scientific data

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of a stored-product psocid, *Liposcelis tricolor* (Psocodea: Liposcelididae)

Shi-Die Jiang[1], Yu-Ang Chen[1], Sheng-Yuan Sun[1], Guy Smagghe[3,4,5], Jin-Jun Wang[1,2,3] & Dan-Dan Wei[1,2,3] ✉

Liposcelis tricolor **(Psocoptera: Liposcelididae) is a significant pest affecting stored products globally. However, due to the lack of a detailed genomic reference, the mechanisms of sex determination, stress resistance, and potential control methods for this booklouse remain poorly understood. In this study, the chromosome-level genome of *L. tricolor* was assembled by employing Illumina, Nanopore, and Hi-C sequencing technologies. The final genome size was determined to be 229.33 Mb, anchored to 9 pseudo-chromosomes. BUSCO analysis showed that 99.2% of complete BUSCOs were identified, suggesting the high completeness of the genome. A total of 91.49 Mb of repetitive sequences, accounting for 38.84% of the total genome, were annotated, and 15,647 protein-coding genes were predicted, with 88.17% functionally annotated. Additionally, we identified 25 typical sex-determining genes based on the genomic data. This high-quality genome assembly provides a crucial foundation for advancing our comprehension of the molecular biology, genetics, and potential control strategies for psocid *L. tricolor*.**

Background & Summary

Booklice, a stored-product pest of significant economic importance within the order Psocoptera, pose a serious threat to global grain and food security[1]. In recent years, controlling booklice has become increasingly challenging due to their rapid development of resistance, wide distribution, broad feeding range, ability to transmit pathogens, and strong reproductive capacity[2,3]. Currently, control of booklice primarily relies on contact insecticides and fumigants. However, under the circumstance of widespread insecticide resistance[4–8], there is an urgent need to identify new targets for environmentally friendly control methods. Among the booklice, *Liposcelis tricolor* is a representative species that is widespread and problematic in China[9].

Throughout evolution, insects have developed diverse sex determination systems, each governed by complex internal molecular regulatory mechanisms[10,11]. The study of these mechanisms has long been a global research focus. While current studies primarily center on model organisms, there is an increasing trend toward leveraging sex-determining genes for pest control[12]. Genetic engineering interventions aimed at manipulating insect sex for disrupting reproduction offer innovative solutions for pest control. For example, gene-editing technologies such as precision guided sterility insect technology (pgSIT) have been effectively applied to *Aedes aegypti*, providing a promising strategy for large-scale agricultural pest control[13]. Meanwhile, comprehensive analysis of sex determination mechanisms in pests could also shed light on their role in the evolution of pest resistance. Interestingly, a unique sex determination pattern, known as paternal genome elimination (PGE), was identified in a booklouse species (*Liposcelis* sp.)[14]. However, whether the PGE pattern is widespread in booklice remains to be investigated.

[1]Key Laboratory of Entomology and Pest Control Engineering, College of Plant Protection, Southwest University, Chongqing, 400715, China. [2]Key Laboratory of Agricultural Biosafety and Green Production of Upper Yangtze River (Ministry of Education), Academy of Agricultural Sciences, Southwest University, Chongqing, 400715, China. [3]Department of Plants and Crops, Ghent University, Ghent, Belgium. [4]Institute of Entomology, Guizhou University, Guiyang, 550025, China. [5]Cellular and Molecular Life Sciences, Department of Biology, Vrije Universiteit Brussel (VUB), Brussels, Belgium. ✉e-mail: wdd@swu.edu.cn

Library	Clean data(Gb)	Coverage(X)	Usage
Illumina	40.33	183.19	Genome survey
Nanopore	43.83	186.09	Scaffold assembly
Hi-C	27.41	119.52	Chromosome assembly
Illumina RNA-Seq	2.50		Gene structure annotation
Pacbio RNA-Seq	10.00		Gene structure annotation

Table 1. Statistics for the sequencing data of *L. tricolor* genome.

Type	Item	Feature
Genome survey	Genome size (Mb)	220.15
	Heterozygosity	0.36%
	GC content	33.64%
Assembled genome	Genome size (Mb)	229.33
	Chromosome number	9
	Contig number	110
	Contig N50 (Mb)	13.12
	GC content	33.90%
	BUSCO completeness	C:99.2% [S:98.0%, D:1.2%], F:0.5%, M:0.3%, n:1013

Table 2. Statistics for *L. tricolor* genome assembly.

Unfortunately, there is a significant gap in research on the sex determination mechanism in booklice, which hampers the application of emerging technologies like gene drive for their control. As sequencing technology advances, genomic data provide important basic data for entomological research. The assembly of high-quality genomes will be crucial for elucidating the molecular regulatory mechanisms of sex determination in booklice. Within booklice, only the genomes of *Liposcelis brunnea*[15] and *Liposcelis bostrychophila*[16] have been reported to date.

In this study, we used Illumina, Nanopore, and Hi-C sequencing technologies to assemble the chromosome-level genome of *L. tricolor*. The final genome size was determined to be 229.33 Mb, anchored to 9 pseudo-chromosomes, with a genome integrity of 99.2%. Genome annotation identified 15,647 protein-coding genes, 88.17% of which were functionally annotated. The genome assembly in this study provides a worthy foundation for evolutionary investigation of *L. tricolor* and the development of novel control strategies.

Methods

Sample collection and sequencing. *L. tricolor* was collected from a wheat warehouse in Heze, Shandong, China, in 2002 and has since been propagated in the laboratory using artificial diet. In brief, the rearing conditions were maintained at 27.5 ± 0.5 °C, 75–80% relative humidity, and total darkness. The artificial feed was composed of whole wheat flour, yeast powder, and skimmed milk powder in a ratio of 10:1:1.

Genomic DNA was extracted from approximately 1,000 female adults of *L. tricolor* using the Qiagen DNeasy extraction Kit (Qiagen, cat. Nos. 69504). For short-read sequencing, a 350 bp paired-end Illumina library was prepared and then sequenced on an Illumina NovaSeq platform (Illumina, CA, USA). For Nanopore sequencing, a large-insert (>10 kb) library was prepared following the protocol provided by the manufacturer and sequenced using the PromethION sequencer (Oxford Nanopore). For Hi-C sequencing, we constructed the Hi-C library according to the standard protocol[17] and sequenced on the Illumina HiSeq platform at Novogene Science and Technology Co., Ltd (Beijing, China). After removing the low-quality sequences, we obtained 40.33 Gb (183.19×) Illumina data, 43.83 Gb (186.09×) Nanopore data, and 27.41 Gb (119.52×) of the Hi-C data (Table 1). Total RNA was extracted from 500 adults of *L. tricolor* using the TRIzol reagent (Invitrogen, MA, USA). RNA-seq libraries for conventional transcriptome sequencing were constructed by Biomarker Technologies (Beijing, China) and then sequenced on the Illumina HiSeq 2000 platform, yielding 2.50 Gb of 150-bp paired-end reads (Table 1). For full-length transcriptome sequencing, RNA-seq libraries were generated by Novogene Science and Technology Co., Ltd (Beijing, China) using the SMRTbell™ Template Prep Kit 1.0 (Pacific Biosciences, USA). Sequencing was conducted on the PacBio Sequel II platform, resulting in 10.00 Gb of RNA-Seq data (Table 1).

Genome feature estimation. The short Illumina reads were quality-filtered by fastp[18] with the parameters '-q 10 -u 50 -y -g -Y 10 -e 20 -l 100 -b 150 -B 150' and then used for genome survey. We performed K-mer counting (k = 19) by Jellyfish (v2.2.10)[19] and calculated genome characteristics with Genomescope (v2.0)[20]. The genome size of *L. tricolor* was estimated to be 220.15 Mb, with 24.94% repeat sequence content and 0.36% heterozygosity (Table 2; Fig. 1A).

Draft Genome assembly and chromosome assembly. The long reads from Nanopore were used to assemble a contig-level genome of *L. tricolor*. After filtering adapters and low-quality reads, the data were error-corrected with Canu (v1.7.1)[21] and assembled using Smartdenovo[22] with default parameters. Pilon (v1.22)[23] was then employed for three rounds of polishing, followed by final error correction using short reads from

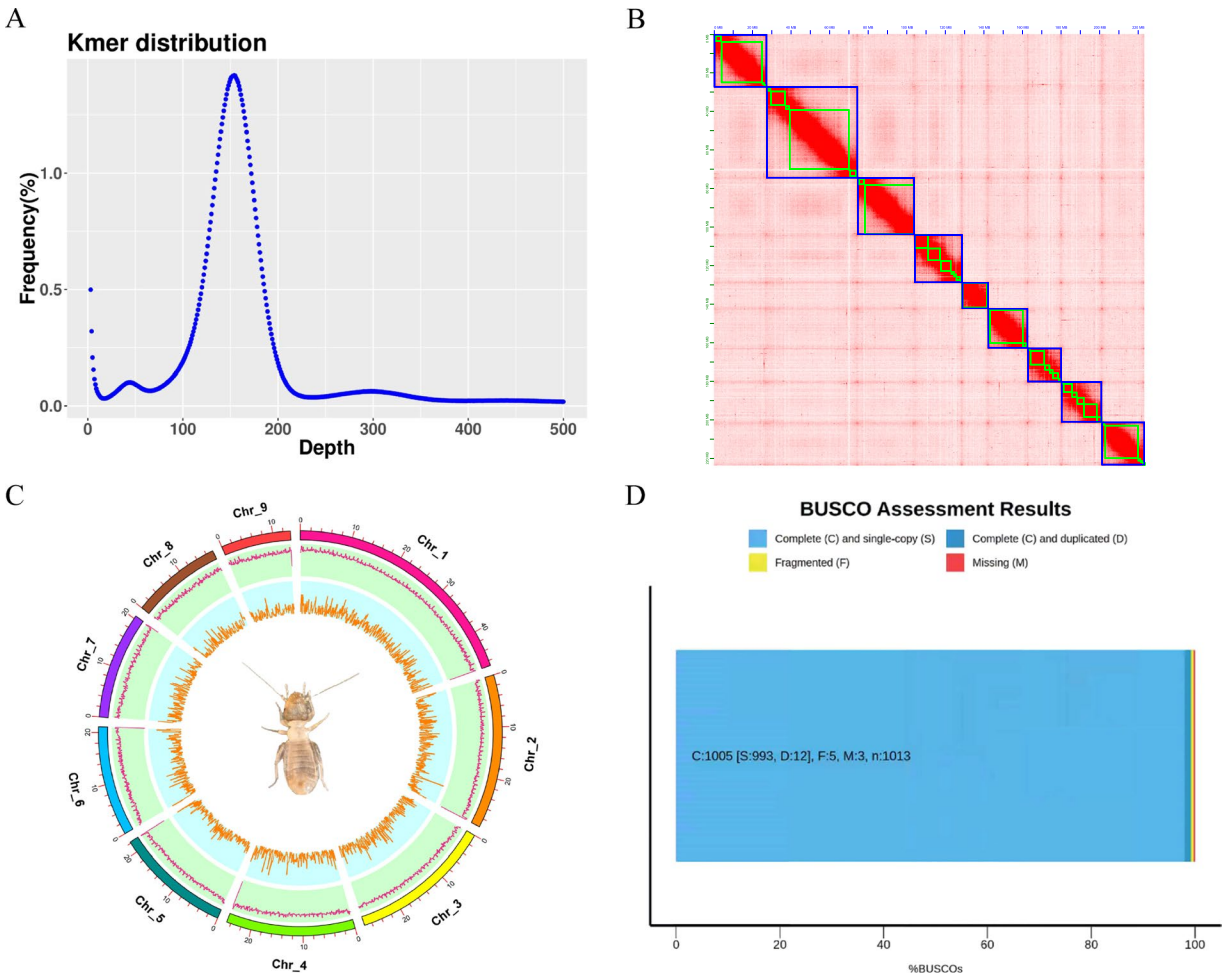

Fig. 1 Overview of the *L. tricolor* genome assembly and features. (**A**) GenomeScope estimation of genome size and heterogeneity using a k-mer of 19. (**B**) HI-C interactive heat map of *L. tricolor*. (**C**) Circos plot of the linkage groups. From outside to inside, it is chromosome length, GC content, and gene density. (**D**) BUSCO assessment of genome assembly integrity.

Illumina. The resulting draft genome was 235.54 Mb in size, consisting 110 contigs with a contig N50 of 13.12 Mb (Table 2).

The chromosome-level genome of *L. tricolor* was assembled using Hi-C sequencing data. After cleaning, the Hi-C reads were aligned to the contigs by Juicer (v1.9.9)[24] and then 110 contigs were anchored to pseudo-chromosomes using 3D-DNA[25]. Manual corrections were made using Juicebox (v1.11.08)[26]. This process successfully anchored the *L. tricolor* genome into 9 pseudo-chromosomes, with a final genome size was 229.33 Mb (Table 2; Fig. 1B,C). Genome completeness was assessed using BUSCO (v5.4.5)[27] with both the arthropoda_odb10 database and insecta OrthoDB v10 database. The results demonstrated that 99.2% and 98.76% of complete BUSCOs were included in this assembly, respectively (Table 2; Fig. 1D; Supplementary Table S1).

Repeat annotation and gene prediction. Transposon elements (TEs) were annotated using homology-based and *de novo* approaches. First, RepeatModeler (v2.0.3)[28], RECON (v1.08)[29], RepeatScout (v1.05)[30], LTR_retriever (v2.8)[31] and LTR_finder (v1.07)[32] were used to customize a *de novo* repeat library. The repeat libraries were then categorized using RepeatClassifier (v1.0)[33]. By merging the *de novo* TE sequences library with the existing databases, including Repbase (v19.06)[34], REXdb (v3.0)[35] and Dfam (v3.2)[36], a non-redundant species-specific TE library was constructed. RepeatMasker (v4.10)[37] was finally employed to identify and classify the TE sequences in the *L. tricolor* genome. In total, 91.49 Mb repeat sequences were identified, representing 38.84% of the assembled *L. tricolor* genome (Table 3). Among these, 59.33 Mb were TE sequences, representing 25.19% of the genome, and 32.16 Mb were tandem repeats, accounting for 13.65% of the assembled *L. tricolor* genome. Additionally, the proportion of repeat sequences in the *L. tricolor* genome (38.84%) exhibited a higher similarity with *L. bostrychophila* (37.88%)[16], compared to *L. brunnea* (15.9%)[15].

A thorough method was used to identify protein-coding genes in the *L. tricolor* genome, combining *de novo* prediction, homology search, and Unigenes from the transcriptome. *De novo* predictions were made by Augustus (v2.4)[38] and SNAP (v2006-07-28)[39]. For the homolog-based searching, GeMoMa (v1.7)[40] was employed using

Repeat type	Sequence type	number	Total length /bp	Percentage/%
Transposable elements	ClassI:Retroelement	105,510	22,770,499	9.67
	ClassI/DIRS	788	94,560	0.04
	ClassI/LINE	18,423	4,251,852	1.81
	ClassI/LTR/Cassandra	4	238	0.00
	ClassI/LTR/Caulimovirus	13	852	0.00
	ClassI/LTR/Copia	1,498	508,694	0.22
	ClassI/LTR/ERV	19,853	5,159,655	2.19
	ClassI/LTR/Gypsy	7,958	2,342,537	0.99
	ClassI/LTR/Pao	891	274,255	0.12
	ClassI/LTR/Unknown	52,632	9,576,286	4.07
	ClassI/LTR/Viper	10	2,369	0.00
	ClassI/SINE	3,440	559,201	0.24
	ClassII:DNA transposon	172,286	36,557,336	15.52
	ClassII/Academ	23	3,234	0.00
	ClassII/CACTA	5,738	1,079,952	0.46
	ClassII/Crypton	271	65,746	0.03
	ClassII/Dada	205	47,657	0.02
	ClassII/Ginger	171	52,455	0.02
	ClassII/Helitron	23,031	5,315,165	2.26
	ClassII/IS3EU	3,576	1,269,943	0.54
	ClassII/Kolobok	200	13,602	0.01
	ClassII/MITE	31	4,420	0.00
	ClassII/Maverick	497	276,841	0.12
	ClassII/Merlin	330	62,240	0.03
	ClassII/Mutator	1,491	201,718	0.09
	ClassII/Novosib	61	3,826	0.00
	ClassII/P	86	6,499	0.00
	ClassII/PIF-Harbinger	799	89,620	0.04
	ClassII/PiggyBac	484	133,110	0.06
	ClassII/Sola	74	4,794	0.00
	ClassII/Tc1-Mariner	2,593	818,468	0.35
	ClassII/Unknown	112,725	23,551,799	10.00
	ClassII/Zator	2,794	444,190	0.19
	ClassII/Zisupton	63	4,942	0.00
	ClassII/hAT	17,043	3,107,115	1.32
Tandem repeats	microsatellite(1–9 bp units)	360,331	7,054,330	3.00
	minisatellite(10–99 bp units)	10,699	16,041,633	6.81
	satellite(>=100 bp units)	8,162	9,061,682	3.85
Total		656,988	91,485,480	38.84

Table 3. Statistics of repeat sequences information in *L. tricolor* genome.

reference gene models from four other species (*Acyrthosiphon pisum, Drosophila melanogaster, L. bostrychophila,* and *Pediculus humanus*). Unigene prediction was carried out using PASA (v2.0.2)[41] to predict genes based on the Unigene assemblies without reference. Gene models generated from these diverse methods were consolidated by Evidence Modeler (EVM) (v1.1.1)[42] and further refined with PASA (v2.0.2)[41]. The predicted gene sequences were functionally annotated by searching the Gene Ontology (GO)[43], Kyoto Encyclopedia of Genes and Genomes (KEGG)[44], eukaryotic orthologous groups (KOG)[45], Pfam[46], SwissProt[47], TrEMBL[48], eggNOG[49], and GenBank Non-Redundant (NR) databases[50]. In total, 15,647 protein-coding genes (PCGs) were annotated (Table 4), with 88.17% (13,796 genes) having functional annotations in the *L. tricolor* genome (Table 5).

Synteny analysis. Synteny analysis of the chromosome-level genomes of *L. tricolor*, *L. bostrychophila,* and *L. brunnea* was conducted using TBtools[51]. The analysis employed the One Step MCScanX plugin for synteny detection, the Fasta Stats plugin for chromosome backbones generation, and the multiple synteny plot plugin for visualization. The results revealed several instances of chromosome fusion and division among the three species. The *L. tricolor* genome exhibited a higher degree of chromosomal synteny with *L. bostrychophila* compared to *L. brunnea* (Fig. 2). Specifically, except for chr9 of *L. bostrychophila*, which was not aligned, chr8 and 9 of *L. tricolor* corresponded to chr6 of *L. bostrychophila*, while the remaining chromosomes showed good collinearity and matched one-to-one.

Method of prediction	Software	Species	Gene number
Ab initio	Augustus	—	16,542
	SNAP	—	35,016
Homology-based	GeMoMa	*Acyrthosiphon pisum*	10,802
		Drosophila melanogaster	9,454
		Liposcelis bostrychophila	10,420
		Pediculus humanus	11,034
RNAseq	GeneMarkS-T	—	12,926
	PASA	—	12,306
Integration	EVM	—	15,647

Table 4. Statistics of gene prediction results.

Annotation database	Number of annotated genes	Proportion of annotated genes /%
GO_Annotation	11,904	76.08
KEGG_Annotation	11,795	75.38
KOG_Annotation	10,380	66.34
Pfam_Annotation	12,922	82.58
Swissprot_Annotation	11,716	74.88
TrEMBL_Annotation	11,654	74.48
EggNOG_Annotation	10,358	66.20
Nr_Annotation	11,862	75.81
All_Annotated	13,796	88.17

Table 5. Gene function annotation statistics information.

Sex-determining genes in *L. tricolor*. Based on the classification of "sex determination" in the NCBI BioSystem database, 30 insect sex determination genes were identified[52]. These genes are shown to regulate sex determination in model insects such as *D. melanogaster* and *B. mori*[53,54]. The amino acid sequences of the corresponding genes from *D. melanogaster* and *B. mori* were downloaded from the NCBI database and used as query sequences. Local BLAST searches were performed against the *L. tricolor* genome and transcriptome sequences, initially with a threshold of E-value $< 10^{-5}$. If no matches were found, the threshold was adjusted to 0.01 and BLAST was repeated to ensure no genes were missed. Following this, target sequences were manually inspected, and their conserved structural domains were analyzed using NCBI CD-search[55]. Molecular characterization parameters, such as molecular weight and isoelectric point, were calculated using the Expasy ProtParam tool[56].

A total of 25 typical insect sex-determining genes were identified in the *L. tricolor* genome and transcriptome (Supplementary Table S2). These include genes involved in the somatic cell sex determination pathway: *da*, *dpn*, *dsf*, *dsx*, *emc*, *fl(2)d*, *fru*, *gro*, *Imp*, *ix*, *PSI*, *run*, *snf*, *Sxl*, *tra-2*, and *vir*; Dose-compensated genes included *mle*, *mof*, *msl-1*, *msl-2*, and *msl-3*. Genes involved in germline sex determination are *otu*, *ovo*, *stat92E*, and *hop*. Chromosomal localization showed that 25 sex-determining genes are distributed across chr1, 2, 3, 5, 6 and 8 of *L. tricolor*, with the highest concentration on chr1, which contains 8 genes.

Data Records

All sequencing data has been uploaded to the NCBI SRA database: transcriptomic sequencing data (SRR30337849[57] and SRR30347065[58]), Illumina sequencing data (SRR30159735[59]), Nanopore sequencing data (SRR30167526[60]-SRR30167533[61–67]), and Hi-C sequencing data (SRR30169228[68]). The chromosome assembly is available in the GenBank database under the accession JBGGLZ000000000[69]. The genome annotations have been deposited in the Figshare database[70].

Technical Validation

Evaluation of the genome assembly. First, the quality of the *L. tricolor* assembly was evaluated using BUSCO (v5.4.5)[27] with arthropoda_odb10 and the insecta OrthoDB v10 databases, identifying 99.2% and 98.76% of the complete BUSCO genes, respectively (Table 2; Fig. 1D; Supplementary Table S1). Additionally, BWA (v0.7.15)[71] was employed to map Illumina short reads onto the assembled genome. Of these, 99.24% were aligned to the genome, with 97.58% properly mapped (Supplementary Table S3). These statistics indicated a high level completeness and accuracy of the *L. tricolor* genome assembly.

Evaluation of the genome annotation. BUSCO (v5.4.5)[27] with vertebrata database was employed to assess the completeness of the *L. tricolor* genome annotation. The analysis indicated that 96.12% of conserved orthologous genes were complete in the predicted PCGs, comprising 93.42% single-copy and 2.71% duplicated genes (Supplementary Table S4).

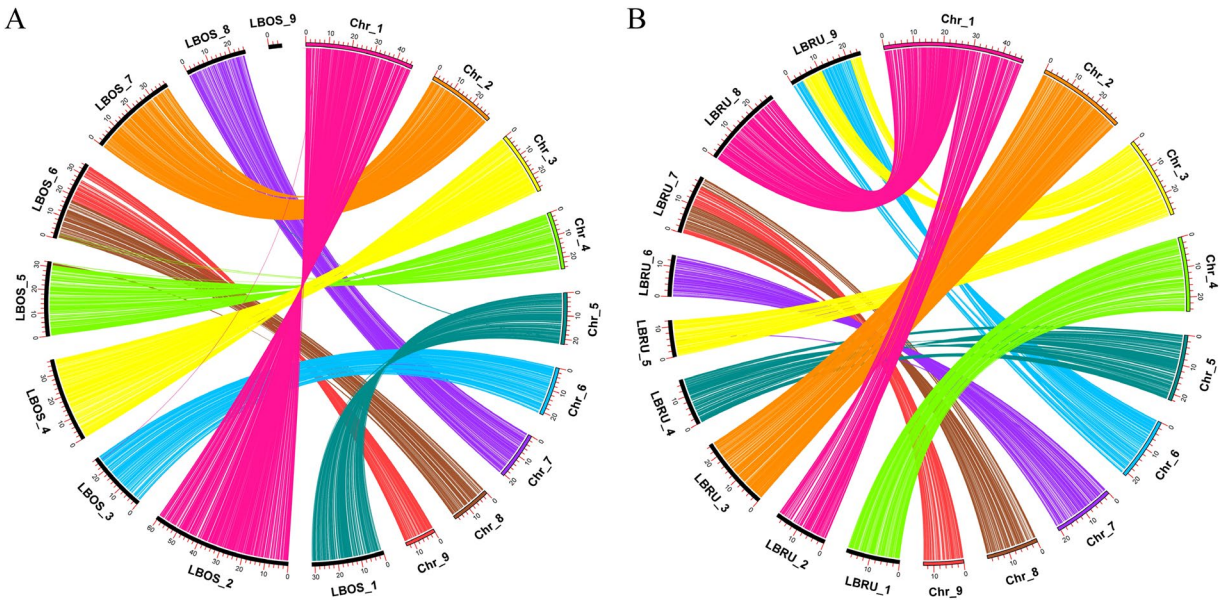

Fig. 2 Syntenic analysis. (**A**) Synteny blocks between *L. tricolor* and *L. bostrychophila* (LBOS). (**B**) Synteny blocks between *L. tricolor* and *L. brunnea* (LBRU).

Code availability

The versions of all software used in this study, along with their parameters, are detailed in the Methods section. The default parameters were utilized for software without detailed parameters. No custom code was employed in the generation or processing of the data described.

Received: 8 October 2024; Accepted: 26 November 2024;
Published online: 02 December 2024

References

1. Nayak, M. K., Collins, P. J., Throne, J. E. & Wang, J. J. Biology and management of psocids infesting stored products. *Annual Review of Entomology* **59**, 279–297, https://doi.org/10.1146/annurev-ento-011613-161947 (2014).
2. Mediannikov, O. *et al*. Booklice *Liposcelis bostrychophila* naturally infected by *Rickettsia felis* cause fever and experimental pneumonia in mammals. *The Journal of Infectious Diseases* **226**, 1075–1083, https://doi.org/10.1093/infdis/jiac282 (2022).
3. Guo, P. Y. *et al*. Identification and characterization of a novel *Iflavirus* in a stored-product psocid, *Liposcelis bostrychophila* (Psocodea: Liposcelididae). *Journal of Stored Products Research* **104**, 102185, https://doi.org/10.1016/j.jspr.2023.102185 (2023).
4. Nayak, M. K., Collins, P. J., Pavic, H. & Kopittke, R. A. Inhibition of egg development by phosphine in the cosmopolitan pest of stored products *Liposcelis bostrychophila* (Psocoptera: Liposcelididae). *Pest Management Science* **59**, 1191–1196, https://doi.org/10.1002/ps.753 (2003).
5. Wang, J. J. & Zhao, Z. M. Accumulation and utilization of triacylglycerol and polysaccharides in *Liposcelis bostrychophila* (Psocoptera, Liposcelididae) selected for resistance to carbon dioxide. *Journal of Applied Entomology* **127**, 107–111, https://doi.org/10.1046/j.1439-0418.2003.00718.x (2003).
6. Ren, Y., Wei, X. Q., Wu, S., Dou, W. & Wang, J. J. Comparison of acetylcholinesterase from three field populations of *Liposcelis paeta Pearman* (Psocoptera: Liposcelididae): Implications of insecticide resistance. *Pesticide Biochemistry and Physiology* **90**, 196–202, https://doi.org/10.1016/j.pestbp.2007.11.008 (2008).
7. Danso, J. K., Opit, G. P., Noden, B. H. & Giles, K. L. Estimating discriminating doses of phosphine for adults of eight species of psocids of genera Liposcelis (Psocodea: Liposcelididae) and Lepinotus (Psocodea: Trogiidae). *Journal of Stored Products Research* **99**, 102025, https://doi.org/10.1016/j.jspr.2022.102025 (2022).
8. Yang, B. B. *et al*. Involvement of Methoprene-tolerant and Krüppel homolog 1 in juvenile hormone-mediated vitellogenesis of female *Liposcelis entomophila* (End.) (Psocoptera: Liposcelididae). *Archives of Insect Biochemistry and Physiology* **112**, e21973, https://doi.org/10.1002/arch.21973 (2023).
9. Wei, D. D. *et al*. Sequence analysis of the ribosomal internal transcribed spacers region in psocids (Psocoptera: Liposcelididae) for phylogenetic inference and species discrimination. *Journal of Economic Entomology* **104**, 1720–1729, https://doi.org/10.1603/ec11177 (2011).
10. Baird, R. B., Mongue, A. J. & Ross, L. Why put all your eggs in one basket? Evolutionary perspectives on the origins of monogenic reproduction. *Heredity* **131**, 87–95, https://doi.org/10.1038/s41437-023-00632-7 (2023).
11. Bachtrog, D. *et al*. Sex Determination: Why so many ways of doing it? *PLOS Biology* **12**, e1001899, https://doi.org/10.1371/journal.pbio.1001899 (2014).
12. Siddall, A., Harvey-Samuel, T., Chapman, T. & Leftwich, P. T. Manipulating Insect Sex Determination Pathways for Genetic Pest Management: Opportunities and Challenges. *Frontiers in Bioengineering and Biotechnology* **10**, 867851, https://doi.org/10.3389/fbioe.2022.867851 (2022).
13. Li, M. *et al*. Targeting sex determination to suppress mosquito populations. *eLife* **12**, RP90199, https://doi.org/10.7554/eLife.90199 (2024).
14. Hodson, C. N. *et al*. Paternal Genome Elimination in Liposcelis booklice (Insecta: Psocodea). *Genetics* **206**, 1091–1100, https://doi.org/10.1534/genetics.117.199786 (2017).
15. Feng, S., Opit, G., Deng, W., Stejskal, V. & Li, Z. A chromosome-level genome of the booklouse, *Liposcelis brunnea*, provides insight into louse evolution and environmental stress adaptation. GigaScience 11, https://doi.org/10.1093/gigascience/giac062 (2022).

16. Jiang, S. D. *et al*. Chromosome-level genome of the parthenogenetic booklouse *Liposcelis bostrychophila* reveals high heterozygosity and a nonhomologous chromosome. Insect Science n/a, https://doi.org/10.1111/1744-7917.13387.
17. Rao, S. S. P. *et al*. A 3D Map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680, https://doi.org/10.1016/j.cell.2014.11.021 (2014).
18. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890, https://doi.org/10.1093/bioinformatics/bty560 (2018).
19. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).
20. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* **11**, 1432, https://doi.org/10.1038/s41467-020-14998-3 (2020).
21. Koren, S. *et al*. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**, 722–736, https://doi.org/10.1101/gr.215087.116 (2017).
22. Liu, H., Wu, S., Li, A. & Ruan, J. SMARTdenovo: a *de novo* assembler using long noisy reads. *GigaByte* **2021**, gigabyte15, https://doi.org/10.46471/gigabyte.15 (2021).
23. Walker, B. J. *et al*. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE* **9**, e112963, https://doi.org/10.1371/journal.pone.0112963 (2014).
24. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* **3**, 95–98, https://doi.org/10.1016/j.cels.2016.07.002 (2016).
25. Dudchenko, O. *et al*. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95, https://doi.org/10.1126/science.aal3327 (2017).
26. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Systems* **3**, 99–101, https://doi.org/10.1016/j.cels.2015.07.012 (2016).
27. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212, https://doi.org/10.1093/bioinformatics/btv351 (2015).
28. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
29. Bao, Z. & Eddy, S. R. Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res* **12**, 1269–1276, https://doi.org/10.1101/gr.88502 (2002).
30. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358, https://doi.org/10.1093/bioinformatics/bti1018 (2005).
31. Ou, S. & Jiang, N. LTR_retriever: A highly accurate and sensitive program for identification of Long Terminal Repeat retrotransposons. *Plant Physiology* **176**, 1410–1422, https://doi.org/10.1104/pp.17.01310 (2017).
32. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**, W265–W268, https://doi.org/10.1093/nar/gkm286 (2007).
33. Wicker, T. *et al*. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* **8**, 973–982, https://doi.org/10.1038/nrg2165 (2007).
34. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* **110**, 462–467, https://doi.org/10.1159/000084979 (2005).
35. 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**, 1, https://doi.org/10.1186/s13100-018-0144-1 (2019).
36. Wheeler, T. J. *et al*. Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Research* **41**, D70–D82, https://doi.org/10.1093/nar/gks1265 (2012).
37. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* **25**, 4.10.11–14.10.14, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
38. 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, https://doi.org/10.1093/bioinformatics/btn013 (2008).
39. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59, https://doi.org/10.1186/1471-2105-5-59 (2004).
40. Keilwagen, J. *et al*. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research* **44**, e89–e89, https://doi.org/10.1093/nar/gkw092 (2016).
41. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**, 5654–5666, https://doi.org/10.1093/nar/gkg770 (2003).
42. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biology* **9**, R7, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
43. Dimmer, E. C. *et al*. The UniProt-GO Annotation database in 2011. *Nucleic Acids Research* **40**, D565–D570, https://doi.org/10.1093/nar/gkr1048 (2011).
44. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27–30, https://doi.org/10.1093/nar/28.1.27 (2000).
45. Tatusov, R. L. *et al*. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research* **29**, 22–28, https://doi.org/10.1093/nar/29.1.22 (2001).
46. Finn, R. D. *et al*. Pfam: the protein families database. *Nucleic Acids Research* **42**, D222–230, https://doi.org/10.1093/nar/gkt1223 (2014).
47. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research* **28**, 45–48, https://doi.org/10.1093/nar/28.1.45 (2000).
48. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* **31**, 365–370, https://doi.org/10.1093/nar/gkg095 (2003).
49. Jensen, L. J. *et al*. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Research* **36**, D250–D254, https://doi.org/10.1093/nar/gkm796 (2007).
50. Marchler-Bauer, A. *et al*. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research* **39**, D225–D229, https://doi.org/10.1093/nar/gkq1189 (2010).
51. Chen, C. *et al*. TBtools-II: A "one for all, all for one" bioinformatics platform for biological big-data mining. *Molecular Plant* **16**, 1733–1742, https://doi.org/10.1016/j.molp.2023.09.010 (2023).
52. Liu, Y. *et al*. Genome-wide dissection of sex determination genes in the highly invasive whitefly species Bemisia tabaci Q/MED. *Insect Molecular Biology* **28**, 509–519, https://doi.org/10.1111/imb.12568 (2019).
53. Nagaraju, J., Gopinath, G., Sharma, V. & Shukla, J. N. Lepidopteran sex determination: a cascade of surprises. *Sexual Development* **8**, 104–112, https://doi.org/10.1159/000357483 (2014).
54. Schütt, C. & Nöthiger, R. Structure, function and evolution of sex-determining systems in Dipteran insects. *Development* **127**, 667–677, https://doi.org/10.1242/dev.127.4.667 (2000).
55. Wang, J. *et al*. The conserved domain database in 2023. *Nucleic Acids Research* **51**, D384–D388, https://doi.org/10.1093/nar/gkac1096 (2022).

56. Artimo, P. *et al*. ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Research* **40**, W597–W603, https://doi.org/10.1093/nar/gks400 (2012).
57. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30337849 (2024).
58. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30347065 (2024).
59. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30159735 (2024).
60. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30167526 (2024).
61. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30167527 (2024).
62. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30167528 (2024).
63. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30167529 (2024).
64. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30167530 (2024).
65. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30167531 (2024).
66. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30167532 (2024).
67. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30167533 (2024).
68. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30169228 (2024).
69. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBGGLZ000000000 (2024).
70. Jiang, S. D. Genome structure annotation of *Liposcelis tricolor*. *figshare* https://doi.org/10.6084/m9.figshare.26526388.v1 (2024).
71. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324 (2009).

Acknowledgements

This study received funding from the National Natural Science Foundation of China (32272528) and the Natural Science Foundation of Chongqing (CSTB2024NSCQ-MSX0374).

Author contributions

S.D.J. and D.D.W. coordinated the study and participated in its design. S.D.J., Y.A.C. and S.Y.S. performed experiments, analyzed the data and generated the figures. S.D.J. wrote the first draft of the manuscript. G.S., J.J.W. and D.D.W. revised and edited the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-04179-y.

Correspondence and requests for materials should be addressed to D.-D.W.