Repeated origin of the W chromosome from the Z chromosome in Lepidoptera

Wenting Dai[1], Judith E. Mank[2], Liping Ban[1, *]

[1] Department of Grassland Resources and Ecology, College of Grassland Science and Technology, China Agricultural University, Beijing, China

[2] Department of Zoology and Biodiversity Research Centre, University of British Columbia, Vancouver, Canada

[*] To whom correspondence should be addressed: liping_ban@163.com (LB).

Abstract

We built a chromosome-level genome assembly of the Asian corn borer, *Ostrinia furnacalis* Guenée (Lepidoptera: Pyralidae, Pyraloidea), an economically important pest in corn, from a female, including both the Z and W chromosome. Despite deep conservation of the Z chromosome across Lepidoptera, our chromosome-level W assembly reveals this chromosome has recently derived from the *O. furnicalis* Z chromosome. We also recover evidence for convergent formation of the W in two other related families, all independently derived from the Z chromosome. Despite its recent origin, the W chromosome has accumulated significant repetitive elements and experienced rapid pseudogenization, illustrating that degeneration can occur rapidly following recombination suppression. Our results, taken together, reveal remarkable dynamics for the sex chromosomes in this otherwise highly-conserved clade.

Introduction

The Asian corn borer *Ostrinia furnacalis* Guenée (Lepidoptera: Pyralidae, Pyraloidea), is a major insect pest of corn, widespread in the Asian-Western Pacific region. Like other Lepidoptera insects, it has a female heterogametic sex chromosome **s**ystem, with ZW females and ZZ males[1,2]. The conservation of the Z chromosome has been well established in Lepidoptera[3], however, the W chromosome in Lepidoptera is unusual in that it was recruited into the genome well after the Z chromosome[4] , as the basal lineages in the clade are Z0/ZZ. Available evidence suggests that, at least in some lineages, the W chromosome bears no homology to the Z[3,5,6], and may actually have originated from a B chromosome[3]. Within the Lepidoptera, at the base of the Ditrysia group, complex sex chromosomes, including neo-W chromosomes are observed in many clades based on cytogenetic analysis[4]. However, a detailed genomic understanding of origin and evolution of the Lepidopteran W chromosome, has remained elusive due to the difficulties of assembling sex-limited chromosomes.

Recently, third-generation sequencing advances have permitted partial or chromosome-level assemblies of the W chromosome in a limited number of Lepidoptera, including *Cydia pomonella* (Torticidae, Tortricoideae)[6], *Cnaphalocrocis medinalis*[7] (Crambidae, Pyraloidea), *Spodoptera exigua*[8], *Spodoptera frugiperda*[9] (Noctuidea, Noctuoidea), *Trichoplusia ni*[10] (Noctuidae, Noctuoidea) and *Dryas iulia*[5] (Nymphalidae, Papilionoidea). This work has revealed no detectible homology between the Z and W in *C. pomonella*[6], *T. ni*[11] and *D. iulia*[5]. All this is consistent with a non-canonical origin of the W chromosome in Lepidoptera, where the W has been recruited from a B chromosome and it therefore not homologous to the deeply conserved corresponding Z chromosome.

Consistent with the cytogenetic evidence within the Ditrysia of highly variable W chromosome karyotype morphologies[4,12], substantial synteny between the W chromosomes of *S. exigua* and *C. medinalis*[7] clades has been noted[12]. Whether these nascent W chromosomes share a common origin from the Z chromosome, is not known. To answer questions about the conservation of the W, as well as the Z

chromosome, we built a chromosome-level genome of *O. furnacalis* though the combination of PacBio HiFi (CCS) sequencing and Hi-C. We couple this with extensive RNA-Seq analysis of multiple developmental stages and tissues in both sexes to reveal that although the Z chromosome of the species shares a highly conserved synteny with the Z in other Lepidopterans, the W has formed recurrently across the lineage from the Z chromosome, most notably in the recent ancestor of *O. furnicalis*. Despite this recent origin, the *O. furnicalis* W has accumulated significant repetitive elements and pseudogenes, suggesting that degeneration of the sex-limited chromosome can occur rapidly following recombination suppression.

Methods

Samples

O. furnacalis larvae were collected from a corn field at Beijing, China, in July 2020, and fed with an artificial diet in the laboratory of China Agricultural University. The incubator environment was set at 26 ± 1℃ and 50 ± 5% relative humidity on a photoperiod (Light: Dark = 16:8).

Genome sequencing and assembly

Genomic DNA was extracted from a female laboratory-fed pupa with the SDS method. From this single individual, we obtained 32.96 Gb circular consensus sequencing (CCS) data from one cell of the PacBio Sequel II platform at Biomarker Co., Ltd., Qingdao, China (Table S1). High accuracy PacBio Hifi data were assembled using hifiasm (v0.12) software to obtain a draft genome[13]. We used purge_haplotigs[14] to remove redundant contigs and generate a non-redundant assembled genome.

Hi-C scaffolding

We constructed Hi-C libraries (300-700 bp insert size), following Rao et al.[15], and sequenced them with pair-end 150 Illumina Hiseq at Biomarker Co., Ltd., Qingdao, China. 53.05Gb of clean data were produced after filtering adapter sequences, primer

82 sequences and low-quality data (Table S2). The resulting trimmed reads were aligned
83 to the draft assembly with BWA (v0.7.10-r789), retaining only uniquely aligned read
84 pairs with mapping quality >20 for further analysis. Invalid read pairs, including
85 dangling-end, self-cycle, re-ligation and dumped products, were removed by HiC-Pro
86 (v2.8.1)[16].

87 The 56.25% of unique mapped read pairs represent valid interaction pairs and were
88 used for correction, clustering, ordering and orientation of scaffolds into
89 chromosomes with LACHESIS[17]. Before chromosome assembly, we performed a
90 preassembly for error correction of contigs which required the splitting of contigs into
91 segments of 50kb on average. Then Hi-C data were mapped to these segments using
92 BWA (v0.7.10-r789). The uniquely mapped data were retained for the assembly, and
93 any two segments which showed inconsistent connection with information from the
94 raw scaffolds were checked manually. Parameters for running LACHESIS included:
95 CLUSTER_MIN_RE_SITES = 27; CLUSTER_MAX_LINK_DENSITY = 2;
96 ORDER_MIN_N_RES_IN_TRUNK = 15; ORDER_MIN_N_RES_IN_SHREDS = 15. After this
97 step, placement and orientation errors exhibiting obvious discrete chromatin
98 interaction patterns were manually adjusted. Finally, we constructed a heatmap based
99 on the interaction signals of valid mapped read pairs between chromosomes.

100 ***Genome annotation***

101 Transposable elements (TEs) were identified by a combination of homology-based and
102 de novo approaches. We first carried out a de novo repeat prediction using
103 RepeatModeler2 (v2.0.1)[18], which implements RECON (v1.0.8)[19] and RepeatScout
104 (v1.0.6)[20]. Then the predicted results were classified using RepeatClassifier[18] by means
105 of repbase (v19.06)[21], REXdb (v3.0)[22] and Dfam (v3.2)[23]. Secondly, we performed a de
106 novo prediction for long terminal repeats (LTRs) using LTR_retriever (v2.8)[24] via
107 LTRharvest (v1.5.9)[25] and LTR_finder (v2.8)[26]. A non-redundant species-specific TE
108 library was constructed by combining the de novo TE library above with the known
109 databases. Final TE sequences in the *O. furnacalis* genome were identified and
110 classified by homology search against the library using RepeatMasker (v4.10)[27].

111 Tandem repeats were annotated by Tandem Repeats Finder (TRF v409)[28] and
112 MIcroSAtellite identification tool (MISA v2.1)[29].

113 We integrated de novo prediction, homology search, and transcript-based assembly to
114 annotate protein-coding genes. The de novo gene models were predicted using
115 Augustus (v2.4)[30] and SNAP (2006-07-28)[31], both ab initio gene-prediction approaches.
116 For the homolog-based approach, we used GeMoMa (v1.7)[32] with reference gene
117 models from *Drosophila melanogaster*, *Bombyx mori*, *Chilo suppressalis* and *Galler*
118 *mellonella*. For the transcript-based prediction, RNA-sequencing data were mapped to
119 the reference genome using Hisat (v2.0.4)[33] and assembled with Stringtie (v1.2.3)[34].
120 GeneMarkS-T (v5.1)[35] was used to predict genes based on the assembled transcripts.
121 PASA (v2.0.2)[36] was used to predict genes based on the unigenes assembled by Trinity
122 (v2.11)[37]. Gene models from these different approaches were combined with EVM
123 (v1.1.1)[38] and updated by PASA (v2.0.2)[36]. The final gene models were annotated by
124 searching the GenBank Non-Redundant (NR, 20200921), TrEMBL (202005), Pfam
125 (v33.1)[39], SwissProt (202005)[40], eukaryotic orthologous groups (KOG, 20110125), gene
126 ontology (GO, 20200615) and Kyoto Encyclopedia of Genes and Genomes (KEGG,
127 20191220) databases.

128 We used tRNAscan-SE (v1.3.1)[41] with default parameters to identify transfer RNAs
129 (tRNAs), and barrnap (v0.9) with default parameters to identify the ribosomal RNAs
130 (rRNAs) based on Rfam (v12.0)[42]. miRNAs were identified with miRbase[43]. Small
131 nucleolar RNA (snoRNAs) and small nuclear RNA (snRNAs) were identified with
132 Infernal (v1.1.1)[44], using the Rfam (v12.0) database[42].

133 Pseudogenes have similar sequences to functional genes, but may have lost their
134 biological function because of mutations, such as indels. GenBlastA (v1.0.4)[45] was used
135 to scan the whole genome after masking predicted functional genes. Putative
136 candidates were then analyzed by searching for non-mature mutations and frame-shift
137 mutations using GeneWise (v2.4.1)[46].

138 ***Comparative genomics and phylogenetic reconstruction***

139 Protein sequence alignments between *O. furnacalis* and seven other Lepidopteran
140 species (*Bombyx mori*, *Chilo suppressalis*, *Cydia pomonella*, *Cnaphalocrocis medinalis,*
141 *Spodoptera exigua, Spodoptera fugiperda,* and *Trichoplusia ni*) were performed with
142 diamond (-e<1e-5), then alignment results were analyzed and the homologous
143 chromosomal regions were identified with MCScanX (MATCH_SCORE: 50, MATCH_SIZE:
144 5, GAP_PENALTY: -1, OVERLAP_WINDOW: 5, E_VALUE: 1e-05, MAX GAPS: 15). The
145 synteny relationships among chromosomes were displayed using circos (v0.69–9).

146 We used the protein sequences of *Ostrinia furnacalis and* ten other insect species
147 (*Spodoptera frugiperda, Spodoptera litura, Bombyx mori, Chilo suppressalis, Danaus*
148 *plexippus, Melitaea cinxia, Papilio Xuthus, Cydia pomonella*, *Plutella xylostella* and
149 *Drosophila melanogaster*) for phylogenetic analysis, keeping only the longest
150 transcript of each gene for analysis, and using OrthoFinder (v2.5.4)[47] with default
151 settings to identify orthologues and homologues. To infer the phylogeny of these
152 insects, multiple sequence alignments of single-copy orthologs were performed using
153 MAFFT (v7.471)[48]. Then we extracted conserved sequences with gblocks (v0.91b)[49]
154 and concatenated them to a single sequence alignment. The resulting sequence
155 alignment was used to construct a maximum likelihood phylogenetic tree using IQ-
156 TREE (v1.6.12)[50] (outgroup: *Drosophila melanogaster*). Divergence times between
157 various species were estimated by MCMCtree in PAML (v4.9j)[51]. Three standard
158 divergence time points from the TimeTree database (http://timetree.org/) were used
159 for calibration: *Spodoptera frugiper* - Spodoptera *litura*, 100-122 million years ago
160 (Mya), *Danaus plexippus* - *Melitaea cinxia*, 76-102 Mya, (*Danaus plexippus*, *Melitaea*
161 *cinxia*) - *Papilio xuthus*, 76-146 Mya. The tree was visualized using FigTree (v1.4)
162 (http://tree.bio.ed.ac.uk/software/figtree/).

163 ***Detection of sex chromosomes***

164 In order to identify the sex chromosomes, we performed genome resequencing with
165 five female and male pupae. High-quality clean data (24.95-28.80-fold coverage of the

genome) were obtained through the pair-end 150 Illumina Hiseq platform at Biomarker Co., Ltd., Qingdao, China (Table S3). The clean data were aligned to reference genome with Bwa-men (v0.7.17). We compared the coverage differences between male and female samples[52] to distinguish the Z, W and autosomes. Specifically, we used the genomecov and groupby in Bedtools (v2.30.0) to obtain a per-base median coverage depth for each chromosome and normalized them by the mean of all chromosome median coverages for each sample. Normalized coverage depth was averaged by sex to produce a coverage depth per chromosome for each sex. Then we compared coverage depth between sexes for each chromosome and calculated $\log_2$ male:female (M:F) coverage ratio [$\log_2$(M:F coverage)]. Autosomes have an equal coverage between sexes [$\log_2$ (M:F coverage) = 0], while the Z chromosome should show approximately two-fold greater coverage in the males [$\log_2$(M:F coverage) = 1]. The W chromosome should show a strong female-biased coverage pattern. We also calculated the M:F median coverage ratio along each chromosome with nonoverlapping 1000 bp windows.

Z-W synteny and phylogenetic analysis

Genome DNA sequences of Z and W are aligned using MUMmer (v4.0.0rc1)[53] (nucmer –mum), keeping only homologous DNA sequences >2000bp (delta-filter -1 -l 2000 -o 0). The results of sequence alignments are used to display the Z-W synteny with RectChr (v1.34) (https://github.com/BGI-shenzhen/RectChr).

Predicted protein sequences were used to detect reciprocal best hits between Z- and W-linked genes using getRBH.pl, for *O. furnacalis*, *C. pomonella*, *C. medinalis*, *S. exigua* and *S. fugiperda*, which were considered as Z-W gametologs for each species. The same method was used to identify the Z-Z orthologs between *O. furnacalis* and the other four Lepidopteran species. Based on Z-Z homology among species, we gained 10 groups of orthologous Z-W gametologs, the coding sequence of which was used for constructing maximum-likelihood gene trees. Specifically, for each group of Z-W gametologs, multiple coding sequences from different insect species were aligned using MAFFT (v7.471)[48]. Conserved sequences were then extracted with gblocks

195 (v0.91b)[49] and used to construct a maximum likelihood phylogenetic tree using IQ-
196 TREE (v1.6.12)[50] , with Z chromosome outgroups *D. plexippus, B. mori or T. ni*.

197 **Results**

198 ***Chromosome-level genome assembly of the Asian Corn Borer***

199 Our chromosome assembly strategy employed PacBio HiFi (CCS) sequencing data to
200 assemble the draft genome and Hi–C data to detect chromosomal contact information.
201 These PacBio long-reads were self-corrected using Quiver and assembled into a draft
202 genome assembly with a total length of 493.10Mb, consisting of 57 contigs with an
203 N50 length of 15.69Mb (Table S4). The assembled genome size is similar to that
204 obtained from genome surveys (Supplementary Figure 1). Next, Hi–C linking
205 information further anchored, ordered, and oriented 46 contigs to 32 chromosomes
206 (30 autosomes, with Z and W sex chromosomes), which contained 86.79% of the
207 whole genome assembly (Table S4, Supplementary Figure 2). The chromosome-level
208 genome assembly was 492.57Mb with a scaffold N50 of 16.47Mb (Table S4).

209 We evaluated the completeness of *O. furnacalis* genome assembly with the Metazoan
210 Benchmark of Universal Single-Copy Orthologs (BUSCO v4) and CEGMA core genes
211 (CEGMA v2.5) datasets. Our assembly contained 97.69% of BUSCO genes, of which
212 97.06% were single copy and 0.63% were duplicates, and 96.07% of CEGMA genes, of
213 which 79.03% were highly conserved (Table S5). To further evaluate the genome
214 assembly quality, genomic DNA sequencing data from Illumina HiSeq were mapped to
215 the assembly scaffolds, with a 99.07% coverage rate (Table S5). Finally, Lepidopteran
216 genomes typically exhibit very high levels of synteny. Whole-genome alignment of the
217 *O. furnacalis* assembly to the chromosomes of the *Spodoptera litura* revealed that
218 chromosomal linkage and ordering of genes are highly conserved (Supplementary
219 Figure 3). All these analyses indicated that the genome assembly is both highly reliable
220 and complete for subsequent analyses.

221 ***Genome annotation***

222 In total, 860,391 repeat sequences spanning 204.4Mb were identified, constituting

41.45% of the *O. furnacalis* genome (Table S4, Figure 1a). We integrated the result of *ab initio*, homology-based and RNA-seq methods to annotate protein-coding genes, most of the which came from homology-based and RNA-seq methods, indicating the high-quality of annotation (Supplementary Figure 4). Finally, we obtained 16,509 protein coding genes in the *O. furnacalis* genome, which has similar gene features to other Lepidopteran genomes (gene length, gene number, coding length, intron length of each gene, exon length) (Supplementary Figure 4). Next, we identified different types of noncoding RNAs (ncRNAs), including 7,710 transfer RNAs (tRNA), 73 ribosomal RNAs (rRNA), and 39 microRNAs (miRNA) (Table S6). In addition, we annotated 171 pseudogenes in the *O. furnacalis* genome, most of which were located in the chromosome LG3 (Figure 1c, Table S6), the W chromosome (see below). We performed a functional annotation for all predicted protein coding genes using NCBI nonredundant, EggNOG, GO, KEGG, SWISS-PROT and Pfam databases, with 98.21% of our predicted genes assigned to at least one of these databases (Table S6).

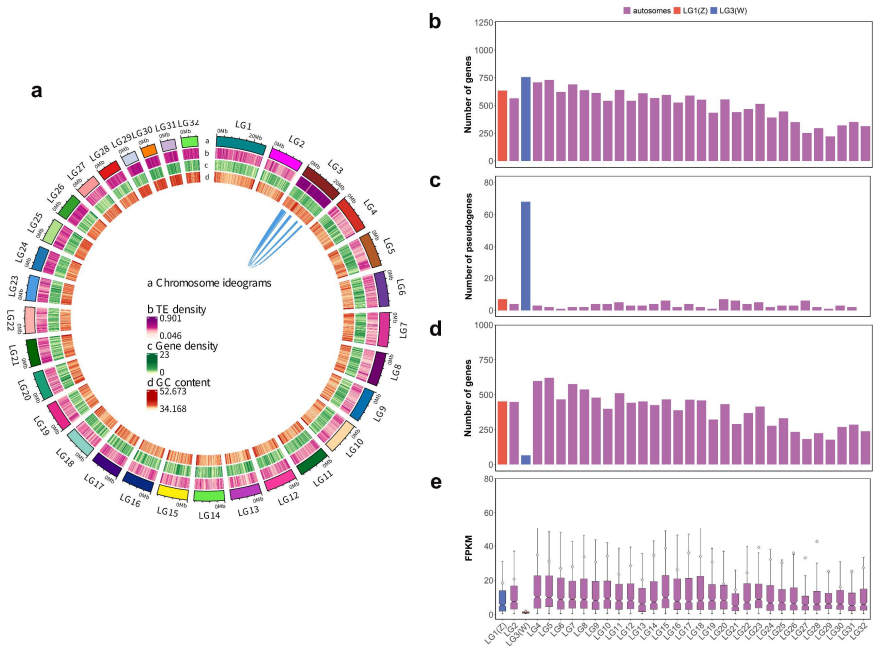

Fig 1. Genomic characterization of *Ostrinia furnacalis*. a. Circular diagram depicting the genomic landscape of the 32 chromosomes (Chr1–Chr32 on a Mb scale). The denotation of each track is listed in the center of the circle. Blue lines in LG3 (the W chromosome) represent collinearity within the W, due to repeat elements. b. Number and distribution of protein-coding genes. c. Number and distribution of pseudogenes. d. Number of genes with FPKM > 0.5. e. Average chromosomal expression level.

We used OrthoFinder to find orthologous genes among *O. furnacalis* and 10 other insect species (see methods). A total of 16231 orthologous groups with 1163 single-copy orthologous genes were identified. We inferred a phylogenic tree and divergence time estimate using the end-to-end concatenated amino acid sequences of 1163 single-copy orthologous genes. The Crambidae lineage to which *O. furnacalis* belongs was estimated to diverge from Obtectomeran Lepidoptera approximately 77.3 Mya ago (Figure 2a).

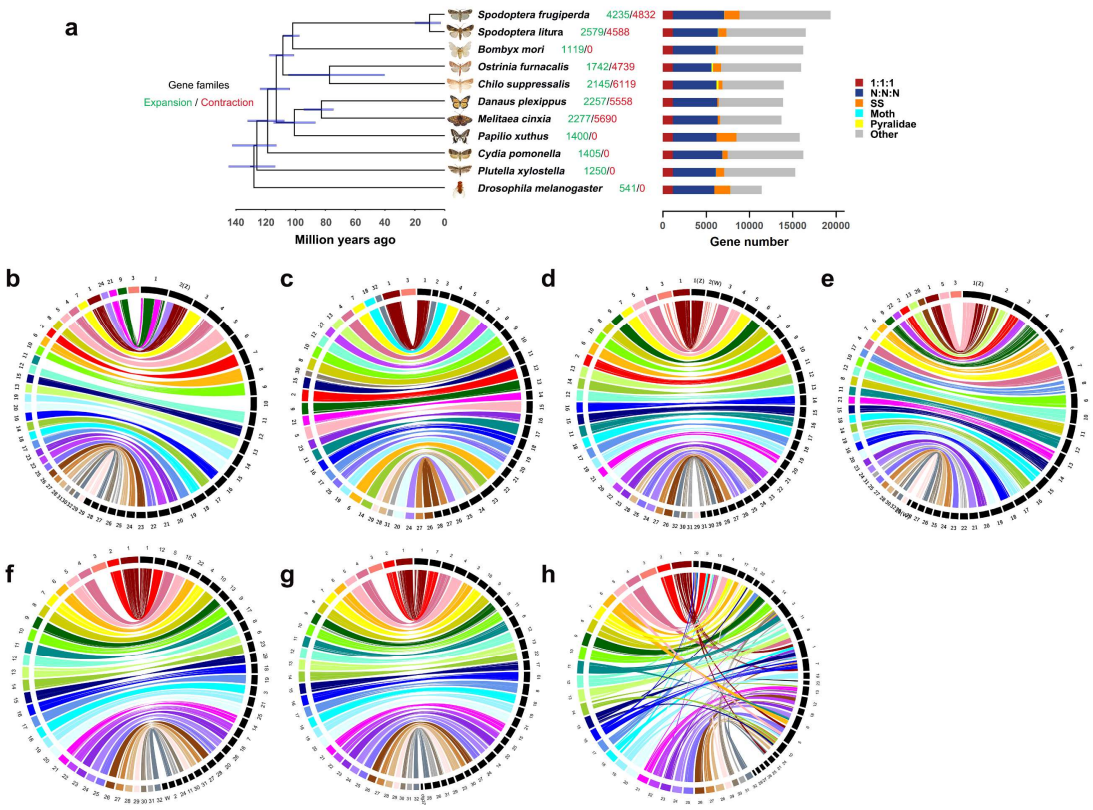

Fig 2. a. Phylogenetic tree and gene orthology of *O. furnacalis* with nine Lepidopteran genomes and *D. melanogaster*. Comparative analysis of synteny between *O. furnacalis* and **b.** *Chilo suppressalis*, **c.** *Bombyx mori*, **d.** *Cnaphalocrocis medinalis*, **e.** *Cydia pomonella,* **f.** *Spodoptera exigua,* **g.** *Spodoptera fugiperda*, and **h.** *Trichoplusia ni*. The chromosomes of *O. furnacalis* are shown in the left, and the other insects' chromosomes are shown in the right. The *O. furnacalis* W chromosome is LG3, and the Z is LG1 (b-h).

261 **Synteny, karyotype evolution, and sex chromosomes**

262 We compared the syntenic relationships between *O. furnacalis* and other
263 Lepidopterans, including *Bombyx mori, Chilo suppressalis, Cnaphalocrocis medinalis*,
264 *Spodoptera litura, Spodoptera exigua, Spodoptera fugiperda, and Trichoplusia ni* (Fig
265 2b-h). In general, the *O. furnacalis* genome shows a high level of synteny with other
266 Lepidopteran genomes, though with some fusion and fission events. *O. furnacalis* LG1
267 is syntenic with the Z chromosome in all other species, suggesting it is the Z
268 chromosome in *O. furnicalis* as well, and this is also evident from M:F coverage analysis
269 (Figure 3a), which reveals two-fold greater coverage in males (ZZ) compared to females
270 (ZW).

271 *O. furnacalis* LG3, the W chromosome, exhibits female-biased coverage consistent
272 with a female-limited chromosome (Figure 3a). The M:F coverage is highly variable
273 across chromosome LG3 compared to all other chromosomes (Figure 3b), due in large
274 part to the abundance of TEs on this chromosome (Supplementary Figure 5,
275 Supplementary Figure 6). The W chromosome (LG3) is the second largest chromosome
276 (22.23 Mb) in our assembly (Figure 1a), with the largest number of predicted protein
277 coding genes compared with other chromosomes (Figure 1b), a large proportion of
278 which are pseudogenized (18.43%) (Figure 1c). Indeed, this chromosome contains 39.8%
279 of all pseudogenes annotated in the genome. Consistent with high levels of
280 pseudogenization, relatively few predicted W genes are expressed (Figure 1d) and
281 those that are expressed show low levels of mRNA (Figure 1e).

282 Furthermore, LG3 has a notably greater repeat density and significantly different
283 repeat composition compared to other chromosomes (Supplementary Figure 5,
284 Supplementary Figure 6) with a larger proportion of satellites, DIRS, LINE, Copia and
285 Gypsy, and an enrichment of *Maverick* elements.

286

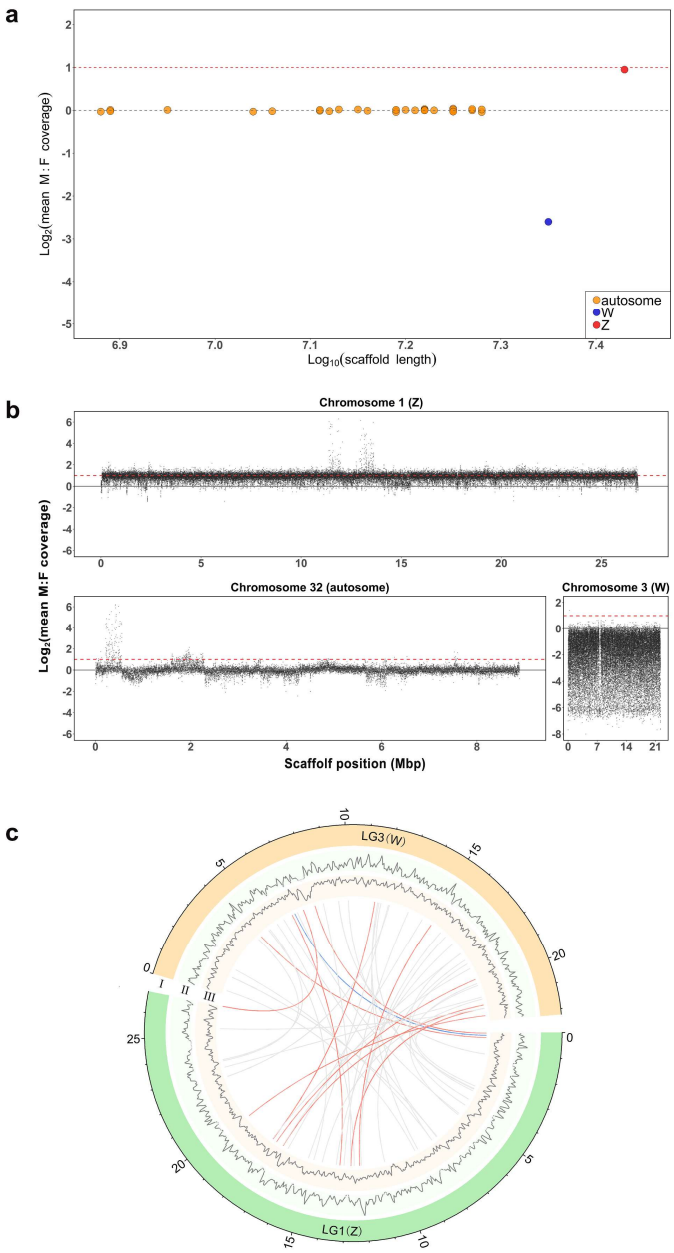

Fig 3. Detection of sex chromosomes and ZW synteny. a. Male:female coverage ratios for each chromosome, plotted by chromosome length. Each point represents a single chromosome. The dashed gray line is the theoretical expectation for autosomes and the dashed red line shows the expectation for the Z chromosome. **b.** Male:female coverage ratios plotted in 500bp windows across scaffolds for Z, W, and a representative autosome (LG32). **c.** A synteny circular plots between the Z and W. From the outside to the center, (I) represents the Z and W chromosome of *Ostrinia furnacalis*, and the outermost scale line shows the physical position in chromosome (Mb), (II) depicts the gene content (0-12) in non-overlapping windows of 100 kb, and (III) indicates the repeat density (0-1) in non-overlapping windows of 100 kb. Links between genes were computed from the reciprocal best BLASTP hits and are color coded relative to the BLASTP alignment percent identity, with percent identity > 80% in blue and > 90% in red.

301 ***Z-W homology***

302 Given that previous reports have found no evidence of homology between the Z and
303 W in some basal Lepidopteran species[5,6,11], we next investigated the relationship
304 between genes on the *O. furnacalis* Z and W chromosomes. Contrary to expectation,
305 we detected significant gene homology (Figure 3c) between the Z and W. Using our W
306 gene set, we further examined homology between the *O. furnacalis* W chromosome
307 and the W chromosome from other Lepidopterans. The *O. furnacalis* W shows
308 substantial homology with the W in *Cnaphalocrocis medinalis*, consistent with some
309 form of shared ancestry within the family Crambidae. However, we observed no
310 discernable homology between the *O. furnacalis* W and the W chromosome of any of
311 the more distantly related Lepidopterans that we queried (Fig 2 b-h).

312 We next investigated the phylogenetic relationships of these genes, as well as
313 homologs identified from related species (Fig 4). *O. furnicalis* Z-W orthologs
314 distributed across the entirely of the Z chromosome clustered most closely with the
315 corresponding W chromosome, consistent with a recent formation of the W from the
316 Z chromosome. Our gene trees suggest additional independent W chromosomes
317 originating from the Z chromosome in at least three families: Crambidae
318 (*Cnaphalocrocis*, Pyraloidea), Noctuidae (*Spodoptera*, Noctuloidea) and Pyralidae
319 (*Ostrinia*, Pyraloidea).

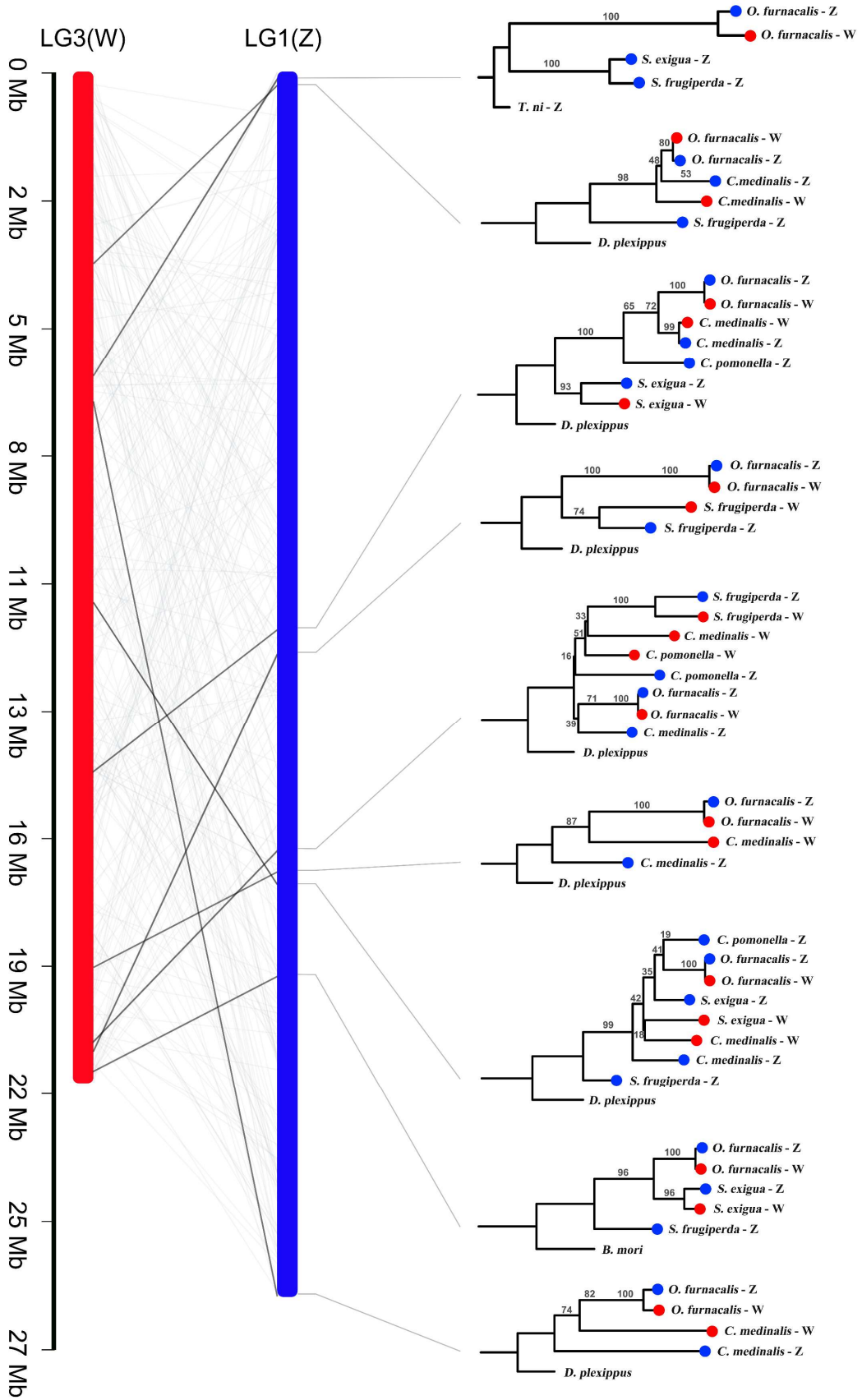

Fig 4. Maximum-likelihood gene trees for *O. furnacalis, S. exigua, S. frugiperda, C.*

322 *pomonella* and *C. medinalis* orthologous ZW gametologs (outgroup: *D.*
323 *plexippus/B.mori/T. ni*), with physical position and synteny (in grey) between the Z and
324 W chromosomes. Black lines represent position on Z and W chromosomes.

325 **Discussion**

326 Here we report a high quality, chromosome-level genome for *O. furnicalis*. Our
327 reference genome includes a single, contiguous sequence for the female-specific W
328 chromosome, allowing us to query the evolutionary history of W chromosome coding
329 content as well as determine Z-W synteny. Our results are consistent with a recent and
330 repeated origin of the neo-W chromosome from the Z in several distinct lineages.

331

332 **Composition and transcription of the W chromosome**

333 W and Y chromosomes are often enriched with repeat sequences[54,55], resulting in
334 significantly higher repeat content compared to other chromosomes[6,7,55]. Consistent
335 with this, we observe enrichment of repetitive elements on the W chromosome in *O.*
336 *furnicalis* (Supplementary Figure 6), with the number of *Maverick* elements
337 particularly higher on the W chromosome compared to the rest of the genome
338 (Supplementary Figure 5). *Maverick*, also known as *Polinton*, is a DNA transposon
339 widespread in eukaryotes[56], and one of the most striking characteristics of *Maverick* is
340 the large size, typically 15–20 kb[56].

341 In addition, we found that the number of protein-coding genes of W chromosome is
342 indeed the largest in the whole genome (Figure 1a). However, the transcription activity
343 of W-linked genes is particularly low (Figure 1d), consistent with previous reports that
344 the W is transcriptionally inactive in somatic tissues[2].

345 **Conservation of sex chromosomes across Lepidoptera species**

346 We searched orthologs to Z-linked genes of *O. furnacalis* in other Lepidopterans, and
347 found that although the Z chromosome shows clear strong conservation[4], the *O.*
348 *furnacalis* W chromosome is only conserved with *C. medinalis*, also a member of
349 Crambidae. This is consistent with previous work suggesting that the composition of
350 the W varies even among species with the same family[57].

351 **Evolutionary history of the sex chromosomes in *O. furnacalis***

352 Karyotype work has revealed a complex history of the W chromosome, including the
353 repeated origin of neo-W chromosomes in many Lepidopteran lineages[4,58-60]. The
354 evolutionary history of the W chromosome in Lepidoptera differs from the canonical
355 model of sex chromosome formation in that it was recruited, possibly from a B
356 element, after the formation of the Z chromosome[4], at the common root of Ditrysia
357 and Tischerioidea[2,61]. Lineages ancestral to this recruitment exhibit Z0:ZZ karyotypes[4].
358 The B-chromosome origin of the W is supported by the fact that in some basal lineages,
359 W chromosome bears no homology to the Z[3,5,6].

360 Subsequent to this, neo-sex chromosomes have formed in several lineages within the
361 Ditrysia[4], and the evolutionary history of these neo-sex chromosomes, as well as the
362 W chromosome itself, remains unclear, with multiple plausible scenarios[59]. Although
363 we cannot rule out that the Z-derived portion of the W chromosome in *O. furnicalis*
364 represents neo-W chromosomes instead of turnovers, our assembly of the W
365 chromosome reveals a single, contiguous molecule (Fig 4), which is not consistent with
366 a physically separated neo-W that co-segregates with a persisting B chromosome. We
367 also observe Z-W synteny across the entirety of the assembled *O. furnicalis* W
368 chromosome (Figure 3c, Figure 4), which is not consistent with a neo-W. Regardless of
369 whether it is a W or neo-W chromosome, our results indicate that peudogenization
370 and repetitive elements accumulate on the W very quickly following recombination
371 arrest.

372 Importantly, in Lepidopterans, recombination is confined to males[12], which means that
373 any element that pairs with the Z chromosome in females during meiosis will remain
374 female-limited in its transmission, and also completely cease to recombine. Our
375 phylogenetic results (Figure 4) suggest that female-limited chromosomes have
376 originated several times independently from the Z chromosome, occurring in at least
377 the Crambidae (*Cnaphalocrocis*, Pyraloidea), Noctuidae (*Spodoptera*, Noctuloidea)
378 and Pyralidae (*Ostrinia*, Pyraloidea).

379 The lack of recombination in females also means that any female-limited chromosome
380 will immediately experience the full degenerative effects of a non-recombining
381 region[62,63], and we do not expect strata of different ages. The rapid accumulation of
382 pseudogenes and repetitive elements on the W chromosome in *O. furnacalis* is
383 analogous to the rapid degeneration of neo-Y chromosomes in *Drosophila* species,
384 which experience male achiasmy[64]. Consistent with this, we observe high levels of
385 pseudogenization and repeat elements (Figure 1) on the W chromosome, despite its
386 recent origin.

387 In addition to female achiasmy, Lepidoptera exhibit a complex form of complete Z
388 chromosome dosage compensation[65,66], and recent theory has suggested that the
389 evolution of dosage compensation may facilitate the rapid degeneration of the sex-
390 limited chromosome[66]. This is consistent with recent reports of rapid degeneration of
391 Y chromosomes following the evolution of complete sex chromosome dosage
392 compensation[67,68], and the data may therefore suggest that the best predictor of the
393 degree of Y or W degeneration is the degree of dosage compensation for the X or Z.

394

395 **Authors' contributions**

396 LB conceived and designed the research, and wrote the manuscript. WD performed
397 the experiments, analyzed and interpreted the data, and wrote the manuscript. JM
398 analyzed the data and wrote the manuscript. All authors read and approved the final
399 manuscript.

400

401 **Acknowledgments**

402 This work was supported by Beijing Agriculture Innovation Consortium (BAIC02-2022).
403 JEM gratefully acknowledges support from NSERC and a Canada 150 Research Chair.

404

405 **Competing interests**

The authors declare that they have no competing interests.

408 **References**

409 1. Blackmon, H., Ross, L. & Bachtrog, D. Sex determination, sex chromosomes, and karyotype
410 evolution in insects. *Journal of Heredity* **108**, 78-93 (2016).
411 2. Traut, W. & Marec, F. Sex chromatin in Lepidoptera. *The Quarterly Review of Biology* **2**, 239-259
412 (1996).
413 3. Fraïsse, C., Picard, M.A.L. & Vicoso, B. The deep conservation of the Lepidoptera Z chromosome
414 suggests a non-canonical origin of the W. *Nature Communications* **8**(2017).
415 4. Sahara, K., Yoshido, A. & Traut, W. Sex chromosome evolution in moths and butterflies.
416 *Chromosome Research* **20**, 83-94 (2012).
417 5. Lewis, J.J. et al. The *Dryas iulia* genome supports multiple gains of a W chromosome from a B
418 chromosome in butterflies. *Genome Biology and Evolution* **13**(2021).
419 6. Wan, F. et al. A chromosome-level genome assembly of *Cydia pomonella* provides insights into
420 chemical ecology and insecticide resistance. *Nature Communications* **10**(2019).
421 7. Zhao, X. et al. A chromosome−level genome assembly of rice leaffolder, *Cnaphalocrocis medinalis*.
422 *Molecular Ecology Resources* **21**, 561-572 (2021).
423 8. Zhang, F., Zhang, J., Yang, Y. & Wu, Y. A chromosome-level genome assembly for the beet
424 armyworm (*Spodoptera exigua*) using PacBio and Hi-C sequencing. *bioRxiv 2019* (2020).
425 9. Xiao, H. et al. The genetic adaptations of fall armyworm *Spodoptera frugiperda* facilitated its rapid
426 global dispersal and invasion. *Molecular Ecology Resources* **20**, 1050-1068 (2020).
427 10. Chen, W. et al. A high−quality chromosome−level genome assembly of a generalist herbivore,
428 *Trichoplusia ni*. *Molecular Ecology Resources* **19**, 485-496 (2019).
429 11. Fu, Y. et al. The genome of the Hi5 germ cell line from *Trichoplusia ni*, an agricultural pest and
430 novel model for small RNA biology. *eLife* **7**(2018).
431 12. MarecKen, F., Sahara, S. & Traut, W. Rise and fall of the W chromosome in Lepidoptera. in
432 *Molecular biology and genetics of the Lepidoptera* (eds. MR, G. & F, M.) 49 – 63 (CRC, Boca Raton,
433 2010).
434 13. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly
435 using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170-175 (2021).
436 14. Roach, M.J., Schmidt, S.A. & Borneman, A.R. Purge Haplotigs: allelic contig reassignment for
437 third-gen diploid genome assemblies. *BMC Bioinformatics* **19**(2018).
438 15. Rao, S.S.P. et al. A 3D map of the human genome at kilobase resolution reveals principles of
439 Cchromatin looping. *Cell* **162**, 687-688 (2015).
440 16. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome*
441 *Biology* **16**(2015).
442 17. Burton, J.N. et al. Chromosome-scale scaffolding of de novo genome assemblies based on
443 chromatin interactions. *Nature Biotechnology* **31**, 1119-1125 (2013).
444 18. Flynn, J.M. et al. RepeatModeler2 for automated genomic discovery of transposable element
445 families. *PNAS* **117**, 9451-9457 (2020).
446 19. Bao, Z. & Eddy, S.R. Automated de novo identification of repeat sequence families in sequenced
447 genomes. *Genome Res* **12**, 1269-76 (2002).
448 20. Price, A.L., Jones, N.C. & Pevzner, P.A. De novo identification of repeat families in large genomes.
449 *Bioinformatics* **21**, i351-i358 (2005).
450 21. J, J. et al. Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome*

Research **110**, 462-467 (2005).
22. Neumann, P., Novák, P., Hoštáková, N. & Macas, J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mobile DNA* **10**(2019).
23. Wheeler, T.J. et al. Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Research* **41**, D70-D82 (2012).
24. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiology* **176**, 1410-1422 (2018).
25. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
26. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**, W265-W268 (2007).
27. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform* **25**, 4.10.1-4.10.14 (2009).
28. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580 (1999).
29. Beier, S., Thiel, T., Munch, T., Scholz, U. & Mascher, M. MISA-web: a web server for microsatellite prediction. *Bioinformatics* **33**, 2583-2585 (2017).
30. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637-644 (2008).
31. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
32. Keilwagen, J. et al. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research* **44**, e89-e89 (2016).
33. Kim, D., Langmead, B. & Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**, 357-360 (2015).
34. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**, 290-295 (2015).
35. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Research* **43**, e78-e78 (2015).
36. Haas, B.J. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**, 5654-5666 (2003).
37. Grabherr, M.G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644-652 (2011).
38. Haas, B.J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol* **9**, R7 (2008).
39. Finn, R.D. Pfam: clans, web tools and services. *Nucleic Acids Research* **34**, D247-D251 (2006).
40. Boeckmann, B. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* **31**, 365-370 (2003).
41. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25**, 955–964 (1997).
42. Griffiths-Jones, S. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research* **33**, D121-D124 (2004).
43. Griffiths-Jones, S. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research* **34**, D140-D144 (2006).

44. Nawrocki, E.P. & Eddy, S.R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933-2935 (2013).
45. She, R., Chu, J.S.C., Wang, K., Pei, J. & Chen, N. genBlastA: enabling BLAST to identify homologous gene sequences. *Genome Research* **19**, 143-149 (2009).
46. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Research* **14**, 988-995 (2004).
47. Emms, D.M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**(2019).
48. Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**, 772-780 (2013).
49. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540-52 (2000).
50. Nguyen, L., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution* **32**, 268-274 (2015).
51. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**, 1586-1591 (2007).
52. Mongue, A.J., Nguyen, P., Voleníková, A. & Walters, J.R. Neo-sex chromosomes in the monarch butterfly, Danaus plexippus. *G3 Genes|Genomes|Genetics* **7**, 3281 – 3294 (2017).
53. Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLOS Computational Biology* **14**, e1005944 (2018).
54. H, A., K, M., Y, Y., T, O. & T, S. Retrotransposable elements on the W chromosome of the silkworm, *Bombyx mori*. *Cytogenetic and Genome Research* **110**, 144-151 (2015).
55. Smeds, L. et al. Evolutionary analysis of the female-specific avian W chromosome. *Nature Communications* **6**(2015).
56. Pritham, E.J., Putliwala, T. & Feschotte, C. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene* **390**, 3-17 (2007).
57. Vítková, M., Fuková, I., Kubíčková, S. & Marec, F. Molecular divergence of the W chromosomes in pyralid moths (Lepidoptera). *Chromosome Research* **15**, 917-930 (2007).
58. Traut, W. & Marec, F. Sex chromosome differentiation in some species of Lepidoptera (Insecta). *Chromosome Research* **5**, 283 – 291 (1997).
59. Dalíková, M. et al. New insights into the evolution of the W chromosome in Lepidoptera. *Journal of Heredity* **108**, 709-719 (2017).
60. Smith, D.A.S. et al. A neo-W chromosome in a tropical butterfly links colour pattern, male-killing, and speciation. *Proceedings of the Royal Society B: Biological Sciences* **283**, 20160821 (2016).
61. Lukhtanov, V.A. Sex chromatin and sex chromosome systems in nonditrysian Lepidoptera (Insecta). *Journal of zoological systematics and evolutionary research* **38**, 73-79 (2000).
62. Bachtrog, D. et al. Sex determination: Why so many ways of doing it. *PLoS biology* **12**, e1001899 (2014).
63. Bachtrog, D. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nature Reviews Genetics* **14**, 113-124 (2013).
64. Mahajan, S., Wei, K.H.C., Nalley, M.J., Gibilisco, L. & Bachtrog, D. De novo assembly of a young *Drosophila* Y chromosome using single-molecule sequencing and chromatin conformation capture. *PLOS Biology* **16**, e2006348 (2018).

539 65. Gu, L. et al. Dichotomy of dosage compensation along the neo Z Chromosome of the monarch
540 butterfly. *Current Biology* **29**, 4071-4077.e3 (2019).
541 66. Duan, J. & Larschan, E.N. Dosage compensation: how to be compensated...or not? *Current Biology*
542 **29**, R1229-R1231 (2019).
543 67. Metzger, D.C.H., Sandkam, B.A., Darolti, I. & Mank, J.E. Rapid evolution of complete dosage
544 compensation in Poecilia. *Genome Biology and Evolution* **13**(2021).
545 68. Darolti, I. et al. Extreme heterogeneity in sex chromosome differentiation and dosage compensation
546 in livebearers. *Proceedings of the National Academy of Sciences* **116**, 19031-19036 (2019).