment metrics were obtained using QUAST v5.0.2 (Gurevich et al. 2013) and BUSCO v5.2.2 (Simão et al. 2015). Assembly graphs were visualized using Bandage v0.8.1 (Wick et al. 2015).

Since the contigs of chromosomes 25 and 29 were broken in the regions of highly repetitive rDNA and histone gene clusters, respectively, they were unambiguously scaffolded based on the unitig assembly graph with the introduction of a gap. The reliability of the scaffolds was confirmed using a whole-genome synteny map with other *Dendrolimus* species. Remaining short, ambiguous contigs containing rDNA and histone clusters were removed from the assembly. The assembly of sex chromosomes is described in the next paragraph.

To polish the primary genome assembly, we searched and eliminated both sequencing errors and true heterozygous sequence variants that distorted the ORFs of protein-coding genes in the assembly. PacBio HiFi data were aligned with minimap2 v2.20-r1061 (Li 2018) to the primary assembly and variant calling was performed using DeepVariant v1.4.0 (Poplin et al. 2018) with "WGS" model type. Variants with "0/0" genotype were removed, and those with allele depth for REF allele less than 2 and ALT/REF depth ratio more than 5 were selected as homozygous alternative. 419 of them were confirmed by accompanying short-read WGS data and introduced into the primary assembly using the bcftools v1.2 consensus tool. For the analysis of deleterious heterozygous variants in the reference genome, coding sequences were extracted from *de novo* transcriptome assembly using TransDecoder v5.5.0 (Haas, BJ. https://github.com/TransDecoder/TransDecoder) and aligned to the genome assembly using gmap v2021-08-25 (Wu and Watanabe 2005) aligner. SAM output with uniquely mapped transcripts was parsed with in-house python script (https://github.com/sashulkaSh/SibMoth/blob/main/parse_cigar.py) in order to find genomic indels that introduce either frameshifts or premature STOP codons in aligned ORFs. The obtained positions were intersected with DeepVariant's dataset and further analyzed in IGV browser (Robinson et al. 2011). The filtered set of alternative alleles was inserted into the primary assembly using the bcftools.

Identification and reassembly of sex chromosomes

The primary and haplotype-aware assemblies of *D. sibiricus* were inconsistent in the sequence of *D. sibiricus* Z chromosome, and the sequence of the potential W chromosome was highly fragmented. A modified chromosome quotient (CQ) approach was applied to reveal genomic sequences of Z and W sex chromosomes. Short-read WGS data for male and female individuals were mapped using bowtie2 v2.4.4 (Langmead and Salzberg 2012) to the set of unitigs (p_utg) which are the most reliable and unbiased representation of the assembled diploid genome. For each library, an array of mean read coverages in 50 Kb windows was computed using bedtools without any mapping quality threshold and normalized to its median. CQ values were computed for each window as the ratio of male (Cm) versus female (Cf) normalized coverage. The windows with $Cf > 0.5$ and $Cm > 1.5$ and $CQ > 1.5$ were defined as Z-positive, while those with $Cf > 0.55$ and $Cm < 0.45$ and $CQ < 0.45$ as W-positive. The windows belonging to mtDNA, histone, and rDNA genes that suffer from high inter-individual copy number variation were masked. The data was further aggregated at the level of unitigs. Unitigs that have Z-positive windows and the median $CQ > 1.5$ were defined as Z-unitigs. Unitigs that have W-positive windows and the median $CQ < 0.55$ were defined as W-unitigs.

Based on the identified list of Z-positive unitigs, the complete sequence of the Z was identified among the haplotype-resolved contigs (hap1/hap2.p_ctg). Raw HiFi data were mapped to the unitigs using minimap2, and those mapped to the W-unitigs were extracted using samtools and reassembled using hifiasm tuned for haploid assembly (-l0 –n-hap 1). The produced contigs were further inspected to detect telomeric signals and unambiguous contig overlaps in order to arrange them in a pseudo-scaffold.

Annotation and analysis of genomic repeats

RepeatModeler v2.0.1 (Flynn et al. 2020) was used to create three custom libraries of repeats for RepeatMasker from genome assemblies of: (i) *D. sibiricus* and *D. punctatus*, (ii) *L. quercus*, and (iii) *H. semele*. The first library used in conjunction with the Master RepeatMasker Database CONS-Dfam_3.1-rb20181026 to annotate and mask repeats in the genome assemblies of *D. sibiricus*, *D. pini*, *D. punctatus*, *D. kikuchii*, and *D. tabulaeformis* (reassembled) using RepeatMasker v4.1.0 (Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015 http://www.repeatmasker.org). Accession numbers of the used assemblies are described in the Table S9.

The second and third libraries were used in the same manner to annotate *L. quercus* and *H. semele* genomes, respectively. The Kimura 2-Parameter divergence metric was calculated separately for the W chromosome and the rest genome using the calcDivergenceFromAlign.pl utility, and the repeat landscape was constructed using createRepeatLandscape.pl, both included with RepeatMasker. The names of short tandem repeats were unified using the lexicographically smallest variant of unit notation.

To analyze the spatial distribution of repeats in chromosomes of *D. sibiricus*, *D. pini*, *D. punctatus*, and *D. kikuchii*, the number and proportion of bases occupied by each repeat family in the whole genome and in 50 kb sliding windows with a 25 kb step were calculated. In the case of the telomere-oriented analysis, for each repeat family, the obtained values were aggregated by the window index counted from each end of the chromosome scaffolds to their middle. The enrichment with a particular repeat family was calculated as the ratio of its proportion in the window to its proportion in the genome.

Gene annotation of *D. sibiricus* genome

The overall pipeline of gene annotation is shown in the Figure S7. The main experimental source for gene structure modeling were two *D. sibiricus* representative transcriptome libraries obtained using polyA-mRNA enrichment and rRNA depletion approaches. The corresponding short-read data were preprocessed to remove low-quality and technical subsequences.

For reference-assisted transcriptome assembly, data were aligned to the *D. sibiricus* genome assembly using STAR v2.7.9a (Dobin et al. 2013). StringTie v2.1.7b (Pertea et al. 2015), PsiCLASS v1.0.3 (Song et al. 2019), and TransMeta v.1.0 (Yu et al. 2022) were applied to the aligned data for genome-assisted reconstruction of transcript structure. Portcullis v1.2.4 (Mapleson et al. 2018) was used to generate a set of reliable splice junctions. Trinity v2.13.2 (Grabherr et al. 2011) was used both for *de novo* and genome-guided transcriptome assembly. In *de novo* mode, the polyA-mRNA library was also assembled separately. The three Trinity assemblies were aligned to the *D. sibiricus* genome assembly using gmap and minimap2. The parallel use of minimap2 is dictated by the fact that, unlike gmap, it outputs all alternative loci of the multi-mappable transcripts.

Additional sources of evidence included transcriptome/proteome data from a closely related *Dendrolimus* species and *B. mori*. Since the assemblies of the *D. punctatus* and *D. kikuchii* genomes available at the time of analysis were not accompanied by gene

annotations, we prepared them ourselves. Raw RNA-seq data of *D. punctatus* and *D. kikuchii* were downloaded from NCBI SRA database (Table S9) and aligned using STAR to the corresponding genome assemblies. Transcript structures were reconstructed using StringTie and the corresponding sequences were extracted using the gtf_genome_to_cdna_fasta.pl script. The longest ORFs were predicted in them by TransDecoder with the aid of blastn and hmmscan searches against UniProtKB/Swiss-Prot (https://www.expasy.org/resources/uniprotkb-swiss-prot) and Pfam-A (https://www.ebi.ac.uk/interpro/download/pfam/) databases, respectively. The resulting protein sets for *D. punctatus*, *D. kikuchii*, and the one for *B. mori* (GCF_014905235.1) were aligned to the *D. sibiricus* assembly using SPALN v2.4.13f (Gotoh 2008), with the best hit with coverage≥90% and identity≥90% retained.

All data obtained above (StringTie, PsiCLASS, TransMeta transcriptome assemblies, Trinity transcriptome alignments, splice junctions, protein alignments of other species) were submitted with different weights to the Mikado v2.3.4 (Venturini et al. 2018) pipeline to generate the best transcript set. The embedded homology searches were run as follows: blastx search was done against the Uniprot:UniRef_90 database (https://www.uniprot.org/help/uniref) limited to Lepidoptera species; TransDecoder was run as described above, but with enabled "complete_orfs_only" and "min_cds_length = 50" parameters; the output of hmmsearch was filtered to exclude HMMs associated with transposable elements. The scoring rules for transcript models by Mikado were adjusted toward more stringent filtering values, which in turn were chosen in such a way as to retain the largest number of BUSCO genes. Also, the optional file of external scores for transcript selection was prepared introducing two alternative filtering parameters, the proportion of repeats in CDS based on RepeatMasker predictions and the coding potential of CDS computed using CPC2 v1.0.1 tool (Kang et al. 2017). The final configuration is listed in the supplementary file (https://github.com/sashulkaSh/SibMoth/blob/main/Dendrolimus.yaml).

The resulting gene models were manually curated and incorrect models were fixed and submitted to the rerun of Mikado as unchangeable reference. In particular, chimeric annotations of tandemly arrayed genes were detected by inspecting adjacent multiple alignments of a same transcript that were produced by minimap2 with maximum gap on the reference (-G) set to 10 or 20 Kb. Predicted non-coding transcripts were also analyzed for coding potential using CPC2 and for conserved protein domains using hmmsearch, with some protein-coding genes being retrieved. Functional annotation of the protein-coding genes was inferred using the InterProScan v5.61-93.0 (Jones et al. 2014).

Analysis of synteny between the W and the other chromosomes in *D. sibiricus* and other Lepidoptera species

The alignment of chrW contigs to the remaining *D. sibiricus* chromosomes was performed using minimap2. Synteny blocks were obtained using the Asynt tool (https://github.com/simonhmartin/asynt). Circular visualization of W−Z synteny was performed using the circlize v0.4.16 (Gu et al. 2014) CRAN R package. The orthology between the W-linked genes and Z-linked genes of *D. sibiricus*, *D. pini*, *D. kikuchii*, *D. punctatus*, *D. tabulaeformis*, *L. quercus*, and *B. mori* was refined using the OrthoFinder v2.5.5 (Emms and Kelly 2019). Out of the found orthogroups, only 14 were represented by all studied species and contained more or less preserved ORF of W-linked orthologs. These genes were aligned using MAFFT v7.475 (Katoh and Standley 2013), trimmed using trimAl v1.4.rev15 (Capella-Gutiérrez et al. 2009), and the alignments were concatenated using AMAS (Borowiec 2016). The phylogenetic tree was constructed using IQ-TREE v2.2.2.6 (Minh et al. 2020) with 10,000 bootstraps without partitioning.

To examine the existence of similar highly preserved W−Z synteny in other Lepidoptera species, a list of Lepidoptera genomes for which both sex chromosome sequences are available was taken from the study of Chen et al. (2023) and the corresponding assemblies were downloaded from the NCBI Genomes database. The alignment of W chromosome sequences to the rest chromosomes of the same species was performed using minimap2 and visualized as dotplots using the pafr v0.0.2 CRAN R package. Since the screening revealed *Hipparchia semele* (GCA_933228805.2) as having the most pronounced W−Z synteny, the W chromosome of this species was also analyzed for repeat content the same way as described for *D. sibiricus.*

The search for publicly available female genomes among the species closely related to *D. sibiricus* was performed. The only found Lasiocampinae genome with the annotated W chromosome was *L. quercus.* A deeper search revealed the genome assembly of *D. tabulaeformis* (GCA_042997425.1; sample of undefined sex) that contained large surplus unplaced scaffolds and accompanied by raw Hi-C and HiFi data. The methods used for confirmation and recovery of the W chromosome of *D. tabulaeformis* are described in Supplementary Methods. The analysis of W−Z synteny and repeat content in these species was performed as described above.

Whole-genome analysis of orthology and collinearity of genes

Using liftoff v1.6.3 (Shumate and Salzberg 2021) with the "polish" option, the *D. sibiricus* annotation was transferred to the genome assemblies of *D. pini*, *D. punctatus*, *D. kikuchii*, *D. tabulaeformis* (reassembled), and *L. quercus*. For *D. punctatus* and *D. kikuchii*, Mikado was further run to aggregate the transferred annotation with transcript models generated using StringTie as described above. For the resulting annotations of the three *Dendrolimus* species, CDSs and UTRs were redefined using GenomeTools v1.5.9 (Gremme et al. 2013) cds tool. A representative set of proteins was extracted from each annotation of *Dendrolimus* species and *B. mori* (GCF_014905235.1). GENESPACE v1.2.3 (Lovell et al. 2022) was used to infer orthologous protein groups and analyze gene collinearity in the studied genomes.

Genome-wide phylogenetic analysis of *Dendrolimus* species

Alignment of *D. sibiricus* and *D. pini* genomes using minimap2 indicated their high whole-genome identity comparable to the identity level of the two partially phased haplotypes of *D. sibiricus*. Therefore, both haplotypes of both species were used in

the phylogenetic inference together with the primary (haploid) genome assemblies of *D. kikuchii*, *D. punctatus*, *D. tabulaeformis*, *L. quercus*, and *B. mori* serving as an outgroup. In particular, gene annotation of *D. sibiricus* was transferred on the alternative haplotypes of *D. sibiricus* and *D. pini* (GCA_949752875.1) and the sets of proteins were extracted as described in the previous section.

OrthoFinder was used to reveal single-copy orthologous genes of the seven studied genomes. The corresponding proteins were aligned using MAFFT, and multiple alignments were trimmed by trimAl. Further analysis was performed in two ways. The first one used the partially phased haplotypes as they come from the assembly pipeline, while the second used reassigned marginal pseudo-haplotypes computed as follows. For each orthogroup, FastTree v2.1.11 (Price et al. 2010) was used to construct individual trees, and the phylogenetically closest pair of *D. sibiricus* and *D. pini* genes was assigned to pseudo-haplotype 1, while the remained pair was assigned to pseudo-haplotype 2. In both cases, the concatenated alignments were subjected to partitioned maximum likelihood phylogenetic analysis with 10,000 bootstrap iterations using the IQ-TREE.

Identification and analysis of *Wolbachia* genomic sequences

The complete circular genome of *Wolbachia pipiens* and the inserts of its fragments into the W chromosome were identified in the *D. sibiricus* genome assembly as a result of blastn search against the NCBI nr database. The reliability of the horizontally transferred sequences was supported by a number of long HiFi reads spanning the insertion sites in the nuclear genome as well as by qPCR analysis as described above.

A search for *Wolbachia* sequences in publicly available WGS data of other *Dendrolimus* species revealed significant matches in *D. pini* HiFi data (ERR11029700) but none in its genome assembly. Therefore, hifiasm reassembly of the *D. pini* genome was performed to recover two unrelated complete circular genomes of *Wolbachia*.

Molecular typing of the three revealed *Wolbachia* genomes was done using the sequence query tool of PubMLST web-service (https://pubmlst.org). Their gene annotation was performed using prokka v1.14.6 (Seemann 2014) pipeline. The resulting gene sequences were blastn-searched for homology to a set of *Wolbachia* cytoplasmic incompatibility (CI) genes, which were taken from the study of Martinez et al. (2021). For each gene, a single best match with the highest identity and coverage was kept. InterProScan was used to reveal conserved domains in the proteins of candidate CI genes of the three *Wolbachia* strains.

Acknowledgments

We thank Yuri S. Tokarev for providing insect samples. We acknowledge the collaborative center “Bioinformatics” ICG SB RAS (FWNR-2022-0020) for providing computational resources of HPC facilities.

Supplementary material

Supplementary material is available at *Molecular Biology and Evolution* online.

Funding

This work was supported by the grant of the state program of the “Sirius” Federal Territory “Scientific and Technological Development of the “Sirius” Federal Territory”, Agreement №24-03 date 27.09.2024 (for bioinformatics study). Genome sequencing was supported by Russian Science Foundation (grant No 20-64-46011). Analysis of karyotype (Vladimir Lukhtanov) was supported by the Russian Science Foundation grant No 24-14-00047 (https://rscf.ru/project/24-14-00047/) awarded to the Zoological Institute of the Russian Academy of Sciences (St. Petersburg, Russian Federation).

Data availability

The raw genome and transcriptome sequencing data generated during this study are available under NCBI BioProject identifier PRJNA1204066. The genome assembly of *Dendrolimus sibiricus* is deposited in NCBI GenBank under the accession GCA_046627935.1. Custom code and configuration files used for the reported analyses are available from the GitHub repository (https://github.com/sashulkaSh/SibMoth).

References

Baranchikov YN, Kondakov YP. Outbreaks of Siberian Moth, *Dendrolimus superans sibiricus* Tschtvrk in Central Siberia. In: Gottschalk KW, Fosbroke SL, editors. *Interagency gypsy moth research forum*. USDA Forest Service, NEFES, GTR-NE-368, Newton Square, Pennsylvania; 1997. p. 15.

Belousova I, Ershov N, Pavlushin S, Ilinsky Y, Martemyanov V. Molecular sexing of Lepidoptera. *J Insect Physiol.* 2019:114: 53–56. https://doi.org/10.1016/j.jinsphys.2019.02.005.

Benetta ED *et al.* Genome elimination mediated by gene expression from a selfish chromosome. *Sci Adv.* 2020:6:eaaz9808. https://doi.org/10.1126/sciadv.aaz9808.

Borowiec ML. AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ*. 2016:4:e1660. https://doi.org/10.7717/peerj.1660.

Botham M, Dainton K. The genome sequence of the Pine-tree Lappet moth, *Dendrolimus pini* (Linnaeus, 1767). *Wellcome Open Res*. 2024:9:538. https://doi.org/10.12688/wellcomeopenres.23053.1.

Bykov RA *et al.* Vysokij uroven’ vstrechaemosti endosimbionta *Wolbachia* v populyaciyah sibirskogo shelkopryada *Dendrolimus superans sibiricus* Tschetverikov, 1908 (Lepidoptera: Lasiocampidae) na territorii Rossii. *Zhurnal Obs. Biol*. 2020:81:387–393. https://doi.org/10.31857/S0044459620050036.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. Trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009:25:1972–1973. https://doi.org/10.1093/bioinformatics/btp348.

Charlesworth D, Bergero R, Graham C, Gardner J, Keegan K. How did the guppy Y chromosome evolve? *PLoS Genet*. 2021:17: e1009704. https://doi.org/10.1371/journal.pgen.1009704.

Chen J *et al.* The genetic mechanism of B chromosome drive in rye illuminated by chromosome-scale assembly. *Nat*

Commun. 2024:15:9686. https://doi.org/10.1038/s41467-024-53799-w.

Chen X *et al.* Unraveling the complex evolutionary history of lepidopteran chromosomes through ancestral chromosome reconstruction and novel chromosome nomenclature. *BMC Biol*. 2023:21:265. https://doi.org/10.1186/s12915-023-01762-4.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods.* 2021:18:170–175. https://doi.org/10.1038/s41592-020-01056-5.

Dobin A *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013:29:15–21. https://doi.org/10.1093/bioinformatics/bts635.

Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 2019:20:238. https://doi.org/10.1186/s13059-019-1832-y.

Flynn JM *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A.* 2020:117:9451–9457. https://doi.org/10.1073/pnas.1921046117.

Fraïsse C, Picard MAL, Vicoso B. The deep conservation of the Lepidoptera Z chromosome suggests a non-canonical origin of the W. *Nat Commun.* 2017:8:1486. https://doi.org/10.1038/s41467-017-01663-5.

Fukui T *et al.* Expression of the *Wolbachia* male-killing factor Oscar impairs dosage compensation in lepidopteran embryos. *FEBS Lett*. 2024:598:331–337. https://doi.org/10.1002/1873-3468.14777.

Gotoh O. A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. *Nucleic Acids Res*. 2008:36:2630–2638. https://doi.org/10.1093/nar/gkn105.

Grabherr MG *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 2011:29:644–652. https://doi.org/10.1038/nbt.1883.

Gremme G, Steinbiss S, Kurtz S. Genome tools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput. Biol. Bioinforma*. 2013:10:645–656. https://doi.org/10.1109/TCBB.2013.68.

Gu Z, Gu L, Eils R, Schlesner M, Brors B. Circlize implements and enhances circular visualization in R. *Bioinformatics*. 2014:30:2811–2812. https://doi.org/10.1093/bioinformatics/btu393.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013:29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086.

Han MJ *et al.* Multiple independent origins of the female W chromosome in moths and butterflies. *Sci Adv.* 2024:10:9851. https://doi.org/10.1126/sciadv.adm9851.

Harvey-Samuel T *et al.* Silencing RNAs expressed from W-linked PxyMasc "retrocopies" target that gene during female sex determination in *Plutella xylostella*. *Proc Natl Acad Sci U S A.* 2022:119:e2206025119. https://doi.org/10.1073/pnas.2206025119.

Hejníčková M *et al.* Degenerated, undifferentiated, rearranged, lost: high variability of sex chromosomes in geometridae (lepidoptera) identified by sex chromatin. *Cells*. 2021:10:2230. https://doi.org/10.3390/cells10092230.

Hockens C, Lorenzi H, Wang TT, Lei EP, Rosin LF. Chromosome segregation during spermatogenesis occurs through a unique center-kinetic mechanism in holocentric moth species. *PLoS Genet*. 2024:20:e1011329. https://doi.org/10.1371/journal.pgen.1011329.

Jeong JS, Kim MJ, Kim SS, Choi SW, Kim I. DNA data and morphology suggest an occurrence of *Dendrolimus sibiricus* Tschetverikov, 1908 (Lepidoptera: Lasiocampidae) instead of *D. superans* Butler, 1877, in South Korea. *Entomol Res.* 2018:48:108–121. https://doi.org/10.1111/1748-5967.12274.

Jones P *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 2014:30:1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

Kageyama D *et al.* Feminizing *Wolbachia* endosymbiont disrupts maternal sex chromosome inheritance in a butterfly species. *Evol Lett.* 2017:1:232–244. https://doi.org/10.1002/evl3.28.

Kang YJ *et al.* CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res*. 2017:45:W12–W16. https://doi.org/10.1093/nar/gkx428.

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 2013:30:772–780. https://doi.org/10.1093/molbev/mst010.

Kernewitz B. Über Spermiogenese bei Lepidopteren. *Zool Anz.* 1914:45:137–139.

Kernewitz B. *Spermiogenese bei Lepidopteren mit besonderer Berücksichtigung der Chromosomen*. Brandstätter; 1915.

Kiuchi T *et al.* A single female-specific piRNA is the primary determiner of sex in the silkworm. *Nature*. 2014:509:633–636. https://doi.org/10.1038/nature13315.

Kononov A *et al.* Genetic diversity among eight *Dendrolimus* species in Eurasia (Lepidoptera: Lasiocampidae) inferred from mitochondrial COI and COII, and nuclear ITS2 markers. *BMC Genet*. 2016:17:173–182. https://doi.org/10.1186/s12863-016-0463-5.

Kristensen N, Scoble MJ, Karsholt O. Lepidoptera phylogeny and systematics: the state of inventorying moth and butterfly diversity. *Zootaxa*. 2007:1668:699–747. https://doi.org/10.11646/zootaxa.1668.1.30.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012:9:357–359. https://doi.org/10.1038/nmeth.1923.

Leclercq S *et al.* Birth of a W sex chromosome by horizontal transfer of *Wolbachia* bacterial symbiont genome. *Proc Natl Acad Sci U S A.* 2016:113:15036–15041. https://doi.org/10.1073/pnas.1608979113.

Lewis JJ *et al.* The *Dryas iulia* genome supports multiple gains of a W chromosome from a B chromosome in butterflies. *Genome Biol Evol.* 2021:13:evab128. https://doi.org/10.1093/gbe/evab128.

Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018:34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Lovell JT *et al.* GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *Elife*. 2022:11:e78526. https://doi.org/10.7554/eLife.78526.

Lukhtanov VA. Sex chromatin and sex chromosome systems in nonditrysian Lepidoptera (Insecta). *J Zool Syst Evol Res.* 2000:38:73–79. https://doi.org/10.1046/j.1439-0469.2000.382130.x.

Lukhtanov VA, Pazhenkova EA. Diversity and evolution of telomeric motifs and telomere DNA organization in insects. *Biol. J. Linn. Soc*. 2023:140:536–555. https://doi.org/10.1093/biolinnean/blad068.

Lukhtanov VA, Vila R, Kandul NP. Rearrangement of the Agrodiaetus dolus species group (Lepidoptera, Lycaenidae) using a new cytological approach and molecular data. *Insect Syst Evol.* 2006:37:325–334. https://doi.org/10.1163/187631206788838563.

Macgregor CJ, Saccheri IJ. The genome sequence of the Rock Grayling, *Hipparchia semele* (Linnaeus, 1758). *Wellcome Open Res*. 2023:8:495. https://doi.org/10.12688/wellcomeopenres.20183.1.

Makino S. On the Chromosomes of *Dendrolimus*-Moths. I. *Proc. Imp. Acad*. 1933a:9:19–22. https://doi.org/10.2183/pjab1912.9.19.

Makino S. On the Chromosomes of *Dendrolimus*-Moths. II. *Proc. Imp. Acad*. 1933b:9:64–67. https://doi.org/10.2183/pjab1912.9.64.

Mapleson D, Venturini L, Kaithakottil G, Swarbreck D. Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. *Gigascience*. 2018:7:giy131. https://doi.org/10.1093/gigascience/giy131.

Martinez J, Klasson L, Welch JJ, Jiggins FM. Life and death of selfish genes: comparative genomics reveals the dynamic evolution of cytoplasmic incompatibility. *Mol Biol Evol.* 2021:38:2–15. https://doi.org/10.1093/molbev/msaa209.

Mc Cartney AM *et al.* Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nat Methods.* 2022:19:687–695. https://doi.org/10.1038/s41592-022-01440-3.

Mikkola K, Ståhls G. Morphological and molecular taxonomy of *Dendrolimus sibiricus* Chetverikov stat.rev. and allied lappet moths (Lepidoptera: Lasiocampidae), with description of a new species. *Entomol. Fenn*. 2008:19:65–85. https://doi.org/10.33338/ef.84417.

Minh BQ *et al.* IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol.* 2020:37:1530–1534. https://doi.org/10.1093/molbev/msaa015.

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B. How many species are there on earth and in the ocean? *PLoS Biol*. 2011:9:e1001127. https://doi.org/10.1371/journal.pbio.1001127.

Moronuki Y, Kasahara R, Naka H, Suzuki MG. Identification and functional analysis of sex-determining genes in the spongy moth, *Lymantria dispar* (lepidoptera: Erebidae). *Insect Biochem Mol Biol.* 2025:177:104219. https://doi.org/10.1016/j.ibmb.2024.104219.

Nichuguti N, Fujiwara H. Essential factors involved in the precise targeting and insertion of telomere-specific non-LTR retrotransposon, SART1Bm. *Sci Rep.* 2020:10:8963. https://doi.org/10.1038/s41598-020-65925-x.

Pazhenkova EA, Lukhtanov VA. Chromosomal conservatism vs chromosomal megaevolution: enigma of karyotypic evolution in Lepidoptera. *Chromosome Res.* 2023:31:16. https://doi.org/10.1007/s10577-023-09725-9.

Pertea M *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol.* 2015:33:290–295. https://doi.org/10.1038/nbt.3122.

Poplin R *et al.* A universal snp and small-indel variant caller using deep neural networks. *Nat Biotechnol.* 2018:36:983. https://doi.org/10.1038/nbt.4235.

Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One*. 2010:5:e9490. https://doi.org/10.1371/journal.pone.0009490.

Robinson JT *et al.* Integrative genomics viewer. *Nat Biotechnol.* 2011:29:24–26. https://doi.org/10.1038/nbt.1754.

Robinson R. *Lepidoptera genetics*. Pergamon Press; 1971.

Rozhkov A. *Sibirskiy shelkopryad*. Izdatelstvo Akademii Nauk SSSR; 1963.

Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014:30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153.

Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. *Bioinformatics*. 2021:37:1639–1643. https://doi.org/10.1093/bioinformatics/btaa1016.

Šichová J *et al.* Fissions, fusions, and translocations shaped the karyotype and multiple sex chromosome constitution of the northeast-Asian wood white butterfly, *Leptidea amurensis*. *Biol. J. Linn. Soc*. 2016:118:457–471. https://doi.org/10.1111/bij.12756.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015:31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

Song L, Sabunciyan S, Yang G, Florea L. A multi-sample approach increases the accuracy of transcript assembly. *Nat Commun.* 2019:10:5000. https://doi.org/10.1038/s41467-019-12990-0.

Traut W, Sahara K, Marec F. Sex chromosomes and sex determination in Lepidoptera. *Sex Dev.* 2007:1:332–346. https://doi.org/10.1159/000111765.

van't Hof AE *et al.* Zygosity-based sex determination in a butterfly drives hypervariability of Masculinizer. *Sci Adv.* 2024:10:6979. https://doi.org/10.1126/sciadv.adj6979.

Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *Gigascience*. 2018:7:giy093. https://doi.org/10.1093/gigascience/giy093.

Voleníková AC *et al.* 2023. Ghost W chromosomes and unique genome architecture in ghost moths of the family Hepialidae [preprint]. bioRxiv 556148. https://doi.org/10.1101/2023.09.03.556148.

Werren JH, Baldo L, Clark ME. *Wolbachia*: master manipulators of invertebrate biology. *Nat Rev Microbiol.* 2008:6:741–751. https://doi.org/10.1038/nrmicro1969.

Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*. 2015:31:3350–3352. https://doi.org/10.1093/bioinformatics/btv383.

Wright CJ, Stevens L, Mackintosh A, Lawniczak M, Blaxter M. Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera. *Nat Ecol Evol.* 2024:8:777–790. https://doi.org/10.1038/s41559-024-02329-4.

Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*. 2005:21:1859–1875. https://doi.org/10.1093/bioinformatics/bti310.

Xiao G. *Forest insects of China*. China Forestry Publishing House; 1992.

Yoshido A, Marec F. Deviations in the Z: a ratio disrupt sexual development in the eri silkmoth, *Samia cynthia ricini*.

Genetics. 2023:224:iyad023. https://doi.org/10.1093/genetics/iyad023.

Yu T, Zhao X, Li G. TransMeta simultaneously assembles multi-sample RNA-seq reads. *Genome Res*. 2022:32:1398–1407. https://doi.org/10.1101/gr.276434.121.

Zhang S *et al.* Chromosome-level genome assembly of an important pine defoliator, *Dendrolimus punctatus* (Lepidoptera; Lasiocampidae). *Mol Ecol Resour.* 2020:20:1023–1037. https://doi.org/10.1111/1755-0998.13169.

Zhao Q-S *et al.* Hybridization experiments with two species of *Dendrolimus*. *Acta Entomol. Sin*. 1992:35:28–32.

Zhou J *et al.* Chromosome-level genome assembly reveals significant gene expansion in the toll and IMD signaling pathways of *Dendrolimus kikuchii*. *Front Genet.* 2021:12:728418. https://doi.org/10.3389/fgene.2021.728418.

Zolotuhin VV. *Lapper Moths (Lepidoptera: Lasiocampidae) of Russia and adjacent territories*. Ulyanovsk Korporatsiya teckhnologiy prodvizheniya; 2015.