scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of *Odontothrips loti* Haliday (Thysanoptera: Thripidae)

Luo Yingning[1], Wei Shuhua[2], Dai Wenting[1], Miao Miao[1], Wang Ying[2], Zhang Rong[2] & Ban Liping[1]✉

As the predominant pest of alfalfa, *Odontothrips loti* Haliday causes great damages over the major alfalfa-growing regions of China. The characteristics of strong mobility and fecundity make them develop rapidly in the field and hard to be controlled. There is a shortage of bioinformation and limited genomic resources available of *O. loti* for us to develop novel pest management strategies. In this study, we constructed a chromosome-level reference genome assembly of *O. loti* with a genome size of 346.59 Mb and scaffold N50 length of 18.52 Mb, anchored onto 16 chromosomes and contained 20128 genes, of which 93.59% were functionally annotated. The results of 99.20% complete insecta_odb10 genes in BUSCO analysis, 91.11% short reads mapped to the ref-genome, and the consistent tendency among the thrips in the distribution of gene length reflects the quality of genome. Our study provided the first report of genome for the genus *Odontothrips*, which offers a genomic resource for further investigations on evolution and molecular biology of *O. loti*, contributing to pest management.

Background & Summary

Odontothrips loti Haliday (Thysanoptera: Thripidae) is a destructive, oligophagous pest that mainly feeds on leguminous crops, particularly alfalfa *Medicago sativa* L.[1,2]. As the predominant pest of alfalfa, in North China, the major alfalfa-growing region, *O. loti* can cause damage to 70%~100% of plants on average[3,4]. Thrips attack the entire life cycle of the host plants, causing the plants to wilt or stop growing and the leaves to turn dry (Fig. 1), which not only leads to severe yield and forage quality reductions but also exacerbates the spread of plant viruses[5–7]. Several features of thrips such as small body size, cryptic behavior, and high fecundity make them difficult to control.

Taking advantages of the low-cost of next generation sequencing (NGS) technology, researchers could identify functional genes related to virus transmission or pesticide resistance from the whole genome level through the construction of genome map, understand the evolution of pesticide resistance and virus transmission mechanisms, and control pest by gene regulation, making it possible to develop new pest management strategies[8–15]. As the genetic information of *O. loti* is still largely unknown currently, we aimed to disclose it for the development of novel *O. loti* control strategies.

In this study, we present a high-quality chromosome-level genome of *O. loti*, which was obtained using a combination of ONT long-read sequencing, Illumina short-read sequencing and chromosome conformation capture (Hi-C) technologies. Comparative genomic analysis was also performed on *O. loti* and another fourteen insect species to explore their phylogenetic relationship and genomic features. We provide the first genome assembly for a thrip in the Odontothrips genus to facilitate better understanding the genome evolution of thrips and developing novel control strategies for this important alfalfa pest.

Methods

Sample preparation. *Odontothrips loti* individuals were initially collected from the alfalfa field at Shangzhuang Experimental Station at the China Agricultural University (40°8'15"N, 116°11'18"E), and the colony was established and maintained for approximately 10 generations in the laboratory using the 'Zhongmu No.1' alfalfa at the temperature of 25 ± 1 °C, the relative humidity of 65 ± 5%, and the light: dark cycle of 16 h:8 h. The

[1]College of Grassland Science and Technology, China Agricultural University, Beijing, 100193, China. [2]Institute of Plant Protection, Ningxia Academy of Agriculture and Forestry Sciences, Yinchuan, 750002, China. ✉e-mail: liping_ban@163.com

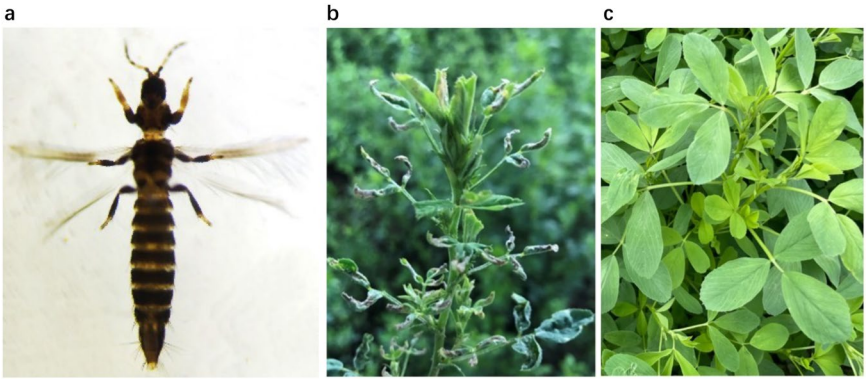

Fig. 1 *Odontothrips loti* (**a**), alfalfa with *O. loti* damage (**b**) and without *O. loti* damage (**c**).

Sample	Nymph /Adult	Sex	The number of thrips
DNA for survey	Adult	Female	1
DNA for assembly	Adult	Female and male	800
DNA for Hi-C	Adult	Female and male	800
RNA for annotation	Nymph and adult	Female and male	240

Table 1. Sample information of *Odontothrips loti* in this study.

Sequencing strategy	Platform	Usage	Insertion size	Clean data (Gb)	Coverage (X)
Short-reads	Illumina	Survey Assembly	150 bp	42.05	123
Long-reads	Oxford Nanopore	Assembly	10–20Kb	39.63	116
Hi-C	Illumina	Hi-C assembly	150 bp	31.78	93
RNA-seq	Oxford Nanopore	Annotation	1–15Kb	10.24	30

Table 2. Library sequencing data and methods used in this study to assemble the *Odontothrips loti* genome.

developmental stages of the thrips were examined under a light microscope. Individuals were collected, flash frozen in liquid nitrogen, and stored at −80 °C until use. Detailed information for *O. loti* sampling was shown in Table 1.

Genomic DNA sequencing. For Illumina short-read sequencing, the genomic DNA was isolated from of a single female adult following Chen's protocol[16], briefly, using sodium dodecyl sulfate (SDS) and proteinase K digestion, followed by phenol-chloroform extraction. The library (150 bp inserts) was constructed with Nextera DNA Flex Library Prep Kit (Illumina, San Diego, CA, USA), and sequenced on the Illumina NovaSeq 6000 (Illumina, San Diego, CA, USA), generating 43.66 Gb of raw data with 150 bp pair-end reads. Adapters and low-quality short reads were removed by Fastp (v0.21.0)[17] with default parameters, resulting in a total of 42.05 Gb (~123 × coverage) of clean data (Table 2). The short-read data was used for genome survey and assembly polish.

For long-read genomic DNA sequencing, we used approximately 800 mixed-sex adult thrips. Genomic DNA was extracted using the SDS method[16], and the DNA fragment size and the degree of degradation were checked on a 0.7% agarose gel. The purity and concentration of extracted DNA were determined with NanoDrop One (Thermo Fisher Scientific). The library was constructed with SQK-LSK109 kit (Oxford Nanopore Technologies, Oxford, UK) according to the manufacturer's instructions and sequenced on the Oxford Nanopore PromethION platform (Oxford Nanopore Technologies, Oxford, UK). We obtained 41.19 Gb (~120 × coverage) of raw long-read data with mean length of 6,182.26 bp (N50 = 16,150 bp). We then used Oxford Nanopore GUPPY (v0.3.0, https://timkahlke.github.io/LongRead_tutorials/BS_G.html) to filter reads with quality score < 7 and obtained 39.63 Gb (~116 × coverage) of clean reads. The cleaned long-read data were used for contig-level genome assembly (Table 2).

Hi-C library preparation and sequencing. The Hi-C sequencing library was prepared with 800 mixed-sex adult thrips. Samples were cross-linked with a 2% formaldehyde isolation buffer and then treated with DpnII (New England Biolabs, Beijing, CN) to digest nuclei. Biotinylated nucleotides were used to repair tails, and the ligated DNA was split into fragments of 300–700 bp in length. The resulting Hi-C library was sequenced in Illumina Novoseq 6000 for 150 bp paired-end reads. After applying the same filter criteria for short reads, a total of 31.78 Gb (~93 × coverage) of clean data was generated to assist the chromosome-level assembly (Table 2).

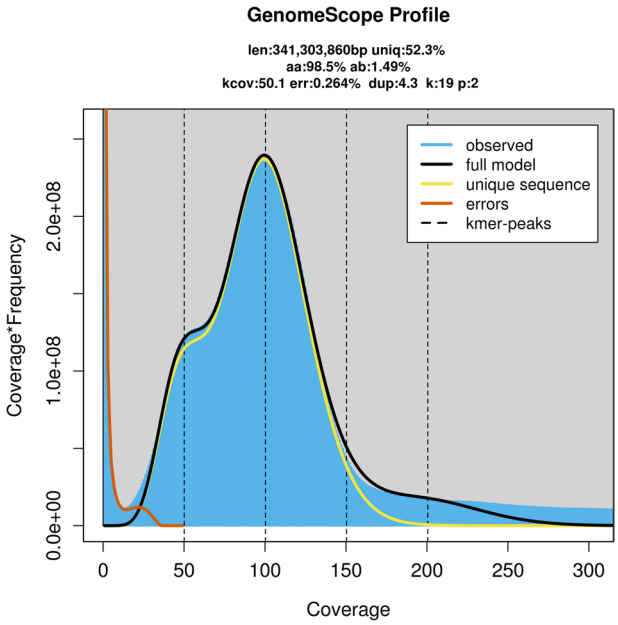

Fig. 2 Characteristics of the Illumina short-read sequencing of the *Odontothrips loti* genome.

ONT-Transcriptome sequencing. For ONT-transcriptome sequencing, approximately 240 thrips including nymph and adult were mixed for RNA extraction with the RNA Easy Fast Tissue/Cell Kit (Tiangen). NanoDrop (Thermo Fisher Scientific) and Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) were used to evaluate the quality of extracted RNA. SQK-PCS109 and SQK-PBK004 kit (Oxford Nanopore Technologies) were used for reverse transcript and construction of cDNA library, and sequencing was proceeded on the PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK). A total of 10.24 Gb of clean reads were generated with mean length of 1,034.61 bp (N50 = 1,238 bp), used to assist genome annotation (Table 2).

Estimation of genomic characteristics. Genomic characteristics were estimated based on 42.05 Gb of short-read data using a K-mer-based statistical analysis in Jellyfish (v2.3.0)[18] and GenomeScope2[19] (p = 2, k = 19). Based on 19-mer depth analysis, the genome size and heterozygosity were estimated to be 341.3 Mb and 1.49%, respectively, therefore, this genome is considered highly heterozygous (Fig. 2).

Genome assembly. *Contig level assembly.* We first used NextDenovo (v2.5.0)[20] to generate a draft assembly, and conducted two rounds of polish with ONT long reads on Racon (v1.4.11, https://github.com/lbcb-sci/racon). Illumina reads were mapped to the assembly using BWA v0.7.17 and another two rounds of contig polishing were performed with Pilon (v1.23)[21]. Owing to its highly heterozygous feature, Purge_haplotigs (v1.0.4, https://github.com/skingan/purge_haplotigs_multiBAM) was applied to de-heterozygosis the draft genome to generate the final contig-level genome, which was 346.58-Mb long and similar to the estimated size, with the N50 contig length of 8.59 Mb (Table 3).

Hi-C scaffolding. Low-quality raw reads (quality score <20,length shorter than 30 bp) and adaptors were removed using Fastp (v0.21.0)[17]. The clean reads were then mapped to the contig assembly using HICUP (v0.8.0)[22] to filter unmapped reads, invalid pairs, dangling end and repeats resulting from PCR amplification. The valid paired-end pairs were used for contig cluster, order and orient by ALLHIC (v0.9.8)[23]. The interaction between contig pairs were converted into binary files by 3D-DNA[24] and Juicer (v1.6)[25]. The HiCExplorer (v3.6)[26] was used to generate the heat maps of contig interaction intensity and location. The Juicebox (v1.11.08)[27] was subsequently employed to review assembly manually. In summary, the resulting chromosome-level genome length was 346.59 Mb with a scaffold N50 of 18.52 Mb (Table 3), around 86.93% (301.28 Mb) of the genome bases were anchored onto 16 chromosomes (Fig. 3a), and most syntenic blocks of genome presents in the low GC content region (Fig. 3b).

Predicting repeats. We used ReaptModeler (v.1.0.11, https://github.com/Dfam-consortium/RepeatModeler) to predict repeat sequence. LTR_FINDER (vOfficial, -size 1000000 -time 300)[28] and LTR_retriever (v2.9.0)[29] were used to find and de-redundant the LTR sequence. These two de novo library were combined with RepBase[30] for further prediction by RepeatMasker (v4.0.9,-nolow -no_is -norna)[31]. RepeatProteinMask (-noLowSimple -pvalue 0.0001) was used for homo-prediction. All results were de-redundant and merged to the final repeat sequence. In summary, 115.26 Mb repeat sequences were identified, accounting for 33.26% of the *O. loti* genome (Table 4). Among these repeat sequences, most (18.85%) are DNA transposon,

Features	Values
Estimated genome size (bp)	341,303,860
Contig-level assembly size (bp)	346,577,358
Chromosome-level assembly size (bp)	346,592,158
Anchored to chromosome (bp)	301,277,358
Contig N50(bp)	8,588,564
Scaffold N50(bp)	18,519,078

Table 3. Major indicators of the *Odontothrips loti* genome.

Type	Length (bp)	Percentage in genome (%)
DNA	65,317,630	18.85
LTR	35,092,753	10.13
LINE	11,957,062	3.45
SINE	1,382,412	0.40
Unknown	14,723,706	4.25
Total	115,261,572	33.26

Table 4. Statistics of the repeat sequences annotation in *Odontothrips loti* genome.

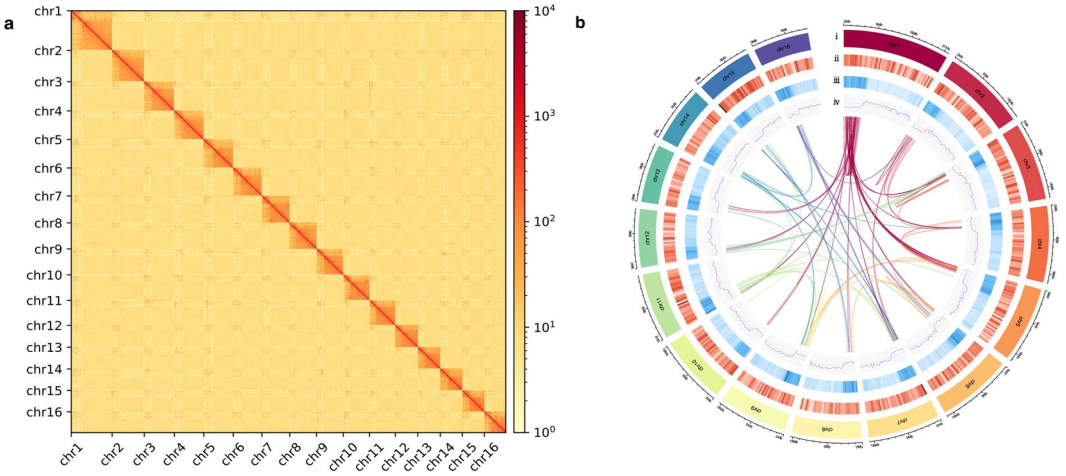

Fig. 3 Heatmap of genome-wide Hi-C data and circular representation of the chromosomes of *Odontothrips loti*. (**a**) The heatmap of chromosome interactions in *O. loti*. The frequency of Hi-C interaction links is represented by colors, which ranges from yellow (low) to red (high). (**b**) Circos plot of distribution of the genomic elements in *O. loti*. The tracks indicate (i) length of the chromosome, (ii) gene density, (iii) distribution of transposable element (TE) density, and (iv) GC density. Center: intra-genomic syntenic blocks of *O. loti*. The densities of genes, TEs, and GC were calculated in 500 kb windows.

followed by 10.13% of long terminal repeats (LTRs), 3.45% of long interspersed nuclear elements (LINEs) and only 0.40% of short interspersed nuclear elements (SINEs) (Table 4).

Protein-coding genes and functional predictions. We utilized a pipeline include three strategies: transcriptome-based prediction, homology-based prediction, and ab initio prediction to annotate protein coding genes. For transcriptome-based prediction, we use NanoFilt (v2.8.0, -q 7 -l 100 -headcrop 30 -minGC 0.3)[32], Pychopper (v2.7.2, https://github.com/epi2me-labs/pychopper), racon (v1.4.11, https://github.com/lbcb-sci/racon), minimap2 (v2.17-r941)[33], stringtie (v2.1.4)[34] and TransDecoder (v5.1.0, https://github.com/TransDecoder/TransDecoder) for ONT-transcriptome reads to predicted protein-coding gene. For homology-based prediction, tblastn (v2.7.1)[35] with an E-value cutoff of 1e-5 and Exonerate (v2.4)[36] were used to predict gene structure by comparing with 3 closely related species (*Megalurothrips usitatus*, *Thrips palmi*, *Frankliniella occidentalis*) and model species *Drosophila melanogaster*. Before ab initio prediction, repetitive elements from the whole genome were soft-masked. Augustus (v3.3.2)[37], GenScan (v1.0)[38] and GlimmerHMM (v3.0.4)[39] were used for de novo prediction. Finally, MAKER (v2.31.10)[40] integrated the above three strategies, resulting in a non-redundant gene set, with weighting as default. Overall, 20,128 protein coding genes were obtained (Table 5).

Database	Number	Percentage (%)
Protein-coding genes	20,128	100.00
Annotated genes	18,837	93.59
Interproscan	17,895	88.91
NR	16,363	81.29
Uniprot	16,241	80.69
Pfam	13,932	69.22
GO	12,229	60.76
KEGG	8,527	42.36
Pathway	4,801	23.85
Unanotated genes	1,291	6.41

Table 5. Statistics for the *Odontothrips loti* functionally annotated protein-coding genes.

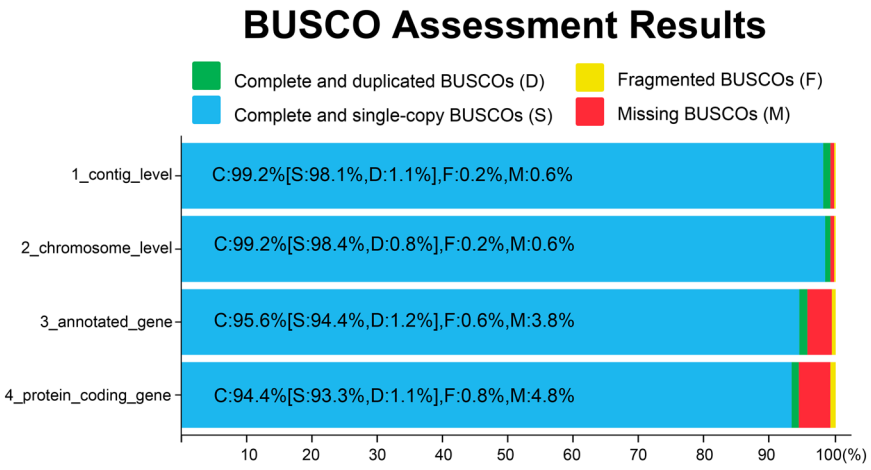

Fig. 4 Benchmarking of genome completeness of *Odontothrips loti* genome assembly and annotation, evaluated by BUSCO based on insect_odb10 database which includes 1,367 genes. C: the number of complete genes, S: the number of complete and single-copy genes, D: the number of complete and duplicated genes, F: the number of incomplete genes, M: the number of missing genes.

For functional annotation, protein sequences were aligned to Non-Redundant protein (NR), Universal Protein (Uniprot), Protein Families Analysis and Modeling (Pfam), Clusters of Orthologous Groups of proteins (COG), Kyoto Encyclopedia of Genes and Genomes (KEGG) and evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG) database. Gene Ontology (GO) terms was obtained from Uniport. InterProScan (v5.52-86.0)[41] was used to search the conserved sequences, motifs and domains. There were 12,229 (60.76%) and 8,527 (42.36%) genes annotated to GO terms and KEGG pathways respectively. A total of 18,837 genes (93.59%) were annotated using at least one public database (Table 5).

Data Records

The assembly genome sequence and annotation data were deposited in Figshare[42] and GenBank[43]. Raw data from Nanopore (CRR997575)[44], Illumina (CRR997573)[45] and Hi-C (CRR997574)[46] genome sequencing and RNA-seq (CRR997576)[47] were deposited in the Genome Sequence Archive (GSA, https://ngdc.cncb.ac.cn/gsa)[48], and were related to the BioProject PRJCA022165.

Technical Validation

Genome quality assessment. We assessed the quality of chromosome-level genome from the three aspects: continuity, consistency, and completeness. First, the scaffold N50 of *O. loti* is 18.52 Mb (Table 3), representing the continuity of genome. Second, we evaluated the consistency of the genome by calculating the comparison rate and coverage of Illumina reads through BWA (v0.7.17)[49], resulting 91.11% short reads were aligned to and covered 94.68% of the ref-genome. Third, we used BUSCO (v4.1.4)[50] to estimate the completeness of chromosome-level genome by searching the 1367 BUSCO genes in insecta_odb10 (https://busco-data.ezlab.org/v5/data/lineages/). The results showed a high completeness level with 99.2%, 99.2%, 95.6%, 94.4% complete genes found in the contig-level genome, chromosome-level genome, annotated gene sets and protein-coding gene sets, respectively (Fig. 4).

Evaluation of gene prediction. To verify the accuracy and reliability of the gene prediction, we determined the distribution of gene length, CDS length, exon length and intron length in *O. loti*, *D. melanogaster*[51] and

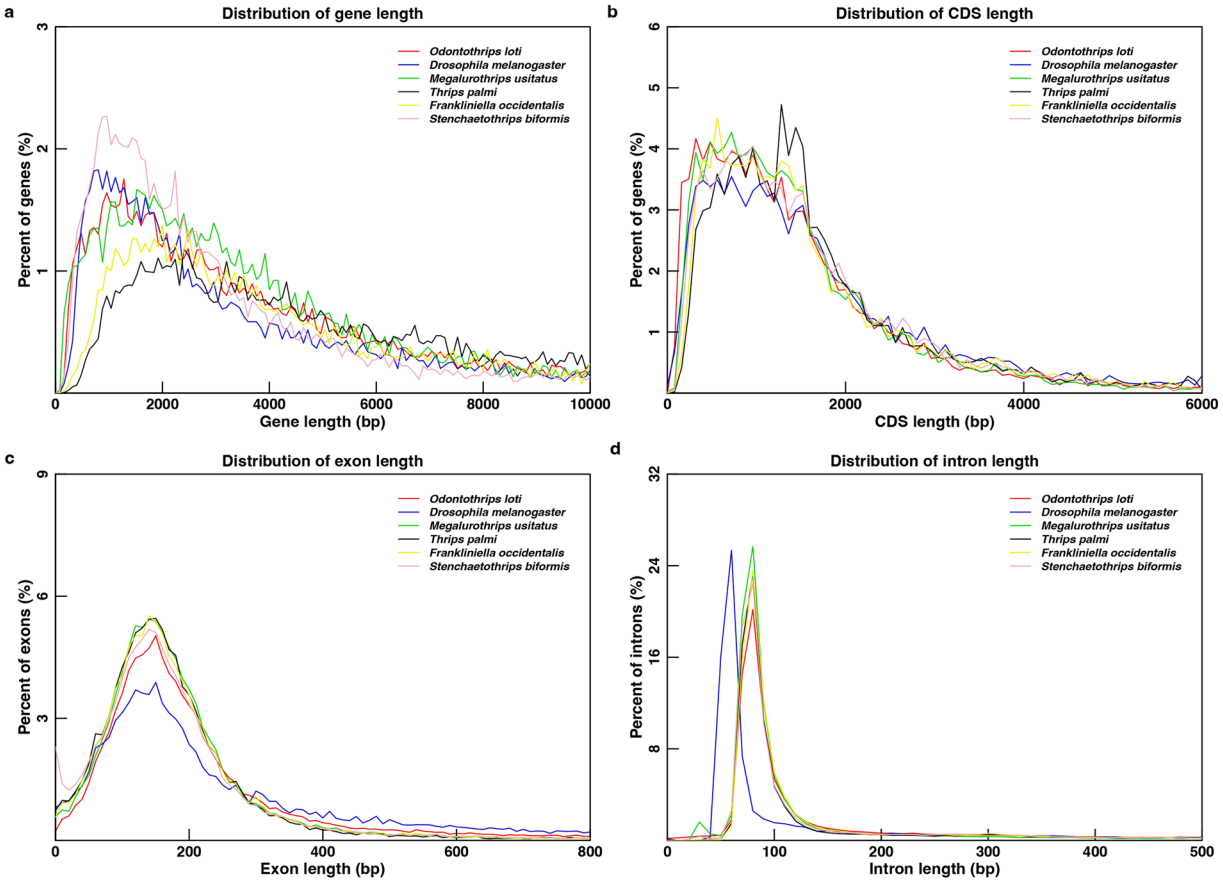

Fig. 5 Annotated genes comparison of the distribution of (**a**) gene length (**b**) CDS length (**c**) exon length (**d**) intron length in *Odontothrips loti* with *Drosophila melanogaster* and four closely related species. The x-axis represents the length, and the y-axis represents the density of genes.

other four related species (*M. usitatus*[8], *T. palmi*[12], *F. occidentalis*[14], *S. biformis*[13]). The consistent tendency among the thrips supported an ideal annotated gene dataset in *O. loti* (Fig. 5).

Code availability

All software and pipelines were executed according to the manual and protocols of the published bioinformatic tools. The version and code/parameters of software have been described in Methods section. No custom code was used.

Received: 22 December 2023; Accepted: 22 April 2024;
Published online: 04 May 2024

References

1. Liu, Y., Luo, Y., Du, L. & Ban, L. Antennal Transcriptome Analysis of Olfactory Genes and Characterization of Odorant Binding Proteins in *Odontothrips loti* (Thysanoptera: Thripidae). *Int J Mol Sci* **24**, 5284 (2023).
2. Liu, Y., Li, J. & Ban, L. Morphology and Distribution of Antennal Sensilla in Three Species of Thripidae (Thysanoptera) Infesting Alfalfa *Medicago sativa*. *Insects* **12**, 81 (2021).
3. Buhe, T. & Wang, X. Breeding research of the variety of anti-thrips alfalfa. *Multifunctional Grasslands In A Changing World, Volume Ii Xxi International Grassland Congress And Viii International Rangeland Congress, Hohhot, China* **29 E 5 Y**, 5–5 (2008).
4. Li, N., Song, X. & Wang, X. The complete mitochondrial genome of *Odontothrips loti* (Haliday, 1852) (Thysanoptera: Thripidae). *Mitochondrial DNA B Resour* **5**, 7–8 (2019).
5. Wu, S. *et al*. A decade of a thrips invasion in China: lessons learned. *Ecotoxicology* **27**, 1032–1038 (2018).
6. Li, J. *et al*. Occurrence, Distribution, and Transmission of Alfalfa Viruses in China. *Viruses* **14**, 1519 (2022).
7. Li, J. *et al*. RNA-seq reveals plant virus composition and diversity in alfalfa, thrips, and aphids in Beijing, China. *Arch Virol* **166**, 1711–1722 (2021).
8. Ma, L. *et al*. Chromosome-level genome assembly of bean flower thrips *Megalurothrips usitatus* (Thysanoptera: Thripidae). *Sci Data* **10**, 252 (2023).
9. Bao, W., Kataoka, Y., Fukada, K. & Sonoda, S. Imidacloprid resistance of melon thrips, *Thrips palmi*, is conferred by CYP450-mediated detoxification. *J. Pestic. Sci.* **40**, 65–68 (2015).
10. Shi, P. *et al*. Variable resistance to spinetoram in populations of *Thrips palmi* across a small area unconnected to genetic similarity. *Evolutionary Applications* **13**, (2020).
11. Xue, B. & Sonoda, S. Resistance to cypermethrin in melon thrips, *Thrips palmi* (Thysanoptera: Thripidae), is conferred by reduced sensitivity of the sodium channel and CYP450-mediated detoxification. *Applied Entomology and Zoology* **47**, (2012).

12. Guo, S. *et al*. Chromosome-level assembly of the melon thrips genome yields insights into evolution of a sap-sucking lifestyle and pesticide resistance. *Molecular Ecology Resources* **20**, 1110–1125 (2020).
13. Hu, Q., Ye, Z., Zhuo, J., Li, J.-M. & Zhang, C. A chromosome-level genome assembly of *Stenchaetothrips biformis* and comparative genomic analysis highlights distinct host adaptations among thrips. *Commun Biol* **6**, 1–10 (2023).
14. Rotenberg, D. *et al*. Genome-enabled insights into the biology of thrips as crop pests. *BMC Biology* **18**, (2020).
15. Zhang, Z. *et al*. The Chromosome-Level Genome Assembly of Bean Blossom Thrips (*Megalurothrips usitatus*) Reveals an Expansion of Protein Digestion-Related Genes in Adaption to High-Protein Host Plants. *Int J Mol Sci* **24**, (2023).
16. Chen, H., Rangasamy, M., Tan, S. Y., Wang, H. & Siegfried, B. D. Evaluation of Five Methods for Total DNA Extraction from Western Corn Rootworm Beetles. *PLoS ONE* **5**, e11963 (2010).
17. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
18. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k* -mers. *Bioinformatics* **27**, 764–770 (2011).
19. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* **11**, 1432 (2020).
20. Hu, J. *et al*. An efficient error correction and accurate assembly tool for noisy long reads. Preprint at https://doi.org/10.1101/2023.03.09.531669 (2023).
21. Walker, B. J. *et al*. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE* **9**, e112963 (2014).
22. Wingett, S. *et al*. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res* **4**, 1310 (2015).
23. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).
24. Dudchenko, O. *et al*. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
25. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–98 (2016).
26. Wolff, J. *et al*. Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Res* **48**, W177–W184 (2020).
27. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**, 99–101 (2016).
28. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* **35**, W265–268 (2007).
29. Ou, S. & Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol* **176**, 1410–1422 (2018).
30. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11 (2015).
31. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* Chapter 4, 4.10.1–4.10.14 (2009).
32. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
33. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).
34. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol* **20**, 278 (2019).
35. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
36. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
37. 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).
38. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**, 78–94 (1997).
39. Delcher, A. L., Bratke, K. A., Powers, E. C. & Salzberg, S. L. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673–679 (2007).
40. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
41. Blum, M. *et al*. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* **49**, D344–D354 (2021).
42. Luo, Y. Chromosome-level reference genome assembly of *O. loti*. *figshare* https://doi.org/10.6084/m9.figshare.24865023.v2 (2024).
43. Luo, Y. & Ban, L. Chromosome-level genome assembly of *Odontothrips loti* Haliday (Thysanoptera: Thripidae). *GenBank* https://identifiers.org/ncbi/insdc:JAZGLN000000000 (2024).
44. *NGDC Genome Sequence Archive (GSA)*. https://ngdc.cncb.ac.cn/gsa/browse/CRA014018/CRR997575 (2024).
45. *NGDC Genome Sequence Archive (GSA)*. https://ngdc.cncb.ac.cn/gsa/browse/CRA014018/CRR997573 (2024).
46. *NGDC Genome Sequence Archive (GSA)*. https://ngdc.cncb.ac.cn/gsa/browse/CRA014018/CRR997574 (2024).
47. *NGDC Genome Sequence Archive (GSA)*. https://ngdc.cncb.ac.cn/gsa/browse/CRA014018/CRR997576 (2024).
48. Chen, T. *et al*. The Genome Sequence Archive Family: Toward Explosive Data Growth and Diverse Data Types. *Genomics, Proteomics & Bioinformatics* **19**, 578–583 (2021).
49. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://doi.org/10.48550/arXiv.1303.3997 (2013).
50. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* **1962**, 227–245 (2019).
51. Hoskins, R. A. *et al*. The Release 6 reference sequence of the *Drosophila melanogaster* genome. *Genome Res* **25**, 445–458 (2015).

Acknowledgements

This work was supported by National Natural Science Foundation of China (no. 31971759 to B.L.), the Beijing Innovation Consortium of Modern Agricultural Industry Technology System (no. BAIC02-2024 to B.L.) and the Ningxia Province Sci-Tech Innovation Demonstration Program of High-Quality Agricultural Development and Ecological Conservation (no. NGSB-2021-15-04 to W.S.). We are grateful to Chaoyang Zhao (National Soil Dynamics Laboratory, USDA-ARS, Auburn, AL, USA) for guidance to improve the language of manuscript. The bioinformatics analysis is supported by High-performance Computing Platform of China Agricultural University.

Author contributions

B.L. conceived of this project. L.Y. and D.W. participated in the data analysis. L.Y., D.W., M.M., W.S., W.Y. and Z.R. collected the samples. L.Y. wrote the manuscript. L.Y. and B.L. revised the manuscript. All authors have read, revised, and approved the final manuscript for submission.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to B.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.