# Gene Gain and Loss from the Asian Corn Borer W Chromosome

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 little conservation with available W chromosome sequence in related species or with the Z chromosome, consistent with a non-canonical origin of the W chromosome. The W chromosome has accumulated significant repetitive elements and experienced rapid gene gain from the remainder of the genome, with most genes exhibiting pseudogenization after duplication to the W. The genes that retain significant expression are largely enriched for functions in DNA recombination, the nucleosome, chromatin and DNA binding, likely related to meiotic and mitotic processes within the female gonad.

## Introduction

The non-recombining, sex-specific portions of the genome, namely Y and W chromosomes, exhibit very different properties from the remainder of the genome[1]. Recent advances in sequencing have made it possible to assemble full, or nearly full, sequences of numerous Y chromosomes[2], and these efforts have revealed general patterns such as the retention of genes related to dosage-sensitivity[3,4] and male function[5-8] as one might expect from a chromosome limited in its inheritance to males.

Studying Y chromosomes alone makes it difficult to differentiate the effects of sex-limitation from the effects of limitation to males, and so W chromosomes can be a powerful contrast to reveal the overall convergence of genomic patterns of these unusual regions of the genome. Despite providing an important comparison, sequencing of W chromosomes has lagged somewhat, with the most complete assemblies in birds[3,9] suggesting that although W chromosomes share some similarities with Y chromosomes, namely an abundance of repetitive elements[10], they may differ from Y chromosomes in that they lack genes with female-specific functions[3,9]. Whether this is unique to birds, or is a broader pattern of W chromosomes, remains unclear.

Lepidoptera, butterflies and moths, provide a key additional female heterogametic system[11,12]. The conservation of the Z chromosome has been well established in Lepidoptera[13], however, the W chromosome in Lepidoptera is unusual in that it was recruited into the genome well after the origin of the Z chromosome[14], 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[13,15,16], and may actually have originated from a B chromosome[13]. 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[14].

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)[16], *Cnaphalocrocis medinalis* (Crambidae, Pyraloidea)[17], *Spodoptera exigua* (Noctuidea, Noctuoidea)[18], *Spodoptera frugiperda* (Noctuidea, Noctuoidea)[19], *Trichoplusia ni* (Noctuidea, Noctuoidea)[20] and *Dryas iulia* (Nymphalidae, Papilionoidea)[15]. This work has revealed no detectible homology between the Z and W in *C. pomonella*[16], *T. ni*[21] and *D. iulia*[15]. In contrast, substantial synteny is observed between the W chromosomes of *S. exigua* and *C. medinalis*[17,22]. 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 is therefore not homologous to the deeply conserved corresponding Z chromosome.

To answer questions about the conservation and gene content of the lepidopteran W, we built a chromosome-level genome of the Asian corn borer *Ostrinia furnacalis* Guenée (Lepidoptera: Pyralidae, Pyraloidea), is a major insect pest of corn, widespread in the Asian-Western Pacific region, using a 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. Our catalog of W gene content reveals extensive duplications from all other chromosomes in the genome, with a relatively low percent of genes with persistent gene activity, which are enriched for functions in DNA recombination, the nucleosome, chromatin and DNA binding. Our results suggest that the W chromosome retains gene related to meiotic and mitotic functions within the female gonad, in contrast to the avian W chromosome[3,9] and more consistent with findings from Y chromosomes[5-8].

## 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

83 incubator environment was set at 26 ± 1℃ and 50 ± 5% relative humidity on a
84 photoperiod (Light: Dark = 16:8).

85 ***Genome sequencing and assembly***

86 Genomic DNA was extracted from a female laboratory-fed pupa. From this single
87 individual, we obtained 32.96 Gb circular consensus sequencing (CCS) data from one
88 cell of the PacBio Sequel II platform at Biomarker Co., Ltd., Qingdao, China (Table S1).
89 High accuracy PacBio Hifi data were assembled using hifiasm (v0.12) software to
90 obtain a draft genome[23]. We used purge_haplotigs[24] to remove redundant contigs
91 and generate a non-redundant assembled genome.

92 ***Hi-C scaffolding***

93 We constructed Hi-C libraries (300-700 bp insert size), following Rao et al.[25],
94 and sequenced them with pair-end 150 Illumina Hiseq at Biomarker Co., Ltd.,
95 Qingdao, China. 53.05 Gb of clean data were produced after filtering adapter
96 sequences, primer sequences and low-quality data (Table S2). The resulting trimmed
97 reads were aligned to the draft assembly with BWA (v0.7.10-r789), retaining only
98 uniquely aligned read pairs with mapping quality >20 for further analysis. Invalid read
99 pairs, including dangling-end, self-cycle, re-ligation and dumped products, were
100 removed by HiC-Pro (v2.8.1)[26].

101 The 56.25% of unique mapped read pairs represent valid interaction pairs and were
102 used for correction, clustering, ordering and orientation of scaffolds into
103 chromosomes with LACHESIS[27]. Before chromosome assembly, we performed a
104 preassembly for error correction of contigs which required the splitting of contigs
105 into segments of 50 kb on average. Then Hi-C data were mapped to these segments
106 using BWA (v0.7.10-r789). The uniquely mapped data were retained for the
107 assembly, and any two segments which showed inconsistent connection with
108 information from the raw scaffolds were checked manually. Parameters for running
109 LACHESIS included: CLUSTER_MIN_RE_SITES = 27; CLUSTER_MAX_LINK_DENSITY = 2;
110 ORDER_MIN_N_RES_IN_TRUNK = 15; ORDER_MIN_N_RES_IN_SHREDS = 15. After

111 this step, placement and orientation errors exhibiting obvious discrete chromatin
112 interaction patterns were manually adjusted. Finally, we constructed a heatmap
113 based on the interaction signals of valid mapped read pairs between chromosomes.

114 ***Genome annotation***

115 Transposable elements (TEs) were identified by a combination of homology-based
116 and de novo approaches. We first carried out a de novo repeat prediction using
117 RepeatModeler2 (v2.0.1)[28], which implements RECON (v1.0.8)[29] and RepeatScout
118 (v1.0.6)[30]. Then the predicted results were classified using RepeatClassifier[28] by
119 means of repbase (v19.06)[31], REXdb (v3.0)[32] and Dfam (v3.2)[33]. Secondly, we
120 performed a de novo prediction for long terminal repeats (LTRs) using LTR_retriever
121 (v2.8)[34] via LTRharvest (v1.5.9)[35] and LTR_finder (v2.8)[36]. A non-redundant species-
122 specific TE library was constructed by combining the de novo TE library above with
123 the known databases. Final TE sequences in the *O. furnacalis* genome were identified
124 and classified by homology search against the library using RepeatMasker (v4.10)[37].
125 Tandem repeats were annotated by Tandem Repeats Finder (TRF v409)[38] and
126 MIcroSAtellite identification tool (MISA v2.1)[39].

127 We integrated de novo prediction, homology search, and transcript-based assembly
128 to annotate protein-coding genes. The de novo gene models were predicted using
129 Augustus (v2.4)[40] and SNAP (2006-07-28)[41], both ab initio gene-prediction
130 approaches. For the homolog-based approach, we used GeMoMa (v1.7)[42] with
131 reference gene models from *Drosophila melanogaster*, *Bombyx mori*, *Chilo*
132 *suppressalis* and *Galleria mellonella*. For the transcript-based prediction, RNA from a
133 mixture sample containing eggs, larvae, pupae and adults whole body of females and
134 males was used for generating sequencing libraries and was sequenced on an
135 Illumina NovaSeq 6000 platform (25.82-fold coverage of the genome, Table S3).
136 RNA-sequencing data were mapped to the reference genome using Hisat (v2.0.4)[43]
137 and assembled with Stringtie (v1.2.3)[44]. GeneMarkS-T (v5.1)[45] was used to predict
138 genes based on the assembled transcripts. PASA (v2.0.2)[46] was used to predict genes
139 based on the unigenes assembled by Trinity (v2.11)[47]. Gene models from these

140 different approaches were combined with EVM (v1.1.1)[48] and updated by PASA
141 (v2.0.2)[46]. The final gene models were annotated by searching the GenBank Non-
142 Redundant (NR, 20200921), TrEMBL (202005), Pfam (v33.1)[49], SwissProt (202005)[50],
143 eukaryotic orthologous groups (KOG, 20110125), gene ontology (GO, 20200615) and
144 Kyoto Encyclopedia of Genes and Genomes (KEGG, 20191220) databases.

145 We used tRNAscan-SE (v1.3.1)[51] with default parameters to identify transfer RNAs
146 (tRNAs), and barrnap (v0.9) with default parameters to identify the ribosomal RNAs
147 (rRNAs) based on Rfam (v12.0)[52]. miRNAs were identified with miRbase[53]. Small
148 nucleolar RNA (snoRNAs) and small nuclear RNA (snRNAs) were identified with
149 Infernal (v1.1.1)[54], using the Rfam (v12.0) database[52].

150 Pseudogenes have similar sequences to functional genes, but may have lost their
151 biological function because of mutations. GenBlastA (v1.0.4)[55] was used to scan the
152 whole genome after masking predicted functional genes. Putative candidates were
153 then analyzed by searching for premature stop codon and frameshift mutations using
154 GeneWise (v2.4.1)[56].

155 ***Comparative genomics and phylogenetic reconstruction***

156 Protein sequence alignments between *O. furnacalis* and seven other lepidopteran
157 species (*B. mori*, *C. suppressalis*, *C. pomonella*, *C. medinalis, S. exigua, Spodoptera*
158 *fugiperda,* and *T. ni*) were performed with diamond (-e < 1e-5), then alignment
159 results were analyzed and the homologous chromosomal regions were identified
160 with MCScanX (MATCH_SCORE: 50, MATCH_SIZE: 5, GAP_PENALTY: -1,
161 OVERLAP_WINDOW: 5, E_VALUE: 1e-05, MAX GAPS: 15). The synteny relationships
162 among chromosomes were displayed using circos (v0.69–9).

163 We used the protein sequences of *O. furnacalis and* ten other insect species (*S.*
164 *frugiperda, Spodoptera litura, B. mori, C. suppressalis, Danaus plexippus, Melitaea*
165 *cinxia, Papilio Xuthus, C. pomonella*, *Plutella xylostella* and *D. melanogaster*) for
166 phylogenetic analysis, keeping only the longest transcript of each gene for analysis,
167 and using OrthoFinder (v2.5.4)[57] with default settings to identify orthologues and

168 homologues. To infer the phylogeny of these insects, multiple sequence alignments
169 of single-copy orthologs were performed using MAFFT (v7.471)[58]. Then we extracted
170 conserved sequences with gblocks (v0.91b)[59] and concatenated them to a single
171 sequence alignment. The resulting sequence alignment was used to construct a
172 maximum likelihood phylogenetic tree using IQ-TREE (v1.6.12)[60] (outgroup: *D.*
173 *melanogaster*). Divergence times between various species were estimated by
174 MCMCtree in PAML (v4.9j)[61]. Three standard divergence time points from the
175 TimeTree database (http://timetree.org/) were used for calibration: *S. frugiperda* – S.
176 *litura*, 100-122 million years ago (Mya), *D. plexippus* – *M. cinxia*, 76-102 Mya, (*D.*
177 *plexippus*, *M. cinxia*) – *P. xuthus*, 76-146 Mya. The tree was visualized using FigTree
178 (v1.4) (http://tree.bio.ed.ac.uk/software/figtree/). The gene count table from
179 OrthoFinder (v2.5.4) was used as inputs to examine the expansion and contraction of
180 each gene family in cafe (v5.1.0)[62].

181 ***Detection of sex chromosomes***

182 In order to identify the sex chromosomes, we performed genome resequencing with
183 five female and male pupae. High-quality clean data (24.95-28.80-fold coverage of
184 the genome) were obtained through the pair-end 150 Illumina Hiseq platform at
185 Biomarker Co., Ltd., Qingdao, China (Table S3). The clean data were aligned to
186 reference genome with Bwa-men (v0.7.17). We compared the coverage differences
187 between male and female samples[63] to distinguish the Z, W and autosomes.
188 Specifically, we used the genomecov and groupby in Bedtools (v2.30.0) to obtain a
189 per-base median coverage depth for each chromosome and normalized them by the
190 mean of all chromosome median coverages for each sample. Normalized coverage
191 depth was averaged by sex to produce a coverage depth per chromosome for each
192 sex. Then we compared coverage depth between sexes for each chromosome and
193 calculated $\log_2$ male:female (M:F) coverage ratio [$\log_2$(M:F coverage)]. Autosomes
194 have an equal coverage between sexes [$\log_2$ (M:F coverage) = 0], while the Z
195 chromosome should show approximately two-fold greater coverage in the males
196 [$\log_2$(M:F coverage) = 1]. The W chromosome should show a strong female-biased

197 coverage pattern. We also calculated the M:F median coverage ratio along each
198 chromosome with nonoverlapping 1000 bp windows.

199 ***W-gene homologs search and calculation of synonymous divergence ($d_S$)***

200 Predicted protein sequences were used to detect reciprocal best hits between the W
201 chromosome and the remainder of the genome using getRBH.pl
202 (https://github.com/Computational-conSequences/SequenceTools) for *O. furnacalis*.
203 For each pair of orthologous genes, we deleted stop codes and aligned the coding
204 sequences using macse (v2.05)[64], then extracted conserved sequences with gblocks
205 (v0.91b)[59]. The resulting alignments were used as inputs of CODEML in PAML (v4.9j)[61]
206 to estimate pairwise synonymous divergence with settings runmode = -2, seqtype = 1
207 and CodonFreq = 2. Since divergence estimates are not reliable for saturated sites,
208 we excluded orthologs with $d_S > 3$[65].
209 The homolog search of *O. furnacalis* W in *C. pomonella*, *C. medinalis*, *S. exigua*, *S.*
210 *fugiperda* and *T.ni* genome also used getRBH.pl. We assessed enrichment of Gene
211 Ontology[66,67] terms for W gene content.

212 ***W-gene expression level***

213 We used two different RNA-Seq datasets to determine W expression, the mixture
214 sample of RNA-Seq data from eggs, larvae, pupae and adults used for genome
215 annotation (described above), and data from female gonad, as sex-limited
216 chromosomes are often largely expressed in the gonad. For female gonad, RNA from
217 five gonad samples of adult females was sequenced on an Illumina NovaSeq 6000
218 platform (10.14-fold coverage of the genome, Table S3). For each dataset separately,
219 we used fastp (v0.20.0)[68] to filter out low quality reads and to remove adapters with
220 the default parameters. Then we mapped the clean reads to the *O. furnacalis*
221 reference genome using HISAT2 (v2.1.0)[43]. The FPKM (fragments per kilobase of exon
222 per million fragments mapped) of each gene was determined using Stringtie
223 (v2.1.4)[44] based on the annotated GFF file. For female gonad, we used averaged
224 FPKM of five samples.

225 ***Copy number variation of W and Z/autosome genes***

226 The amino acid sequences of protein-coding genes from whole genome were used as
227 the input of the blastp mapping against the Swiss-Prot database to assign gene
228 symbols (abbreviations for the gene names). We BLASTed each W gene to the
229 remainder of the genome to identify the best hit and only retained orthologous
230 genes on the W chromosome and Z/autosomes that consistently mapped to the
231 same protein (reciprocal best BLAST hit) and calculated their copy number following
232 the protocol in Mueller et al.[69].
233
234

235 **Results**

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

237 Our chromosome assembly strategy employed PacBio HiFi (CCS) sequencing data to
238 assemble the draft genome and Hi–C data to detect chromosomal contact
239 information. These PacBio long-reads were self-corrected using Quiver and
240 assembled into a draft genome assembly with a total length of 493.10Mb, consisting
241 of 57 contigs with an N50 length of 15.69Mb (Table S4). The assembled genome size
242 is similar to that obtained from genome surveys (Supplementary Figure 1). Next, Hi–C
243 linking information further anchored, ordered, and oriented 46 contigs to 32
244 chromosomes (30 autosomes, with Z and W sex chromosomes), which contained
245 86.79% of the whole genome assembly (Table S4, Supplementary Figure 2). The
246 chromosome-level genome assembly was 492.57Mb with a scaffold N50 of 16.47Mb
247 (Table S4).

248 We evaluated the completeness of *O. furnacalis* genome assembly with the
249 Metazoan Benchmark of Universal Single-Copy Orthologs (BUSCO v4) and CEGMA
250 core genes (CEGMA v2.5) datasets. Our assembly contained 97.69% of BUSCO genes,
251 of which 97.06% were single copy and 0.63% were duplicates, and 96.07% of CEGMA
252 genes, of which 79.03% were highly conserved (Table S5). To further evaluate the
253 genome assembly quality, genomic DNA sequencing data from Illumina HiSeq were

254 mapped to the assembly scaffolds, with a 99.07% coverage rate (Table S5). Finally,
255 lepidopteran genomes typically exhibit very high levels of synteny. Whole-genome
256 alignment of the *O. furnacalis* assembly to the chromosomes of the *S. litura* revealed
257 that chromosomal linkage and ordering of genes are highly conserved
258 (Supplementary Figure 3). All these analyses indicated that the genome assembly is
259 both highly reliable and complete for subsequent analyses.

260 ***Genome annotation***

261 In total, 860,391 repeat sequences spanning 204.4Mb were identified, constituting
262 41.45% of the *O. furnacalis* genome (Table S4, Figure 1a). We integrated the result of
263 *ab initio*, homology-based and RNA-seq methods to annotate protein-coding genes,
264 most of the which came from homology-based and RNA-seq methods, indicating the
265 high-quality of annotation (Supplementary Figure 4). Finally, we obtained 16,509
266 protein coding genes in the *O. furnacalis* genome, which has similar gene features to
267 other lepidopteran genomes (gene length, gene number, coding length, intron length
268 of each gene, exon length) (Supplementary Figure 4). Next, we identified different
269 types of noncoding RNAs (ncRNAs), including 7,710 transfer RNAs (tRNA), 73
270 ribosomal RNAs (rRNA), and 39 microRNAs (miRNA) (Table S6). In addition, we
271 annotated 167 pseudogenes in the *O. furnacalis* genome, defined as any genomic
272 sequence that is similar to another gene but is defective, such as containing a
273 premature stop codon or a frameshifts mutation[70]. Most pseudogenes were located
274 in the chromosome LG3 (Figure 1c, Table S6), the W chromosome (see below). We
275 performed a functional annotation for all predicted protein coding genes using NCBI
276 nonredundant, EggNOG, GO, KEGG, SWISS-PROT and Pfam databases, with 98.21% of
277 our predicted genes assigned to at least one of these databases (Table S6).

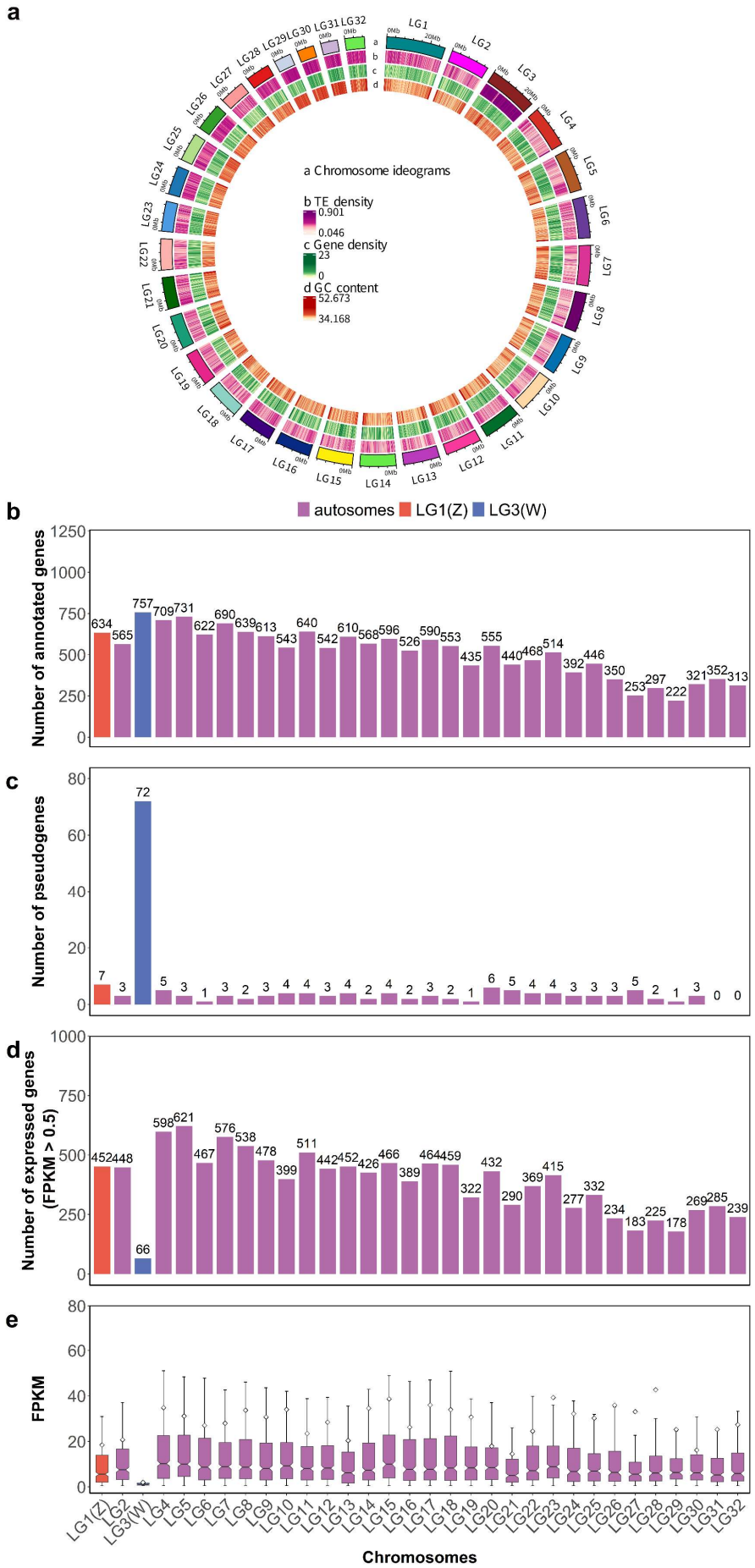

a
a Chromosome ideograms
b TE density
0.901
0.046
c Gene density
23
0
d GC content
52.673
34.168
b
autosomes
LG1(Z)
LG3(W)
Number of annotated genes
c
Number of pseudogenes
d
Number of expressed genes (FPKM > 0.5)
e
FPKM
Chromosomes

**Fig 1. Genomic characterization of *Ostrinia furnacalis*.** a. Circos plot 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 expressed genes (FPKM > 0.5 in mixed sample). e. Average chromosomal expression level (excluding genes with FPKM ≤ 0.5 in mixed sample).

We used OrthoFinder to find orthologous genes among *O. furnacalis* and ten other insect species (see Materials and Methods). A total of 16231 orthologous groups with 1163 single-copy orthologous genes were identified. We inferred a phylogenetic 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).

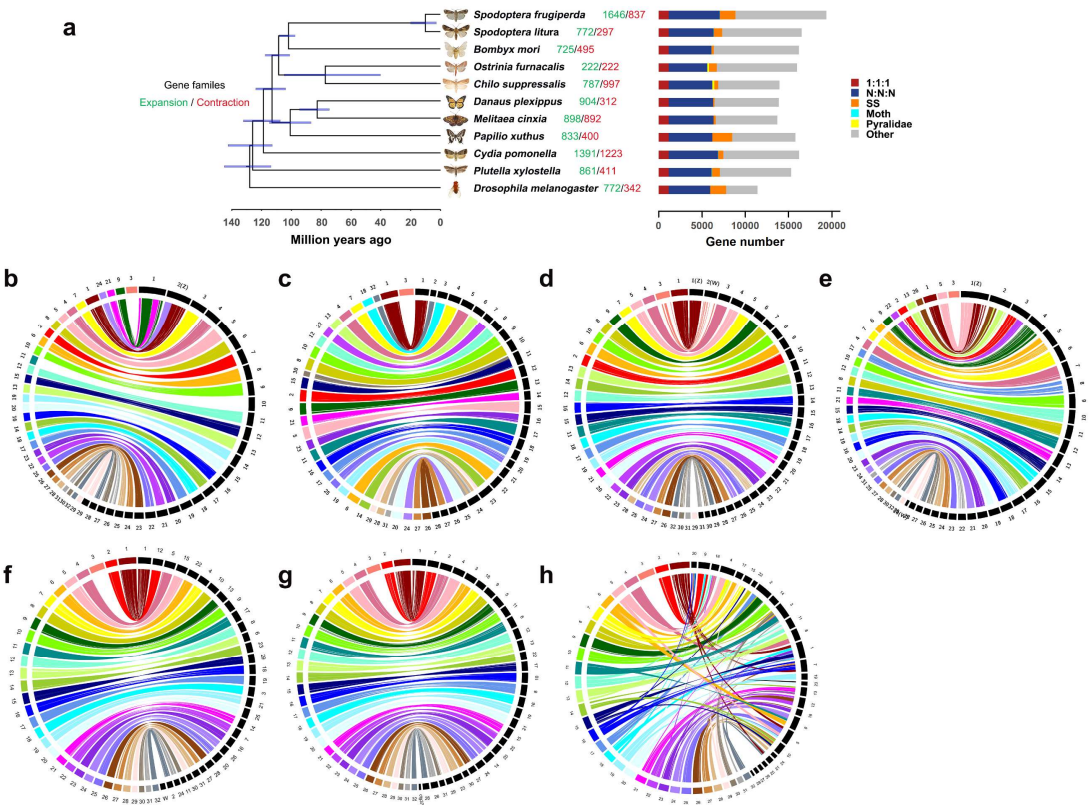



**Fig 2. a.** Phylogenetic tree and gene orthology of *O. furnacalis* with nine lepidopteran genomes and *D. melanogaster*. Comparative analysis of synteny between *O. furnacalis* and **b.** *C. suppressalis*, **c.** *B. mori*, **d.** *C. medinalis*, **e.** *C. pomonella,* **f.** *S.*

*exigua,* **g.** *S. fugiperda*, and **h.** *T. 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).

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

We compared the syntenic relationships between *O. furnacalis* and other lepidopterans, including *B. mori, C. suppressalis, C. medinalis*, *S. litura, S. exigua, S. fugiperda, and T. ni* (Figure 2b-h). In general, the *O. furnacalis* genome shows a high level of synteny with other lepidopteran genomes, though with some fusion and fission events. *O. furnacalis* LG1 is syntenic with the Z chromosome in all other species, suggesting it is the Z chromosome in *O. furnicalis* as well, and this is also evident from M:F coverage analysis (Figure 3a), which reveals two-fold greater coverage in males (ZZ) compared to females (ZW).

*O. furnacalis* LG3, the W chromosome, exhibits female-biased coverage consistent with a female-limited chromosome (Figure 3a). The M:F coverage is highly variable across chromosome LG3 compared to all other chromosomes (Figure 3b), due in large part to the abundance of TEs on this chromosome (Supplementary Figure 5, Supplementary Figure 6). The W chromosome (LG3) is the second largest chromosome (22.23 Mb) in our assembly (Figure 1a), with the largest number of predicted protein coding genes (excluding pseudogenes) compared with other chromosomes (Figure 1b). The W also has the largest number of pseudogenes (Figure 1c), and contains 43.1% of all pseudogenes annotated in the genome. Many genes that are not technically pseudogenized were expressed below our 0.5 FPKM threshold. Only 66 protein coding genes showed significant expression above this threshold in our mixed sample (comprised of eggs, larvae, pupae and adults, see Materials and Methods, Figure 1d, 1e). 48 W genes showed expression >0.5 FPKM in adult female gonads.

Furthermore, LG3 has a notably greater repeat density and significantly different

repeat composition compared to other chromosomes (Supplementary Figure 5, Supplementary Figure 6) with a larger proportion of satellites, DIRS, LINE, Copia and Gypsy, and an enrichment of *Maverick* elements.

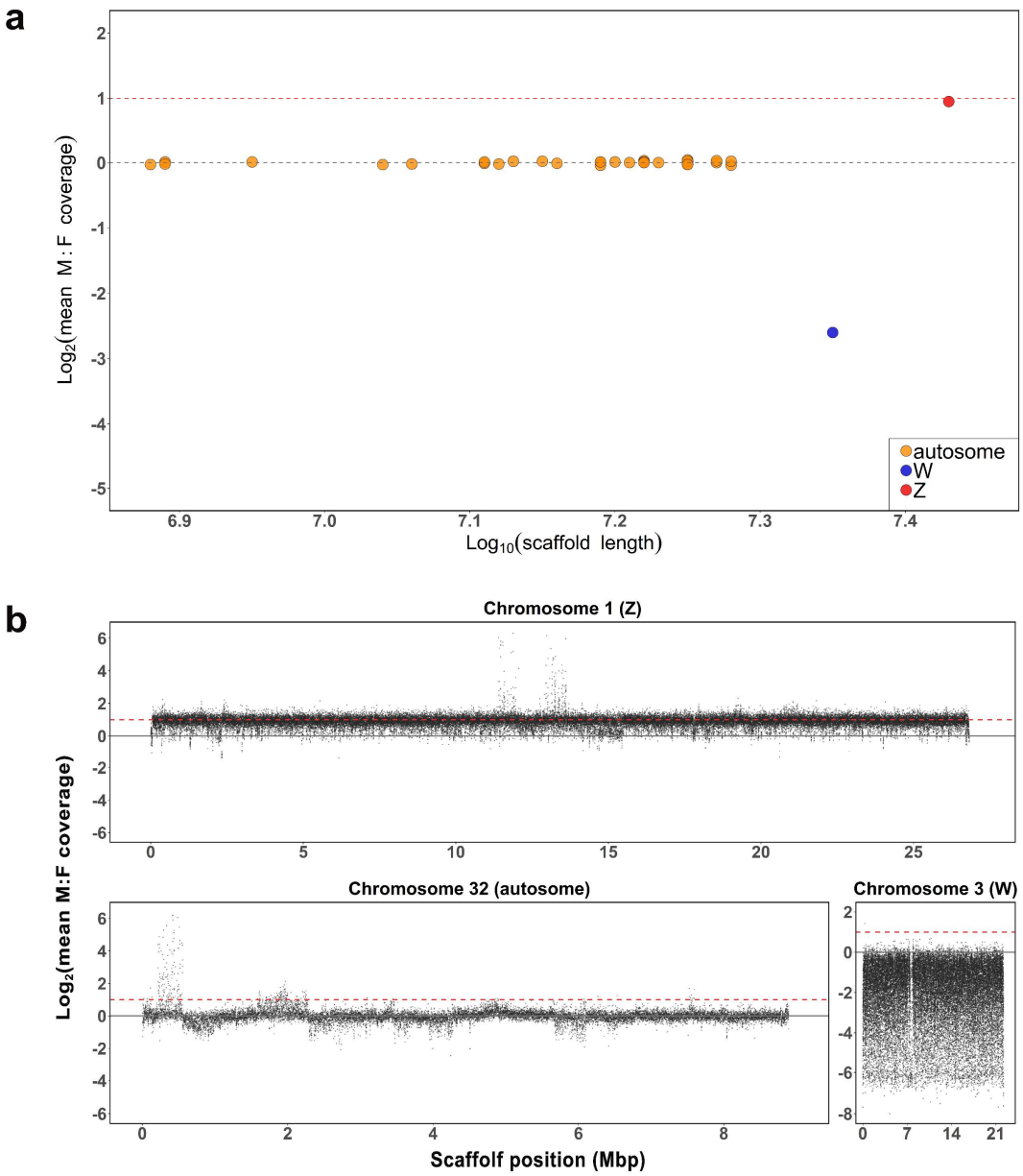



**Fig 3. Detection of sex chromosomes. a.** Male:female coverage ratios for each chromosome, plotted by chromosome length. Each point represents a single chromosome. The dashed gray line is the theoretical expectation for autosomes and the dashed red line shows the expectation for the Z chromosome. **b.** Male:female coverage ratios plotted in 500bp windows across scaffolds for Z, W, and a representative autosome (LG32).

339 ***W homology***

340 Given that previous reports have found no evidence of homology between the Z and
341 W in some basal lepidopteran species[15,16,21], we next investigated the evolutionary
342 history of the *O. furnacalis* W chromosome. Using our W gene set, we first examined
343 homology between the *O. furnacalis* W chromosome and the W chromosome from
344 other lepidopterans. The *O. furnacalis* W shows substantial homology with the W in
345 *C. medinalis*, consistent with some form of shared ancestry within the family
346 Crambidae. However, we observed no discernable homology between the *O.*
347 *furnacalis* W and the W chromosome of any of the more distantly related
348 lepidopterans that we queried (Figure 2b-h).

349

350 Reciprocal best BLAST hits between the W coding catalog and the remainder of the
351 *O. furnicalis* genome reveals similarity throughout the genome (Figure 4), consistent
352 with a B chromosome origin of the W chromosome followed by gene duplication
353 from locations throughout the genome. Of the 482 W coding genes having
354 Z/autosome paralogs (Figure 4), 202 exhibited fewer or no introns in the W paralogs,
355 consistent with retrotransposition.

356

357

358

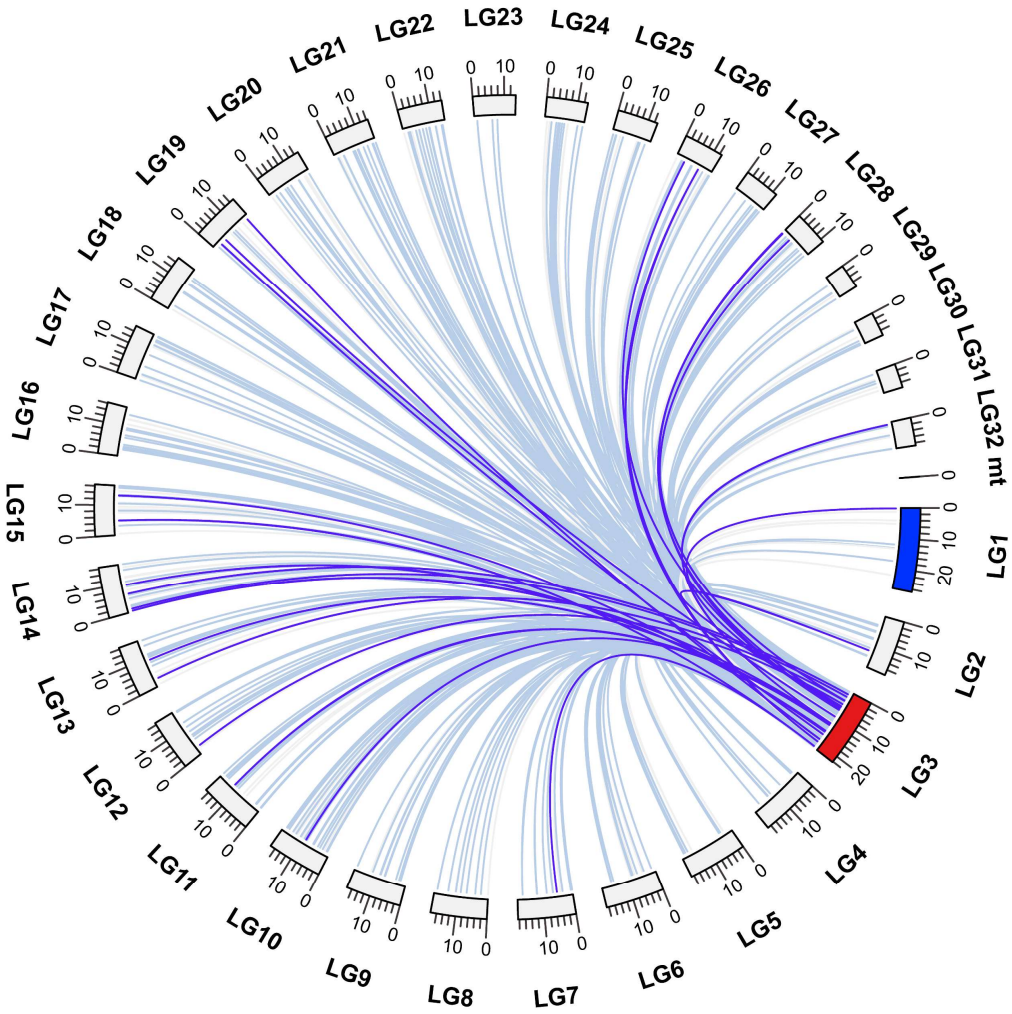
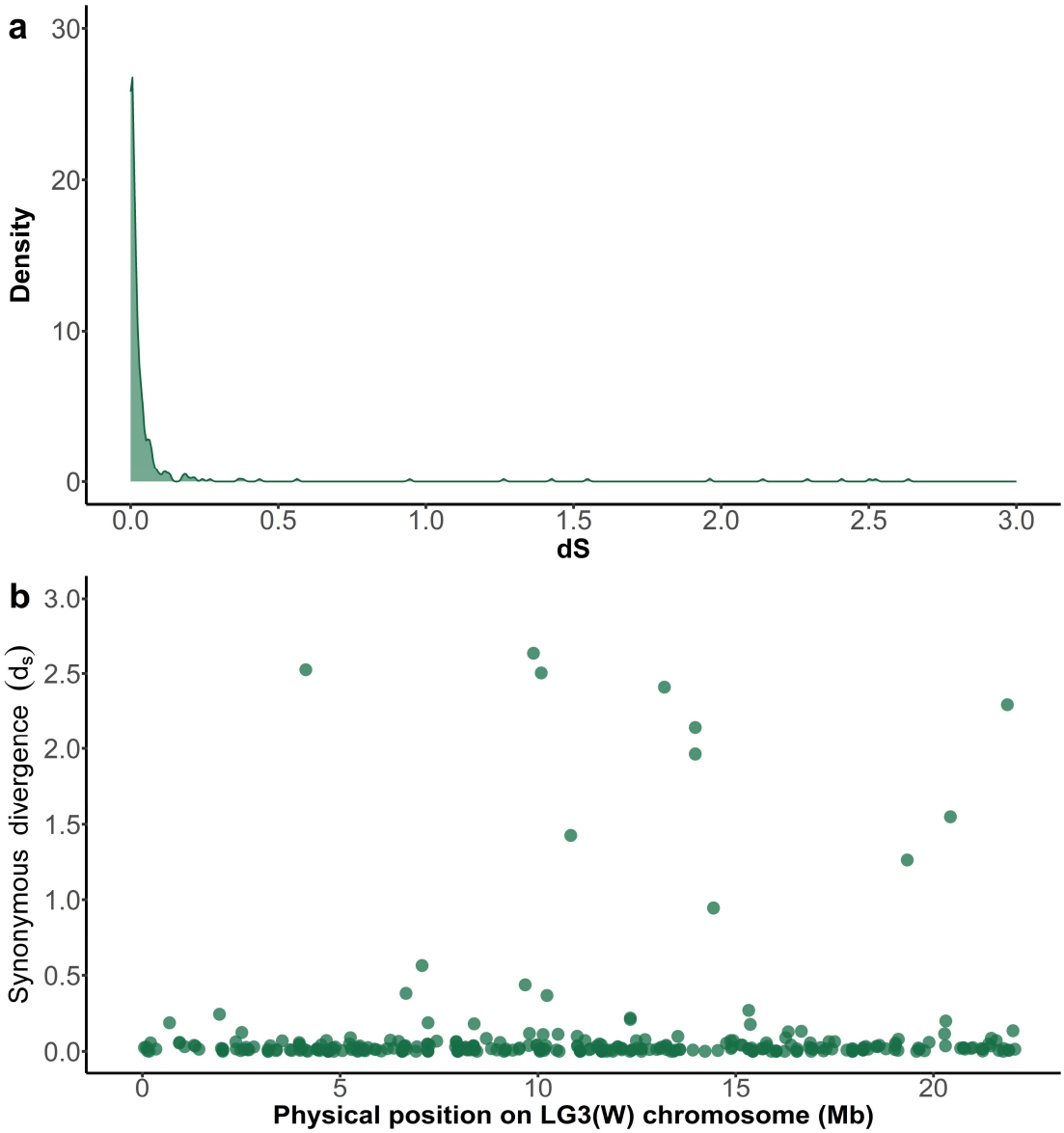



**Fig. 4 The reciprocal best hits search for W (LG3) genes.** W chromosome (LG3) is marked in red and Z chromosome (LG1) in blue, mt represent the mitochondrial genome of *O. furnacalis.* Light blue line: identity ≥ 90%; purple line: 80% ≤ identity < 90%; light grey line: identity < 80%.

We next examined synonymous divergence ($d_S$) for these paralogs in order to determine the timing of gene duplication to the W chromosome. The density of $d_S$ (Figure 5a) shows that most paralogs with W genes have relatively low divergence, and the distribution of paralog $d_S$ across the W chromosome suggests that duplications occur randomly throughout the chromosome. Given the low percent of genes on the W chromosome that exhibit expression >0.5 FPKM and the high

371 number of pseudogenes on the W (Figure 1), it is likely that most duplicates to the W
372 chromosome fail to maintain significant expression and are silenced relatively
373 quickly.

374




375

376 **Fig 5. Synonymous divergence between W gene and best BLAST hit gene from the**
377 **remainder of the genome**. a. Density plot, b. Distribution across the W chromosome.

378

379 Sex-limited chromosomes often contain many copies of some genes[5,71], and so we
380 examined copy number for genes on the W chromosome and their most similar

paralog (Figure 6, Table S7). Overall, we observed lower number of gene copies on the W than for the autosomal/Z chromosome paralog.

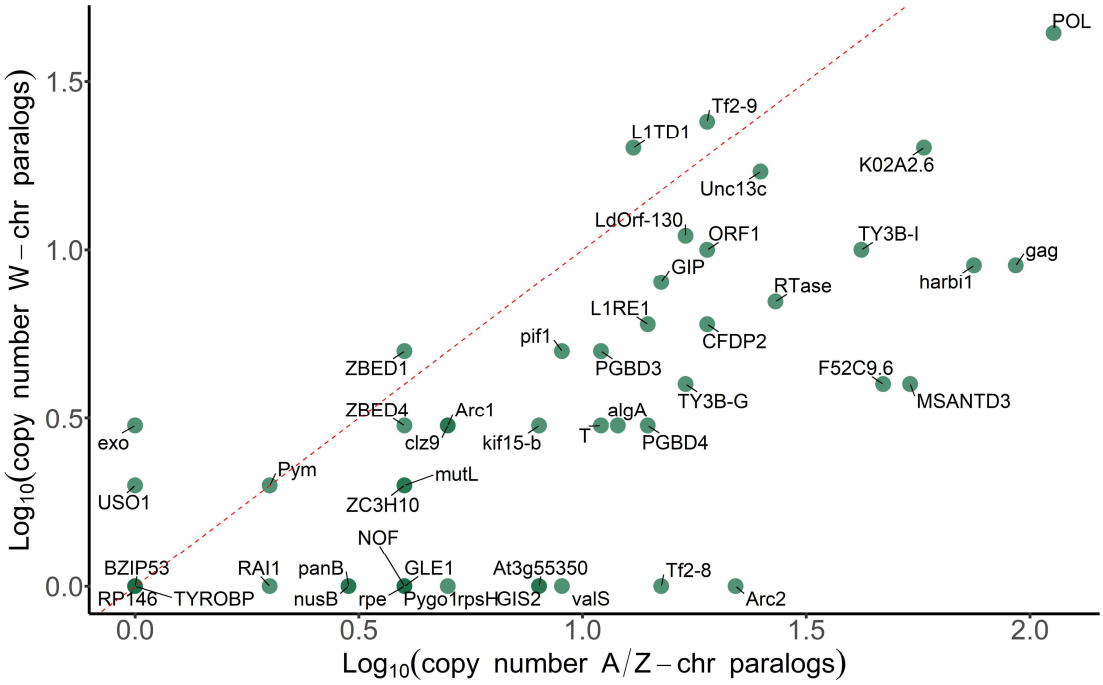



**Fig. 6. Copy number for W and autosomal/Z chromosome paralogs.**

Given the rapid apparent decay of W chromosome paralogs (Figure 5), we examined the Gene Ontology enrichment for W expressed genes (>0.5 FPKM) (Tables 1 and 2), and observed statistical enrichment of terms related to functions in DNA recombination, the nucleosome, chromatin and DNA binding. Together, these results suggest that many genes retained with functional expression on the W chromosome are related to mitotic and meiotic processes, much of it within the female gonad.

**Table 1. Functional Enrichment of Significantly Expressed Genes (FPKM > 0.5 from mixed sample) on *O. furnicalus* W chromosome**

| PANTHER GO | FDR | Expressed W genes |
|---|---|---|
| **Biological** | | |
| Negative regulation of DNA recombination | 7.64 e-11 | 5 |

| | | |
|---|---|---|
| Nucleosome assembly | 8.35 e -10 | 5 |
| Chromosome condensation | 2.48 e-8 | 5 |
| Telomere maintenance | 2.67 e-02 | 2 |
| **Molecular** | | |
| Nucleosomal DNA binding | 1.16 e-11 | 5 |
| Structural constituent of chromatin | 1.13 e-10 | 5 |
| Double-stranded DNA binding | 1.66 e-7 | 7 |
| **Cellular** | | |
| Nucleosome | 1.68 e-10 | 5 |
| Chromatin | 6.12 e-7 | 6 |

**Table 2. Functional Enrichment of Significantly Expressed Genes (FPKM < 0.5 from female gonad) on *O. furnicalus* W chromosome**

| PANTHER GO Term | FDR | Expressed W genes |
|---|---|---|
| **Biological** | | |
| Negative regulation of DNA recombination | 2.04e-10 | 5 |
| Nucleosome assembly | 2.22e-9 | 2 |
| Chromosome condensation | 7.70e-08 | 5 |
| **Molecular** | | |
| Nucleosomal DNA binding | 3.09e-11 | 5 |
| Structural constituent of chromatin | 3.02e-10 | 5 |
| Double-stranded DNA binding | 8.28e-05 | 6 |
| **Cellular** | | |
| Nucleosome | 4.46e-10 | 5 |
| Chromatin | 3.19e-6 | 6 |

## Discussion

Here we report a high quality, chromosome-level genome for *O. furnicalis*. Our reference genome includes a single, contiguous sequence for the female-specific W chromosome, allowing us to query the content of this unique region of the genome.

W and Y chromosomes are often enriched for pseudogenes, as the lack of recombination in these regions leads to high rates of gene silencing[1]. At the same time, these regions often accumulate repeat sequences[9,72], resulting in significantly higher repeat content compared to other chromosomes[9,16,17]. Consistent with this,

409 we although we found that the number of annotated protein-coding genes of W
410 chromosome is indeed the largest of all the chromosomes in the genome (Figure 1),
411 most of these are pseudogenes (Figure 1c), and those genes that retain functional
412 coding sequence have low overall transcriptional activity (Figure 1d)[12].

413 We also found the W is enriched for repetitive elements in *O. furnicalis*
414 (Supplementary Figure 5 and Supplementary Figure 6), with the number of *Maverick*
415 elements particularly higher on the W chromosome compared to the rest of the
416 genome (Supplementary Figure 5). *Maverick*, also known as *Polinton*, is a DNA
417 transposon widespread in eukaryotes[73], and one of the most striking characteristics
418 of *Maverick* is the large size, typically 15–20 kb[73].

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

420 We searched orthologs to Z-linked genes of *O. furnacalis* in other lepidopterans, and
421 found that although the Z chromosome shows clear strong conservation[14] (Figure 2)
422 the *O. furnacalis* W chromosome is only conserved with *C. medinalis*, also a member
423 of Crambidae (Figure 2b-h). This is consistent with previous work suggesting that the
424 composition of the W varies even among species with the same family[74].

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

426 Karyotype work has revealed a complex history of the W chromosome, including the
427 repeated origin of neo-W chromosomes in many lepidopteran lineages. The
428 evolutionary history of the W chromosome in Lepidoptera differs from the canonical
429 model of sex chromosome formation in that it was recruited, possibly from a B
430 element, after the formation of the Z chromosome[14], at the common root of Ditrysia
431 and Tischerioidea[12,75]. Lineages ancestral to this recruitment exhibit Z0:ZZ
432 karyotypes[14]. The B-chromosome origin of the W is supported by the fact that in
433 some basal lineages, W chromosome bears no homology to the Z[13,15,16].

434 We observe little similarity in gene content between the W and Z chromosome
435 (Figure 4), a steady rate of gene duplication to the W chromosome from throughout
436 the genome (Figure 5), and reduced intron number for many W paralogs compared

437 to their most similar autosomal/Z-linked copy suggests that this likely often occurs
438 through retrotransposition across the W chromosome. Genes retrotransposed or
439 otherwise duplicated to the W chromosome will immediately experience the full
440 degenerative effects of a non-recombining region[1,76], and consistent with this, we
441 observe high levels of pseudogenization (Figure 1) on the W chromosome. However,
442 those genes on the W that retain functional coding sequence and expression (e.g.
443 non-pseudogenized) are enriched for mitotic and meiotic functions, and display Gene
444 Ontology term enrichments related to recombination, chromosome packaging and
445 replication (Tables 1 and 2). Although it has been suggested that W chromosomes
446 may differ from Y chromosomes in that they do not acquire genes related to female
447 gonadal function[3,9] in the same way that Y chromosomes retain genes related to
448 male function[5-8], our results suggest that this may not be a generalized pattern of W
449 chromosomes. Indeed, the W chromosome in *O. furnicalis* are enriched for functions
450 related to female meiosis and mitosis.

451

452

453 **Authors' contributions**

454 LB conceived and designed the research. WD performed the experiments, analyzed
455 and interpreted the data, and wrote the manuscript. JEM helped analyze the data
456 and write the manuscript. All authors read and approved the final manuscript.

457

458 **Acknowledgments**

459 This work was supported by the Beijing Agriculture Innovation Consortium (BAIC02-
460 2023) and National Key R&D Program of China (2022YFD1400500). JEM gratefully
461 acknowledges support from NSERC and a Canada 150 Research Chair. This work was
462 supported by the high-performance computing platform of China Agricultural
463 University.

## Competing interests

The authors declare that they have no competing interests.

## References


1. Bachtrog, D. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nature Reviews Genetics* **14**, 113-124 (2013).
2. Tomaszkiewicz, M., Medvedev, P. & Makova, K.D. Y and W chromosome assemblies: approaches and discoveries. *Trends in Genetics* **33**, 266-282 (2017).
3. Bellott, D.W. et al. Avian W and mammalian Y chromosomes convergently retained dosage-sensitive regulators. *Nature Genetics* **49**, 387-394 (2017).
4. Bellott, D.W. & Page, D.C. Dosage-sensitive functions in embryonic development drove the survival of genes on sex-specific chromosomes in snakes, birds, and mammals. *Genome Research* **31**, 198-210 (2021).
5. Rhie, A. et al. The complete sequence of a human Y chromosome. *bioRxiv* (2022).
6. Cechova, M. et al. Dynamic evolution of great ape Y chromosomes. *PNAS* **117**, 26273-26280 (2020).
7. Koerich, L.B., Wang, X., Clark, A.G. & Carvalho, A.B. Low conservation of gene content in the *Drosophila* Y chromosome. *Nature* **456**, 949-951 (2008).
8. Tobler, R., Nolte, V. & Schlötterer, C. High rate of translocation-based gene birth on the *Drosophila* Y chromosome. *Proceedings of the National Academy of Sciences* **114**, 11721-11726 (2017).
9. Smeds, L. et al. Evolutionary analysis of the female-specific avian W chromosome. *Nature Communications* **6**, 7330 (2015).
10. Peona, V. et al. The avian W chromosome is a refugium for endogenous retroviruses with likely effects on female-biased mutational load and genetic incompatibilities. *PHILOSOPHICAL TRANSACTIONSOF THE ROYAL SOCIETY B* **376**, 20200186 (2021).
11. Blackmon, H., Ross, L. & Bachtrog, D. Sex determination, sex chromosomes, and karyotype evolution in insects. *Journal of Heredity* **108**, 78-93 (2016).
12. Traut, W. & Marec, F. Sex chromatin in Lepidoptera. *The Quarterly Review of Biology* **2**, 239-259 (1996).
13. Fraïsse, C., Picard, M.A.L. & Vicoso, B. The deep conservation of the Lepidoptera Z chromosome suggests a non-canonical origin of the W. *Nature Communications* **8**, 1486 (2017).
14. Sahara, K., Yoshido, A. & Traut, W. Sex chromosome evolution in moths and butterflies. *Chromosome Research* **20**, 83-94 (2012).
15. Lewis, J.J. et al. The *Dryas iulia* genome supports multiple gains of a W chromosome from a B chromosome in butterflies. *Genome Biology and Evolution* **13**, evab128 (2021).
16. Wan, F. et al. A chromosome-level genome assembly of *Cydia pomonella* provides insights into chemical ecology and insecticide resistance. *Nature Communications* **10**, 4237 (2019).
17. Zhao, X. et al. A chromosome−level genome assembly of rice leaffolder, *Cnaphalocrocis medinalis*. *Molecular Ecology Resources* **21**, 561-572 (2021).
18. Zhang, F., Zhang, J., Yang, Y. & Wu, Y. A chromosome-level genome assembly for the beet armyworm (*Spodoptera exigua*) using PacBio and Hi-C sequencing. *bioRxiv* (2020).
19. Xiao, H. et al. The genetic adaptations of fall armyworm *Spodoptera frugiperda* facilitated its rapid global dispersal and invasion. *Molecular Ecology Resources* **20**, 1050-1068 (2020).
20. Chen, W. et al. A high-quality chromosome−level genome assembly of a generalist herbivore, *Trichoplusia ni*. *Molecular Ecology Resources* **19**, 485-496 (2019).
21. Fu, Y. et al. The genome of the Hi5 germ cell line from *Trichoplusia ni*, an agricultural pest and novel model for small RNA biology. *eLife* **7**, e31628 (2018).

512 22. MarecKen, F., Sahara, S. & Traut, W. Rise and fall of the W chromosome in Lepidoptera. in
513 *Molecular biology and genetics of the Lepidoptera* (eds. MR, G. & F, M.) 49−63 (CRC, Boca Raton,
514 2010).
515 23. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly
516 using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170-175 (2021).
517 24. Roach, M.J., Schmidt, S.A. & Borneman, A.R. Purge Haplotigs: allelic contig reassignment for
518 third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460 (2018).
519 25. Rao, S.S.P. et al. A 3D map of the human genome at kilobase resolution reveals principles of
520 chromatin looping. *Cell* **162**, 687-688 (2015).
521 26. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome*
522 *Biology* **16**, 259 (2015).
523 27. Burton, J.N. et al. Chromosome-scale scaffolding of de novo genome assemblies based on
524 chromatin interactions. *Nature Biotechnology* **31**, 1119-1125 (2013).
525 28. Flynn, J.M. et al. RepeatModeler2 for automated genomic discovery of transposable element
526 families. *PNAS* **117**, 9451-9457 (2020).
527 29. Bao, Z. & Eddy, S.R. Automated de novo identification of repeat sequence families in sequenced
528 genomes. *Genome Res* **12**, 1269-76 (2002).
529 30. Price, A.L., Jones, N.C. & Pevzner, P.A. De novo identification of repeat families in large genomes.
530 *Bioinformatics* **21**, i351-i358 (2005).
531 31. J, J. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome*
532 *Research* **110**, 462-467 (2005).
533 32. Neumann, P., Novák, P., Hoštáková, N. & Macas, J. Systematic survey of plant LTR-
534 retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a
535 reference for element classification. *Mobile DNA* **10**, 1 (2019).
536 33. Wheeler, T.J. et al. Dfam: a database of repetitive DNA based on profile hidden Markov models.
537 *Nucleic Acids Research* **41**, D70-D82 (2012).
538 34. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long
539 terminal repeat retrotransposons. *Plant Physiology* **176**, 1410-1422 (2018).
540 35. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo
541 detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
542 36. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR
543 retrotransposons. *Nucleic Acids Research* **35**, W265-W268 (2007).
544 37. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic
545 sequences. *Curr. Protoc. Bioinform* **25**, 4.10.1-4.10.14 (2009).
546 38. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*
547 **27**, 573−580 (1999).
548 39. Beier, S., Thiel, T., Munch, T., Scholz, U. & Mascher, M. MISA-web: a web server for
549 microsatellite prediction. *Bioinformatics* **33**, 2583-2585 (2017).
550 40. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped
551 cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637-644 (2008).
552 41. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
553 42. Keilwagen, J. et al. Using intron position conservation for homology-based gene prediction. *Nucleic*
554 *Acids Research* **44**, e89-e89 (2016).
555 43. Kim, D., Langmead, B. & Salzberg, S.L. HISAT: a fast spliced aligner with low memory

requirements. *Nature Methods* **12**, 357-360 (2015).
44. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**, 290-295 (2015).
45. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Research* **43**, e78-e78 (2015).
46. Haas, B.J. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**, 5654-5666 (2003).
47. Grabherr, M.G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644-652 (2011).
48. Haas, B.J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol* **9**, R7 (2008).
49. Finn, R.D. Pfam: clans, web tools and services. *Nucleic Acids Research* **34**, D247-D251 (2006).
50. Boeckmann, B. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* **31**, 365-370 (2003).
51. 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).
52. Griffiths-Jones, S. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research* **33**, D121-D124 (2004).
53. Griffiths-Jones, S. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research* **34**, D140-D144 (2006).
54. Nawrocki, E.P. & Eddy, S.R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933-2935 (2013).
55. 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).
56. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Research* **14**, 988-995 (2004).
57. Emms, D.M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**, 238 (2019).
58. 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).
59. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540-52 (2000).
60. 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).
61. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**, 1586-1591 (2007).
62. Mendes, F.K., Vanderpool, D., Fulton, B. & Hahn, M.W. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* **36**, 5516-5518 (2020).
63. 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).
64. Ranwez, V., Harispe, S., Delsuc, F., Douzery, E.J.P. & Murphy, W.J. MACSE: multiple alignment of coding sequences accounting for frameshifts and stop codons. *PloS one* **6**, e22594-e22594 (2011).
65. Mank, J.E., Axelsson, E. & Ellegren, H. Fast-X on the Z: Rapid evolution of sex-linked genes in

birds. *Genome Research* **17**, 618-624 (2007).
66. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25−29 (2000).
67. Carbon, S. et al. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research* **49**, D325-D334 (2021).
68. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884−i890 (2018).
69. Mueller, J.C. et al. Micro germline-restricted chromosome in blue tits: evidence for meiotic functions. *Molecular Biology and Evolution* **40**, msad096 (2023).
70. Cheetham, S.W., Faulkner, G.J. & Dinger, M.E. Overcoming challenges and dogmas to understand the functions of pseudogenes. *Nature Reviews Genetics* **21**, 191−201 (2020).
71. Rogers, T.F., Pizzari, T. & Wright, A.E. Multi-copy gene family evolution on the avian W chromosome. *Journal of Heredity* **112**, 250-259 (2021).
72. 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).
73. 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).
74. Vítková, M., Fuková, I., Kubíčková, S. & Marec, F. Molecular divergence of the W chromosomes in pyralid moths (Lepidoptera). *Chromosome Research* **15**, 917-930 (2007).
75. Lukhtanov, V.A. Sex chromatin and sex chromosome systems in nonditrysian Lepidoptera (Insecta). *Journal of zoological systematics and evolutionary research* **38**, 73-79 (2000).
76. Bachtrog, D. et al. Sex determination: why so many ways of doing it. *PLoS biology* **12**, e1001899 (2014).