



OPEN

DATA DESCRIPTOR

A chromosome-level genome assembly of the dried fruit mite *Carpoglyphus lactis*

Lang Liang^{1,2}, Hu-Die He^{1,2} , Dao-Chao Jin¹ & Tian-Ci Yi¹

Carpoglyphus lactis is a globally distributed stored-product pest mite that inflicts substantial damage on dried fruits, grains, and other preserved commodities. However, the absence of high-quality genomic resources has long constrained insights into its biology and evolutionary characteristics. In this study, we report a chromosome-level genome assembly for *C. lactis*, generated using an integrative approach that combines Illumina, PacBio, Oxford Nanopore, and Hi-C sequencing technologies. The assembled genome comprises approximately 53.47 Mb, featuring a contig N50 of 713.31 kb and a scaffold N50 of 5.92 Mb. Utilizing Hi-C data, 99.8% of the genome sequences were anchored onto nine chromosomes. Gene annotation revealed a total of 10,330 protein-coding genes, with repetitive sequences accounting for 9.66% of the genome. BUSCO completeness assessment using the arachnida_odb12 database (n = 1,123) indicated high integrity for both the genome assembly (94.6%) and gene annotation (94.7%). This chromosome-level genome serves as a valuable resource for investigating mite biology and facilitating the development of effective control strategies against this economically important pest species.

Background & Summary

Astigmatic mites represent one of the most diverse and speciose lineages within the subclass Acari, encompassing 77 families and over 6,000 described species¹. Traditionally, they are thought to have originated from Oribatid mites in soil ecosystems². However, in contrast to their ancestral forms, astigmatid mites display exceptional ecological adaptability, allowing them to colonize a wide array of fragmented and transient habitats^{3,4}. Furthermore, many species employ a dispersal strategy known as phoresy, wherein they attach to other organisms for transportation^{5,6}. Their pronounced ecological versatility and evolutionary success make astigmatid mites an excellent model system for investigating adaptive evolution within the Acari.

The genus *Carpoglyphus*, belonging to the family Carpglyphidae, includes six valid species, most of which are recognized as economically important pests associated with stored products⁷. Among them, *C. lactis* (Linnaeus, 1767), commonly referred to as the dried fruit mite, is a cosmopolitan species reported across a wide variety of food commodities, including dried fruits, beer, dairy products, jams, honey, and wine^{8,9}. Under the warm and humid conditions prevalent during summer and autumn, this mite proliferates rapidly, leading to significant economic losses in the food processing and sugar industries. Furthermore, *C. lactis* can be transmitted to humans via multiple routes, triggering a spectrum of clinical symptoms, such as mite-induced dermatitis, allergic alveolitis, and food mite syndrome¹⁰. Despite its agricultural and medical relevance, a high-quality reference genome has been lacking, limiting insights into its evolutionary biology and impeding the development of effective management strategies.

In this study, we report a high-quality chromosome-level genome assembly for *Carpoglyphus lactis*. We employed a hybrid sequencing and assembly approach integrating PacBio long-read sequencing for de novo assembly, Illumina short-read data for error correction and polishing, and Hi-C chromatin interaction data to reconstruct chromosome-scale scaffolds. Gene annotation was performed using a comprehensive pipeline that combined ab initio prediction, homology-based evidence from known protein sequences, and transcriptomic data obtained from both Illumina RNA-seq and Nanopore-based full-length transcript sequencing. This

¹Institute of Entomology, Guizhou University, and The Provincial Key Laboratory for Plant Pest Management of Mountainous Region, Scientific Observing and Experimental Station of Crop Pest in Guiyang, Ministry of Agricultural and Rural Affairs, Guiyang, 550025, P. R. China. ²These authors contributed equally: Lang Liang, Hu-Die He. e-mail: dcjin@gzu.edu.cn; tcyi@gzu.edu.cn

Libraries	Clean data (Gb)	Sequencing coverage (×)	SRA accession number
WGS	11.4	200.41	SRR33763638
Hi-C	8.3	130.08	SRR33763639
HiFi	5.5	98.41	SRR33763640
RNA-sr	7.7	—	SRR33763641
RNA-ONT	9.7	—	SRR33763642

Table 1. Summary of sequencing data used for genome assembly and annotation of *C. lactis*. Note: WGS: Whole Genome Sequencing for genome assembly. Hi-C: High-throughput Chromosome Conformation Capture for chromosome-level scaffolding. HiFi: PacBio High-Fidelity long-read sequencing for improving assembly contiguity and accuracy. RNA-sr: Short-read RNA sequencing for gene annotation. RNA-ONT: Oxford Nanopore Technologies long-read RNA sequencing for gene annotation.

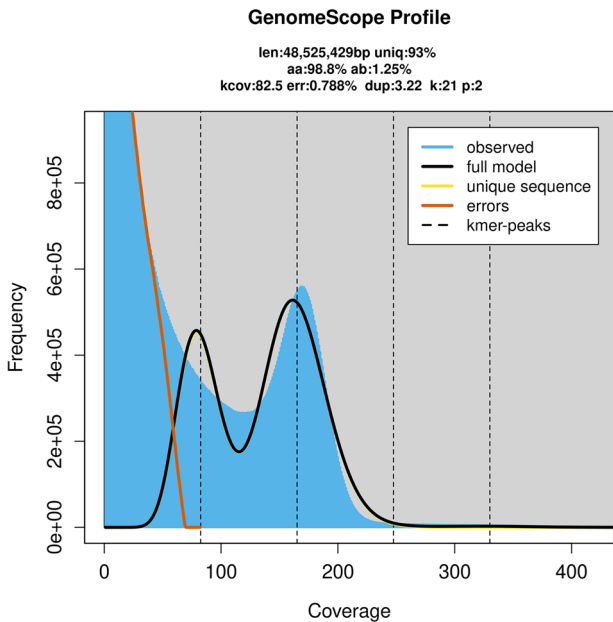


Fig. 1 GenomeScope genome size estimates for *C. lactis*.

chromosome-level reference genome represents a valuable resource for advancing studies on the evolutionary biology of astigmatid mites and for facilitating the development of targeted strategies for managing this economically significant pest.

Methods

Sample preparation. The *C. lactis* population used in this study was originally collected from dried fruits maintained in the insect rearing facility of the Institute of Entomology, Guizhou University (Guiyang, Guizhou, China). In the laboratory, mites were fed with yeast and reared under controlled conditions: a constant temperature of 25 ± 2°C, relative humidity of 75 ± 5%, and kept in complete darkness. To minimize heterozygosity and establish a genetically stable strain, one female and one male were isolated from the original population to initiate a new colony. A five-generation inbreeding program was carried out by pairing one male and one female in each generation to propagate the next generation. From the resulting fifth-generation inbred line, adult females were selected as the source material for whole-genome sequencing analysis.

DNA sequencing. For PacBio long-read sequencing, genomic DNA was extracted from 200 adult females. A 20-kb SMRTbell library was constructed using the SMRTbell® Express Template Prep Kit 2.0 (Pacific Biosciences, Cat. #PN 101-853-100, Menlo Park, CA, USA) and sequenced on the PacBio Sequel IIe platform. After quality filtering, approximately 5.5 Gb of high-quality SMRT sequences were retained, providing ~98 × genome coverage. For Illumina short-read sequencing, genomic DNA was also extracted from 200 adult females. A standard whole-genome sequencing library with an insert size of ~350 bp (paired-end 150 bp) was prepared using the TruSeq DNA PCR-Free Library Preparation Kit and sequenced on the Illumina NovaSeq X Plus platform, generating 11.4 Gb of paired-end reads. To anchor contigs into chromosome-scale scaffolds, we employed high-throughput chromosome conformation capture (Hi-C) sequencing technology. Fresh tissue samples from 1,000 adult females were crosslinked with formaldehyde, followed by digestion with the restriction enzyme MboI and proximity ligation to capture three-dimensional chromatin interactions. The resulting libraries were sequenced on the Illumina NovaSeq X Plus platform, producing 8.3 Gb of clean paired-end reads (150 bp in

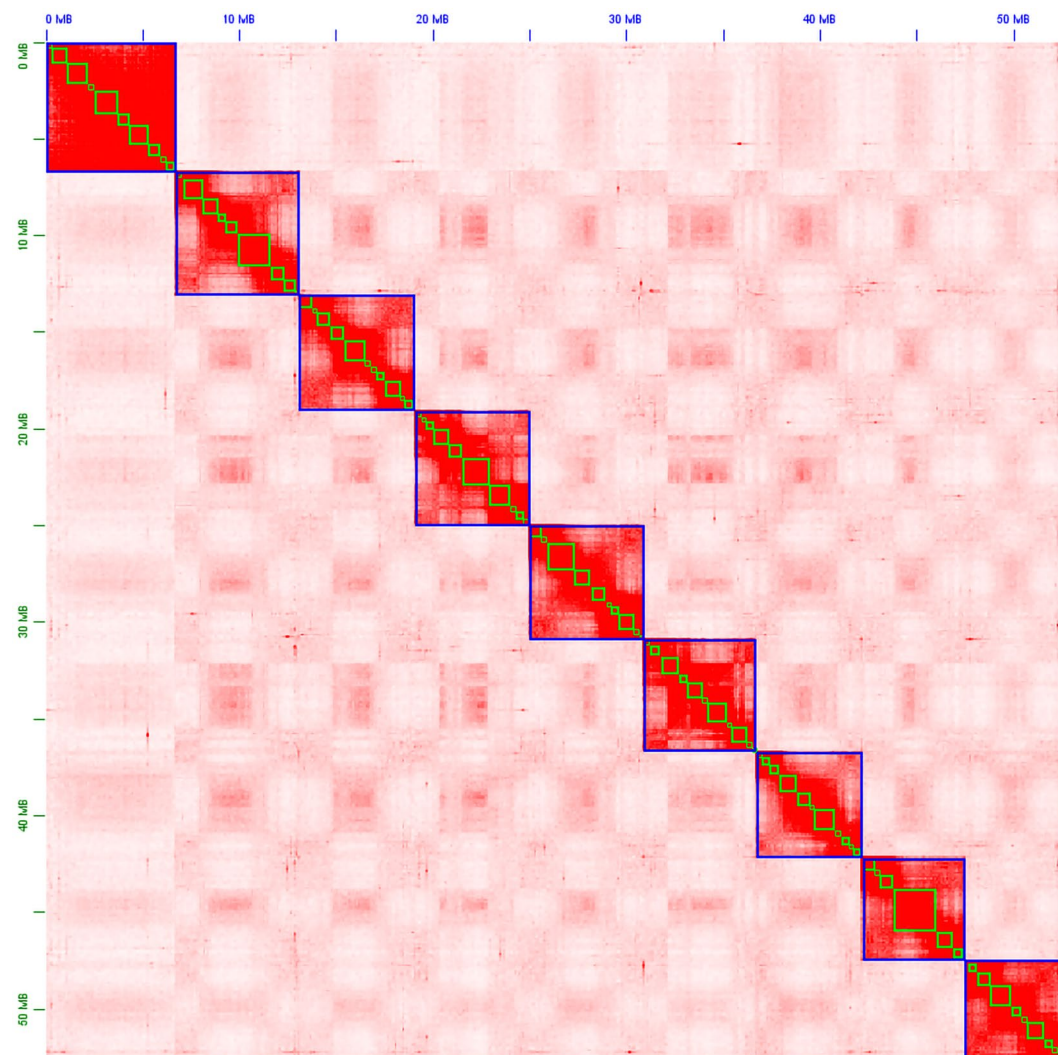


Fig. 2 Genome-wide chromosomal heatmap of *C. lactis*, the blue boxes.

length). All library preparation and sequencing operations were performed by Berry Genomics (Beijing, China). The total amount of raw sequencing data and the corresponding sequencing depth are summarized in Table 1.

RNA sequencing. For RNA short-read sequencing, total RNA was extracted from 1,000 adult females using TRIzol™ Reagent (Thermo Fisher Scientific). Poly(A)-enriched libraries were constructed according to the manufacturer's instructions using the VAHTS mRNA-seq v2 Library Prep Kit (Vazyme, NR603). Sequencing was performed on the Illumina NovaSeq X Plus platform with a paired-end 150 bp strategy, yielding a total of 7.7 Gb of high-quality clean reads. For long-read RNA sequencing, poly(A)-enriched RNA was isolated using the NEBNext Poly(A) mRNA Magnetic Isolation Module. A Nanopore sequencing library was subsequently prepared using the SQK-PCS109 + SQKPBK004 kit. Library construction was carried out by BenaGen (Wuhan, China). The purified library was sequenced on an Oxford Nanopore PromethION platform, generating 9.7 Gb of long-read sequences.

Genome survey. First, we performed quality control on the raw Illumina data using BBDuk v39.25¹¹. Duplicate reads were removed using the “clumpify.sh” script with default parameters. Subsequently, low-quality reads were filtered out using the “bbduk.sh” script, applying criteria such as base quality (>Q20), minimum read length (>15 bp), poly(A/G/C) trimming (>10 bp), and correction of overlapping paired-end reads.

The genome survey was conducted based on k-mer frequency distribution, with 21-mer frequencies assessed using Jellyfish global 1¹². Genome feature analysis was performed with GenomeScope v2.0¹³, setting the maximum k-mer depth threshold to 10,000 and parameters specified as ‘-k 21 -p 2 -m 10,000’. The results indicated a predicted genome size of 48.52 Mb and a heterozygosity level of 1.25%. Detailed analysis results are presented in Fig. 1.

Genome assembly. High-quality HiFi reads were initially assembled using Hifiasm v0.25.0-r726¹⁴ with the parameter “-l 3”. To exclude low-depth sequences potentially originating from contamination or sequencing errors, only contigs with a sequencing depth exceeding 9× (approximately 1/10 of the estimated genome

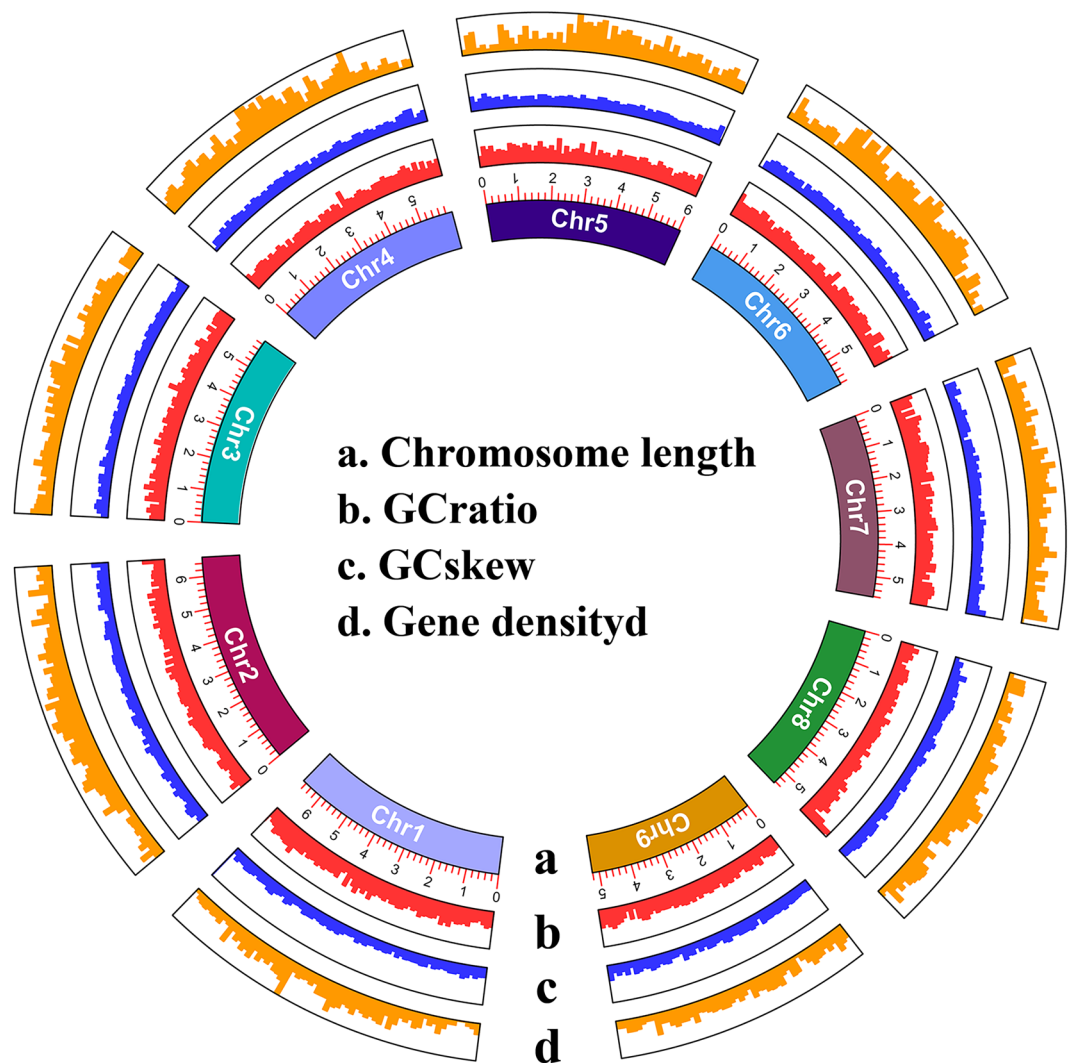


Fig. 3 Gene features in the *C. lactis* genome. The tracks, from innermost to outermost, represent the nine chromosomes (Chr1–Chr9), GC content, GC skew, and gene density.

coverage) were retained. Subsequently, the assembly was further polished using nextPolish v1.4.1¹⁵, incorporating both second and third-generation genomic sequencing data for error correction. To remove redundant sequences from the polished assembly, Purge_dups v1.2.5¹⁶ was applied. Minimap2 v2.24¹⁷ was used as the alignment tool to map HiFi reads to the genome (“-cx map-hifi”) and for self-alignment of the genome (“-x asm5 -DP”). Due to the high heterozygosity of the genome, Purge_dups was run with parameters “-2 -a 60”.

Chromosome-level scaffolding and contig assembly were performed based on Hi-C data using the Chromap and YAHS pipelines. First, Hi-C data were subjected to quality control using Chromap v0.3.0-r509¹⁸, including read alignment, removal of PCR duplicates, and extraction of Hi-C interaction signals. Subsequently, two rounds of scaffolding were carried out using YAHS v1.2¹⁹ with default parameters. Following the initial round of scaffolding, the assembly was manually curated using Juicebox v1.11.08²⁰, followed by a final round of scaffolding. The sequencing depth of the final genome assembly was assessed using SAMtools v1.16.1²¹, with aligned BAM files generated by minimap2 based on HiFi reads (“-ax map-hifi”) or second-generation whole-genome sequencing data (“-ax sr”). As shown in the Hi-C scaffolding heatmap (Fig. 2), the quality of the chromosome-level scaffolding was excellent, resulting in a total of nine chromosome-level assemblies.

Genome completeness was assessed using BUSCO v5.8.3²² with the arachnida_odb12 reference database, which contains 1,123 single-copy orthologous genes. To evaluate both the utility of the raw sequencing data and the integrity of the genome assembly, short-read and long-read sequencing data were aligned to the assembled genome using Minimap2, and alignment rates were calculated with SAMtools. Potential contamination in the assembly was screened using MMseq. 2 v137²³ with a sequence identity threshold of 0.8 (--min-seq-id 0.8) against the NCBI nt_pork (bacteria and archaea) database. Additionally, a BLASTN search (BLAST + v2.12.0²⁴) was performed against the UniVec database and the reference genome of *Saccharomyces cerevisiae* to specifically detect vector and feed-related contaminants, respectively. Sequences showing over 90% similarity to entries in these databases were considered contaminants and removed from the final assembly. Single-base quality

Features	Statistics
Length of scaffold/contig (Mb)	53.47/53.46
Scaffold/Contig N50 (Kb)	713/5,929
Contigs number	121
Scaffolds number	12
chromosomes number	9
Anchored to chromosome (%)	99.86
GC %	40.90
QV	49.40

Table 2. Statistics for the assembly of *C. lactis* genome.

Class	Length (bp)	Percent (%)
Repetitive elements Size	5,167,192	9.66
Unclassified	2,387,275	4.46
LTRs	971,581	1.82
DNA transposons	688,709	1.35
LINEs	259,790	0.49
SINEs	66,483	0.12

Table 3. Statistics of repetitive sequence annotation for *C. lactis*.

Functional Annotation	Number	Percent (%)
Uniprot (sprot)	6,989	67.66
Uniprot (trembl)	9,022	87.34
Nr	9,098	88.07
InterPro	9,621	93.14
Emapper (eggNOG)	8,116	78.57
Total	9,890	95.74

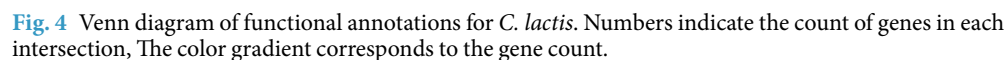
Table 4. Summary of protein-coding genes of the *C. lactis*. Note: Uniprot (sprot): UniProtKB/Swiss-Prot database. Uniprot (trembl): UniProtKB/TrEMBL database. Nr: the NCBI non-redundant protein sequence database. InterPro: a database of protein families, domains, and functional sites. Emapper (eggNOG): a functional annotation tool based on the eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) database.

scores (QV values) and k-mer spectra were evaluated using Merqury v1.4²⁵. The final chromosome-level genome assembly of *C. lactis* is 53.47 Mb, consisting of 12 scaffolds and 121 contigs, with scaffold N50 and contig N50 values of 5.92 Mb and 713.31 kb, respectively. Of these, 118 contigs (accounting for 99.86% of the total length) were anchored onto 9 chromosomes (Fig. 3). The remaining three scaffolds are all short, unplaced nuclear sequences that could not be anchored to a chromosome. Chromosome lengths ranged from 5.21 to 6.71 Mb, with an average GC content of 40.9% and a QV value of 49.4 (Table 2).

Genome annotation. The RepeatModeler v2.0.4²⁶ software was used with the additional LTR search option enabled (-LTRStruct) to construct a species-specific repeat database based on the structural characteristics of repetitive sequences and de novo prediction principles. This database was subsequently combined with the Dfam 3.7²⁷ and RepBase-20181026²⁸ databases to generate a comprehensive reference database for repeat sequence identification and alignment. Repetitive sequence prediction was carried out using RepeatMasker v4.1.5²⁹ with the final constructed repeat database. The results showed that repetitive sequences totaled 5,167,192 bp, accounting for 9.66% of the genome. In the *C. lactis* genome, the top five categories of repetitive elements were: Unknown (4.46%), LTRs (1.82%), DNA transposons (1.35%), LINEs (0.49%), and SINEs (0.12%). The detailed statistical results are presented in Table 3.

Two complementary strategies were employed for the annotation of non-coding RNAs (ncRNAs): (1) Sequences were aligned to the Rfam database, a collection of known non-coding RNA families, using Infernal v1.1.5³⁰ for the identification and annotation of rRNA, snRNA, and miRNA sequences. (2) tRNA genes were predicted using tRNAscan-SE v2.0.1³¹, which is specifically designed for accurate detection of tRNA sequences in genomic data. The annotation results revealed a total of 214 non-coding RNA genes, including 49 rRNAs, 9 miRNAs, 12 snRNAs, 123 tRNAs, and 3 ribozymes.

Protein-coding gene structure annotation was performed using the MAKER v3.01.04³² pipeline, integrating three types of evidence to improve prediction accuracy. The specific workflow included the following steps: (1)



Gene function annotation was performed by aligning gene sequences to established databases. We used Diamond v2.1.8³⁷ to search the UniProtKB database (including SwissProt and TrEMBL) and the NCBI nr database (February 7, 2024) for functional gene information. Subsequently, InterPro 5.74-105.06³⁸ and eggNOG-mapper v2.1.12³⁹ were employed to further enrich functional annotations and identify protein domains. The results showed that the five databases annotated 6,989, 9,022, 9,098, 9,621, and 8,116 functional genes, respectively. After removing duplicates, a total of 9,890 genes were functionally annotated, with 6,446 genes being commonly identified across all four databases. Detailed results are presented in Table 4 and visualized in Fig. 4.

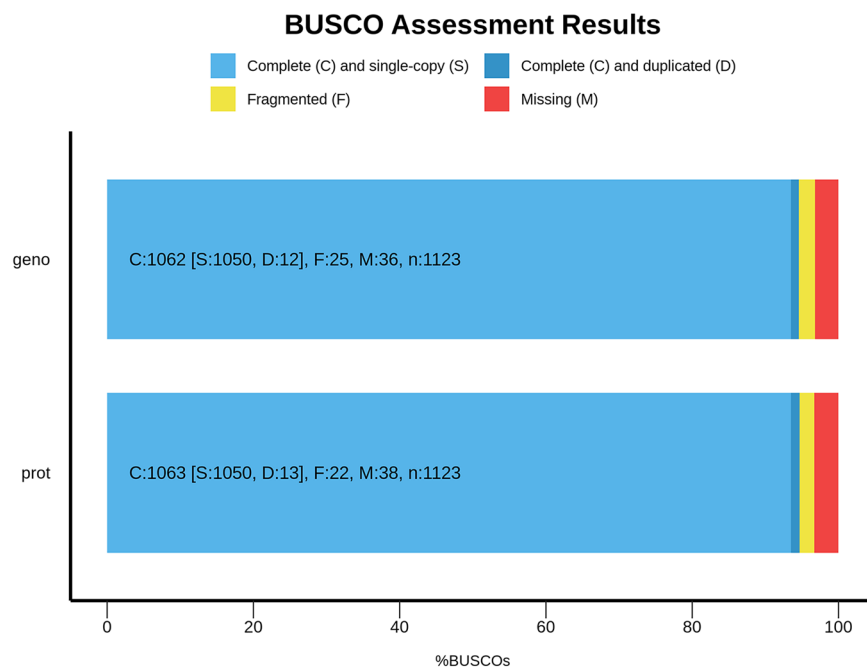


Fig. 5 BUSCO assessment results for the *C. lactis* genome assembly and annotated protein set.

Data Records

The raw sequencing reads and genome assembly have been deposited in the NCBI database under BioProject accession number PRJNA1260190⁴⁰. The WGS, Hi-C, HiFi, RNA-seq, and ONT data are publicly available under accession number SRP588867⁴¹. The complete genome assembly is accessible through NCBI with the accession number GCA_051106175.1⁴². Additionally, the genome assembly and annotation files have been made publicly available via Figshare⁴³.

Technical Validation

The quality of the *C. lactis* genome assembly was assessed using two complementary approaches. First, completeness was evaluated using BUSCO v5.8.3 with the arachnida_odb12 dataset (n = 1,123) (Fig. 5), revealing a total completeness of 94.6% (93.6% single-copy and 1.0% duplicated). Second, base-level accuracy was estimated using Merquy v1.4, yielding a high-quality value (QV) of 49.9, indicating exceptional sequence fidelity. These results collectively demonstrate that the genome assembly is highly contiguous and complete. Additionally, the gene prediction achieved a BUSCO completeness of 94.7% (93.5% single-copy and 1.2% duplicated), further confirming the overall high quality of both the assembly and annotation.

Data availability

The data presented in this manuscript have not been previously published in any form. The raw sequencing reads and the assembled genome described in this study have been deposited in the National Center for Biotechnology Information (NCBI) databases under BioProject accession number PRJNA1260190. The corresponding genome annotation files are available in the figshare repository (<https://doi.org/10.6084/m9.figshare.29815598>).

Code availability

No custom scripts were used in this study. All data processing was carried out using standardized pipelines based on the bioinformatics tools described in the Methods section.

Received: 5 August 2025; Accepted: 24 September 2025;

Published online: 06 November 2025

References

- Schatz, H., Behan-Palletier, V. M., O'Connor, B. M., & Norton, R. A. Suborder Oribatida van der Hammen, 1968. In: Zhang, Z.-Q. (Ed.) Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness. *Zootaxa*. 3148 (1), 141–148 (2011).
- Klimov, P. B. *et al.* A transitional fossil mite (Astigmata: Levantoglyphidae fam. n.) from the early Cretaceous suggests gradual evolution of phoresy-related metamorphosis. *Sci Rep-UK*. **11**(1), 15113 (2021).
- O'Connor, B. M. Evolutionary ecology of astigmatid mites. *Annu Rev Entomol*. **27**(1), 385–409 (1982).
- Xiong, Q. *et al.* Comparative genomics reveals insights into the divergent evolution of astigmatic mites and household pest adaptations. *Mol Biol Evol*. **39**(5), msac097 (2022).
- Houck, M. A. & O'Connor, B. M. Ecological and evolutionary significance of phoresy in the Astigmata. *Annu Rev Entomol*. **36**(1), 611–636 (1991).
- Seeman, O. D. & Walter, D. E. Phoresy and mites: More than just a free ride. *Annu Rev Entomol*. **68**(1), 69–88 (2023).
- Aliakbarpour, H. & Fan, Q. H. The genus *Carpoglyphus* (Acariformes: Carpoglyphidae). *Zoosymposia*. **22**, 231–231 (2022).

8. Hubert, J., Nesvorna, M., Green, S. J. & Klimov, P. B. Microbial communities of stored product mites: variation by species and population. *Microb Ecol.* **81**(2), 506–522 (2021).
9. Hubert, J., Nesvorna, M., Kopecký, J., Ságová-Marečková, M. & Poltronieri, P. *Carpoglyphus lactis* (Acari: Astigmata) from various dried fruits differed in associated micro-organisms. *J. Appl. Microbiol.* **118**(2), 470–484 (2015).
10. Li, C. P., Cui, Y. B., Wang, J., Yang, Q. G. & Tian, Y. Acaroid mite, intestinal and urinary acariasis. *World J Gastroentero.* **9**(4), 874 (2003).
11. Bushnell, B. BBtools. Available online: <https://sourceforge.net/projects/bbmap/> (accessed on 5 May 2025) (2014).
12. Lee, S. H. *et al.* The global spread of jellyfish hazards mirrors the pace of human imprint in the marine environment. *Environ Int.* **171**, 107699 (2023).
13. 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).
14. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods.* **18**, 170–175 (2021).
15. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics.* **36**(7), 2253–2255 (2020).
16. Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics.* **36**, 2896–2898 (2020).
17. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics.* **34**, 3094–3100 (2018).
18. Zhang, H. *et al.* Fast alignment and preprocessing of chromatin profiles with Chromap. *Nat Commun.* **12**(1), 6566 (2021).
19. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: Yet another Hi-C scaffolding tool. *Bioinformatics.* **39**, btac808 (2023).
20. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3**, 95–98 (2016).
21. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience.* **10**, giab008 (2021).
22. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol.* **38**, 4647–4654 (2021).
23. Steinegger, M. & Söding, J. MMseqs. 2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol.* **35**, 1026–1028 (2017).
24. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**(3), 403–410 (1990).
25. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
26. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci.* **117**, 9451–9457 (2020).
27. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA.* **12**, 2 (2021).
28. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA.* **6**, 11 (2015).
29. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. <http://www.repeatmasker.org> (2013/2015).
30. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics.* **29**, 2933–2935 (2013).
31. Chan, P. P. & Lowe, T. M. TRNAscan-SE: Searching for tRNA genes in genomic sequences. *Methods Mol Biol.* **1962**, 1–14 (2019).
32. Holt, C. & Yandell, M. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics.* **12**, 491 (2011).
33. Gabriel, L. *et al.* BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Research* **34**(5), 769–777 (2024).
34. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol.* **37**, 907–915 (2019).
35. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
36. Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S. O. & Grau, J. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinformatics.* **19** (2018).
37. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods.* **18**, 366–368 (2021).
38. Finn, R. D. *et al.* InterPro in 2017—Beyond protein family and domain annotations. *Nucleic Acids Res.* **45**, D190–D199 (2017).
39. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol.* **38**, 5825–5829 (2021).
40. NCBI BioProject <https://identifiers.org/ncbi/bioproject:PRJNA1260190> (2025).
41. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRP588867> (2025).
42. NCBI GenBank https://identifiers.org/ncbi/insdc.gca:GCA_051106175.1 (2025).
43. Liang, L. *et al.* Genome Assembly and Annotation of *Carpoglyphus lactis*. Figshare <https://doi.org/10.6084/m9.figshare.29815598> (2025).

Acknowledgements

This study was supported by the National Natural Science Foundation of China (Grant No.32160118 and 32470478), and the National Key Research and Development Program of China (Grant No. 2022FY202100). We gratefully acknowledge Prof. Feng Zhang from Nanjing Agricultural University for his expert technical guidance and insightful suggestions during the preparation of this manuscript.

Author contributions

Tian-Ci Yi and Dao-Chao Jin provided financial support and technical guidance for this study. Lang Liang conducted the data analysis, as well as drafted and revised the manuscript. Hu-Die He contributed to sample collection, preparation, and also participated in manuscript writing. All authors have read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.-C.J. or T.-C.Y.

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.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025