scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of the Adonis ladybird *Hippodamia variegata*

Hong-Ling Liu[1,5]✉, Yan-Ping Yong[1,5], Xing-Long Wu[1], Zhi-Teng Chen[2], Shu-Jun Wei[3], Peng Cai[4] & De-Qiang Pu[1]✉

The Adonis ladybird (*Hippodamia variegata*), an important predator in agricultural ecosystems, plays a crucial role in biological control and is a significant model for evolutionary and genomic studies within Coccinellidae. Despite its ecological importance, the lack of a reference genome for *H. variegata* has limited in-depth investigations into its biology and potential as a biocontrol agent. Here, we present a high-quality, chromosome-level genome assembly of *H. variegata*. The final assembly spans 493.01 Mb, with a scaffold N50 of 28.19 Mb and a GC content of 36.41%. Hi-C sequencing data enabled the anchoring of 343.20 Mb to 10 chromosomes. Repetitive elements accounted for 258.56 Mb (52.44%) of the genome, with long interspersed nuclear elements (LINEs) being the most prevalent. We identified 37,348 protein-coding genes, of which 78.55% were functionally annotated in public protein databases. This high-quality genome assembly will serve as a valuable resource for furthering our understanding of Adonis ladybird's evolutionary biology, enhancing its utility in pest management, and supporting future research on ladybird genomics.

Background & Summary

Ladybird beetles (Coleoptera: Coccinellidae) are a highly diverse family within the Polyphaga suborder, comprising over 6000 species across 360 genera and eight subfamilies, with a wide global distribution[1,2]. These beetles play an essential role as generalist predators, primarily feeding on various agricultural pests such as aphids, whiteflies, scale insects, and mites[3]. Many species within Coccinellidae are integral to Integrated Pest Management (IPM) programs globally due to their predatory habits, which help control pest populations while minimizing the need for insecticides[4,5]. However, the extensive use of chemical insecticides has led to resistance development in many pest species, increasing the importance of IPM strategies that incorporate natural predators such as ladybirds alongside reduced pesticide use[5,6]. Despite their essential ecological and agricultural functions, our understanding of the genetic and genomic basis of their adaptability and biological roles remains limited.

Hippodamia variegata (Goeze, 1777)[7], also known as the Adonis ladybird, the variegated ladybug, and spotted amber ladybeetle, is one of the most commercially important ladybird species. It is widely recognized for its effectiveness in controlling aphid populations in economically significant crops like cotton, peach, and tobacco[8,9]. This species originated in the Palearctic region but has now spread globally due to its high biotic potential and predatory aggressiveness[10,11]. This species is considered the most effective natural enemy of aphids in many countries, such as Bulgaria, Ukraine, Italy, India, Turkmenistan, and China, where it plays a critical role in maintaining pest populations at manageable levels in various crops like wheat, cotton, vegetables, and orchards[12–20]. Numerous studies have investigated the biological characteristics of *H. variegata*, such as its life table parameters, functional responses to prey, prey suitability, and predatory efficiency, highlighting its utility in biological control programs[20–24]. Recent studies have revealed that *H. variegata* has a diploid chromosome number ($2n = 20$), an XY sex determination system, and a genome size estimated at 284 Mb with a G + C content of

[1]Institute of Plant Protection, Sichuan Academy of Agricultural Sciences, Chengdu, 610066, China. [2]School of Grain Science and Technology, Jiangsu University of Science and Technology, Zhenjiang, 212004, China. [3]Institute of Plant and Environmental Protection, Beijing Academy of Agriculture and Forestry Sciences, Beijing, 100097, China. [4]Horticultural Institute, Sichuan Academy of Agricultural Sciences, Vegetable Germplasm Innovation and Variety Improvement Key Laboratory of Sichuan Province, Chengdu, 610066, China. [5]These authors contributed equally: Hong-Ling Liu, Yan-Ping Yong. ✉e-mail: liuhongling@ippsaas.org.cn; pdqpudeqiang@163.com

36.4%[25]. Despite these studies, genomic resources for *H. variegata* remain scarce, limiting our ability to understand its predatory and adaptive traits at the genetic level.

Although there have been significant advances in genomic research on insects, the genomic resources for Coccinellidae remain scarce, with very few chromosome-level genome assemblies available[26]. High-quality genome assemblies can provide insights into the genetic basis of key biological traits such as pest resistance, feeding behavior, and environmental adaptability[27], which are crucial for understanding and enhancing the efficacy of ladybirds in biological control programs. In this study, we present the chromosome-level genome assembly of *H. variegata*, utilizing a combination of Illumina short-read sequencing, PacBio HiFi (high-fidelity) long-read sequencing, and Hi-C (high-throughput chromosome conformation capture) data. We aim to generate a high-quality genome assembly to facilitate further research on its genetic underpinnings and functional genomics. This genome resource will serve as a foundation for comparative studies within Coccinellidae and contribute to the development of more effective biological control strategies.

Methods

Sample and DNA preparation. The specimens of *H. variegata* used in this study were maintained under controlled laboratory conditions at the Sichuan Academy of Agricultural Sciences in Chengdu, China. Genomic DNA was extracted from a healthy adult female using the TIANGEN Blood & Tissue kit (Tiangen, Beijing, China), following the manufacturer's protocol. The DNA quality was assessed with NanoDrop (NanoDrop products, Wilmington, DE, USA), Qubit 3.0 Fluorometer (Life Technologies Corporation, Eugene, OR, USA), and 1% agarose gel electrophoresis. All three sequencing libraries (Illumina, PacBio, and Hi-C) were constructed from the same genomic DNA sample. Different amounts of DNA were used for each library type to optimize the sequencing protocols.

Paired-end library preparation, sequencing and quality control. A paired-end Illumina sequencing library with an insert size of 400 bp was constructed from genomic DNA using the Illumina TruSeq DNA PCR-free prep kit (Illumina Inc., San Diego, CA, USA). The library quality was assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA), and the quantification was performed with the Promega QuantiFluor system (Thermo Fisher Scientific, Waltham, MA, USA). Sequencing was performed on an Illumina NovaSeq. 6000 platform (Illumina Inc., San Diego, CA, USA) with a 150-bp paired-end strategy. The library preparation and sequencing were performed at Shanghai Personal Biotechnology Co., Ltd. (Shanghai, China). This process generated approximately 70.8 Gb of raw sequences (Table S1). Quality control and filtering of the raw data were performed using Fastp v0.23.1[28]. The reads with adapter contamination, low-quality reads with a mean PHRED score below 20%, reads containing poly-N, and reads shorter than 150 bp, are all filtered out. A random subset of 10,000 high-quality reads was aligned against public GenBank data to check for potential external contamination. A total of approximately 68.7 Gb of high-quality Illumina sequencing data was obtained for genome survey analysis (Table S2).

Genome survey. The genome survey was conducted using k-mer analysis (k = 19) performed with Jellyfish v2.3.0[29]. The genome size, heterozygosity rate, and duplication rates were estimated using GenomeScope[30]. The estimated genome size of *H. variegata* was approximately 672.49 Mb, with a heterozygosity rate of 1.01% and a repetitive fraction of 83.61% (Fig. 1a, Table S3). The estimated genome size is larger than the final assembled genome size of 493.01 Mb. The discrepancy is likely due to the fact that the genome survey method (k-mer analysis) may overestimate the genome size, particularly due to the presence of repetitive regions that were not fully assembled in the final genome. The ploidy level was confirmed to be diploid through k-mer analysis using Smudgeplot v0.2.3[31] (Fig. 1b).

PacBio library construction and sequencing. For long-read sequencing, the PacBio single molecule real-time (SMRT) library was constructed according to the standard PacBio Template Prep Kit 1.0 protocol (Pacific Biosciences, Menlo Park, CA, USA). The PacBio library quality was assessed with an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA), and quantification was performed using the Promega QuantiFluor system (Thermo Fisher Scientific, Waltham, MA, USA). Sequencing was carried out on a PacBio Revio platform (Pacific Biosciences, Menlo Park, CA, USA) in circular consensus sequencing (CCS) mode, generating 5,139,915 HiFi reads with a read length N50 of 19,515 bp, which were used for genome assembly (Table S4).

Hi-C library preparation and sequencing. The Hi-C library was constructed following the TruSeq DNA PCR-free prep kit protocol (Illumina Inc., San Diego, CA, USA) and based on a standardized procedure as described in a previous study[26]. The library was sequenced on an Illumina NovaSeq. 6000 platform (Illumina Inc., San Diego, CA, USA), which generated approximately 68.9 Gb of raw sequencing data (Table S5). After quality control and data filtration using Fastp v0.23.1, we obtained approximately 65.3 Gb of high-quality Hi-C data (Table S6). HiC-PRO v3.1.0[32] was used to align the Hi-C data, with the restriction enzyme junction site set to 'GATCGATC'. All other parameters were remained default. The Hi-C alignment process generated approximately 71.5 million uniquely mapped paired-end reads, out of which 81.5% were valid interaction pairs useful to assist further genome assembly (Table S7).

Transcriptome library preparation and sequencing. Total RNA was extracted from the whole body of an adult female using the Trizol Reagent kit (Invitrogen, Carlsbad, CA, USA). The quality and quantity of RNA were evaluated using a NanoDrop spectrophotometer (NanoDrop products, Wilmington, DE, USA), a Qubit 3.0 Fluorometer (Life Technologies Corporation, Eugene, OR, USA), and 1% agarose gel electrophoresis. Subsequently, an RNA library was constructed using the SMRTbell™ Template Prep Kit 1.0 (Illumina Inc., San Diego, CA, USA) following the manufacturer's instructions. Sequencing was performed in CCS mode on a PacBio

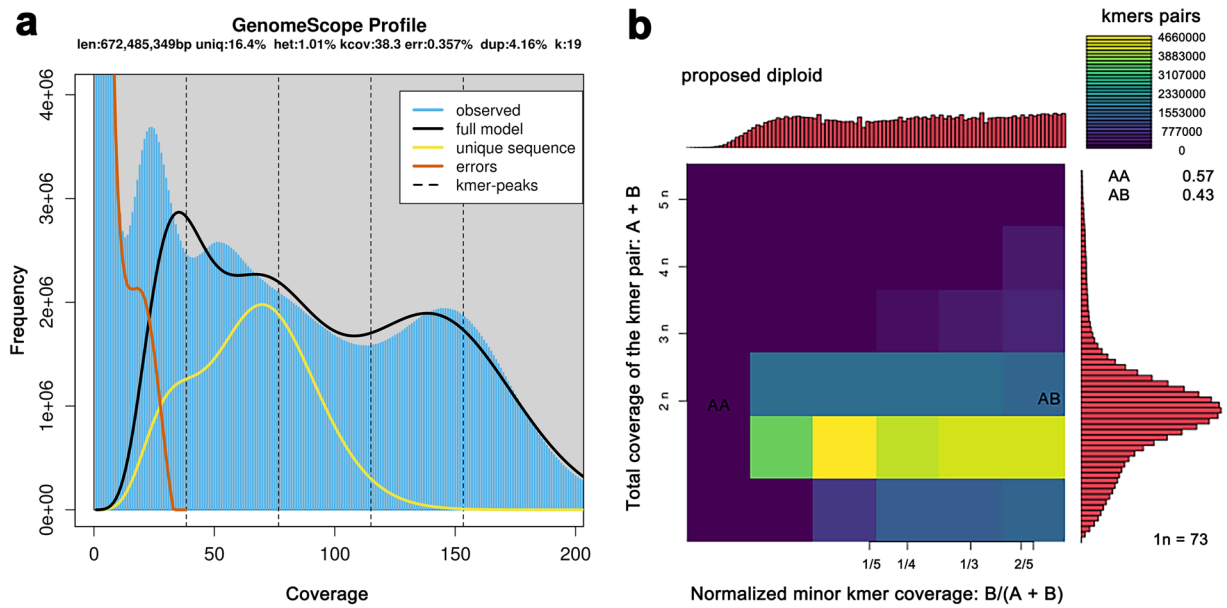

Fig. 1 Genome profiling plots of *Hippodamia variegata*. (**a**) GenomeScope plot showing K-mer distribution curve of Illumina paired-end reads based on a k value of 19. (**b**) Smudgeplot proposing *H. variegata* as diploid. Estimated ploidies are indicated on the upper left of each graph, with the likelihood of various ploidies shown on the right.

Revio platform (Pacific Biosciences, Menlo Park, CA, USA). This process yielded 4,614,863 high-quality CCS reads, amounting to approximately 6 Gb of data with a read N50 of 1,431 bp (Table S8). These sequences were used to assist genome annotation.

Genome assembly. The PacBio HiFi reads were assembled using the Improved Phased Assembly (IPA) method (Pacific Biosciences, Menlo Park, CA, USA), which generated phased assemblies including accurate haplotigs and overlaps. To identify and eliminate haplotypic duplications, Purge_Dups v1.2.5 (https://github.com/dfguan/purge_dups) was employed[33]. High-quality Hi-C reads were utilized to scaffold the haploid contigs into chromosomes using the 3D de novo assembly software, 3D-DNA v201008[34]. In total, 343,197,613 bp of sequence data were anchored to 10 chromosomes (Fig. 2a, Table S9), representing 9 autosomes and the X sex chromosome. The assembly was polished in three rounds of correction using Racon v1.4.20[35]. The scaffolded chromosomes were subsequently reordered in accordance with the reference genome (GenBank No. GCF_002263795.3). The final genome size of *H. variegata* was determined to be 493,014,220 bp, with a scaffold N50 length of approximately 28.19 Mb and a GC content of 36.41% (Fig. 2b, Table 1).

Repeat annotation. Repetitive elements (REs) in the genome were identified through both de novo annotation and homology-based search. The de novo repeat library was constructed using RepeatModeler v2.0.4[36], while protein-coding sequences with similarity to those in the Swiss-Prot database were removed from the library. Homology-based searches were conducted using RepeatMasker v4.1.4[36] against the RepBase-20150807 database. A total of 258.56 Mb of repetitive sequences were identified, comprising 52.44% of the genome (Table 2, Table S10). Long interspersed nuclear elements (LINEs) represented the largest proportion, comprising 99.25 Mb (20.13%). Additionally, 100.71 Mb (20.43%) of unclassified repetitive elements were detected, which could not be assigned to any known repeat families.

Non-coding RNA annotation. Transfer RNA (tRNA) and ribosomal RNA (rRNA) genes were identified using tRNAscan-SE v1.3.1[37] and RNAmmer v1.2[38], respectively. Other non-coding RNA (ncRNA) genes were identified using Infernal v1.1.3[39] against the Rfam v1.0 database[40]. As a result, 3,778 ncRNAs were identified in the *H. variegata* genome, including 510 18S rRNAs, 461 28S rRNAs, 2,039 tRNAs, and 768 other ncRNAs which further contain 192 miRNAs, 66 snRNAs, 71 snoRNAs, and 95 sRNAs functional in gene regulation, pre-mRNA splicing, rRNA processing, and the regulation of chromatin structure and gene expression (Table 2, Table S11).

Protein-coding genes prediction. The prediction of protein-coding genes (PCGs) was performed using a combination of *ab initio* prediction, homology-based prediction, and transcriptome-based approaches. *Ab initio* gene prediction was conducted using Augustus v3.3.2[41], GeneID v1.4[42], and GeneMark v4.71[43]. For homology-based prediction, GeMoMa v1.9[44] and Exonerate v2.2.0[45] were employed, along with amino acid sequences from related species. Transcriptome-based predictions were conducted using PASA v2.5.2[46]. The integration of these gene predictions into a final consensus set was performed using EVidenceModeler (EVM) vr2012-06-25[47]. A total of 37,348 genes were predicted (Table 2), with an average gene length of 5,253.3 bp (Table S12). Exons and introns had average lengths of ca. 441 bp and ca. 1433 bp, respectively, with each gene

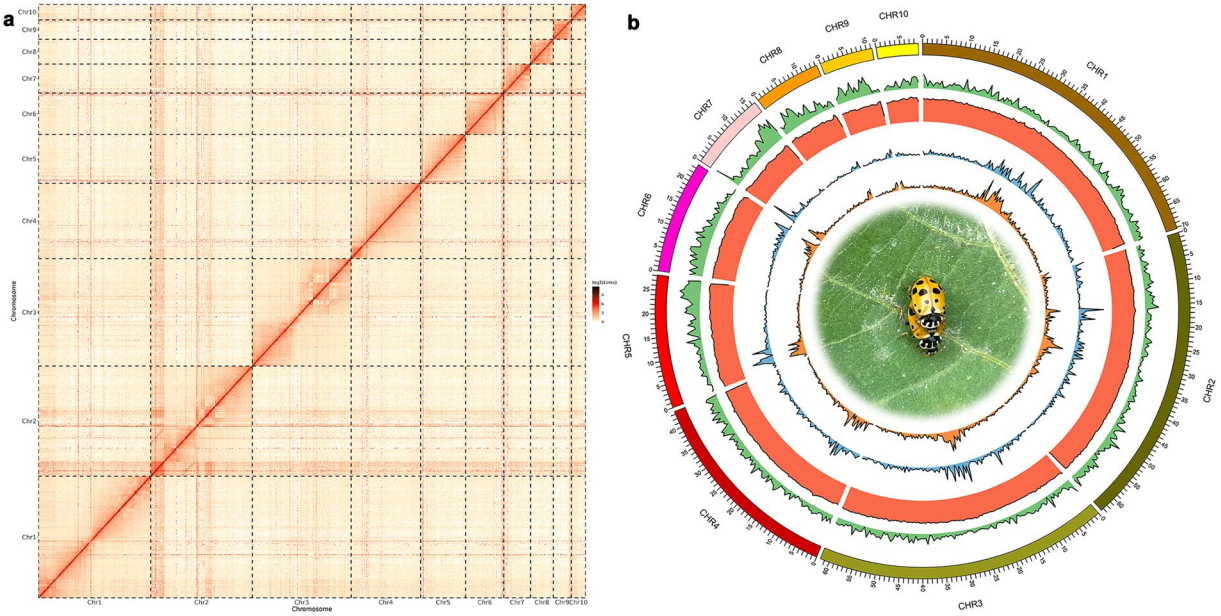

Fig. 2 Summary of the final genome assembly results. (**a**) Hi-C interaction heatmap of the *H. variegata* genome. The colour block indicates the intensity of interaction from yellow (low) to red (high). Chr1 to Chr9 represent the 9 autosomes and Chr10 is the X sex chromosome. (**b**) Circular diagram of genomic features of *H. variegata*. Circos plots from outer to inner ring represent chromosome karyotype analysis results, PCG density, GC content, LTR/Copia transposable elements density, and LTR/Gypsy transposable elements density, respectively.

Parameter	Number
Total sequence length (bp)	493,014,220
Total sequence number	1,953
Min sequence length (bp)	8,303
Max sequence length (bp)	70,261,243
N20 (bp)	63,576,082
N20 number	2
N50 (bp)	28,192,547
N50 number	5
N90 (bp)	72,960
N90 number	423
N number	3,066
N rate %	6.2188875606874e-06
GC content %	36.41
Sequences greater than 1 kb	1,953
Complete BUSCOs (%)	98.1%
Complete single-copy BUSCOs (%)	96.4%
Complete duplicated BUSCOs (%)	1.7%

Table 1. Genome assembly and assessment of *Hippodamia variegata*.

containing an average of 3.5 exons. The predicted genes contained 133,237 coding sequences (CDSs), which had an average length of ca. 441 bp and contributed to a total CDS length of 58,781,628 bp, accounting for 11.92% of the genome.

Gene function annotation. Putative gene functions were annotated by aligning gene sequences to the NCBI non-redundant protein sequence (NR) database and the Swiss-Prot database using Diamond v2.0.14.152[48]. The Pfam database was used to identify protein domains through InterProScan v5.61–93.0[49], which also provided Gene Ontology (GO) term assignments. KEGG pathway analysis was conducted using the KEGG Automatic Annotation Server (KAAS; https://www.genome.jp/kegg/kaas/)[50]. Carbohydrate-active enzymes (CAZy)[51] were predicted using HMMER v3.3.2[52].

Annotation	Number
Repetitive sequences (%)	52.44%
Retroelements (%)	22.56%
SINEs (%)	0.01%
LINEs (%)	20.13%
Ty1/Copia (%)	0.32%
Gypsy/DIRS1 (%)	1.33%
DNA transposons (%)	9.46%
18 s rRNA	510
28 s rRNA	461
tRNA	2039
Other ncRNA	768
Total genes length (bp)	196,200,843
Number of protein-coding genes	37,348
Average gene length (bp)	5,253.3
Average CDS length (bp)	441.1
Average exon length (bp)	441.1
Average intron length (bp)	1,433.1
Average exon number per gene	3.5
Complete BUSCOs (%)	97.9%
GO	23446
KEGG	6385
NR description	21992
SwissProt description	24599
Pfam description	29337
Functional annotation genes (%)	88.32

Table 2. Genome annotation of *Hippodamia variegata*.

Among the 37,348 genes predicted in the *H. variegata* genome, 21,992 (58.88%) were annotated in the NR database, 24,599 (65.86%) in the Swiss-Prot database, and 29,337 (78.55%) in the Pfam database (Table 2, Table S13). GO terms were assigned to 23,446 (62.78%) genes (Fig. S1), while 6,385 (17.1%) genes were assigned KEGG Orthology (KO) terms (Fig. S2). The CAZy annotation predicted 344 glycosyl transferase (GT) genes, 21 polysaccharide lyase (PL) genes, 281 carbohydrate esterase (CE) genes, 115 auxiliary activity (AA) genes, 124 carbohydrate-binding module (CBM) genes, and 363 glycoside hydrolase (GH) genes (Fig. S3, Table S14).

Data Records

The genomic (Illumina, PacBio, Hi-C) and transcriptomic sequencing data was deposited at the NCBI Sequence Read Archive (SRA) database under BioProject ID PRJNA1172969[53]. The accession numbers of the Illumina sequencing data, HiFi sequencing data, Hi-C sequencing data, and transcriptomic data are SRR31065962[54], SRR31065963[55], SRR31065964[56], and SRR31065965[57], respectively. The accession number of the genome assembly is JBJYYB000000000[58].

Technical Validation

The completeness and continuity of the *H. variegata* genome assembly were evaluated using Benchmarking Universal Single-Copy Orthologues (BUSCO) v5.4.5[59] against the Insecta_odb10 database, which contains 1,367 conserved gene sets. The BUSCO analysis revealed that 1,341 (98.1%) of the genes were complete (Table S15), thus confirming the high quality of the assembled genome. For the predicted PCGs, BUSCO analysis identified 1,338 (97.9%) complete genes, of which 92.3% were single-copy and 5.6% were duplicated (Table S16). Collectively, these results indicate that the *H. variegata* genome assembly achieved in this study is of high quality, with substantial contiguity and completeness.

Code availability

The bioinformatic analyses were performed using the manuals and protocols by the software developers. If manually adjusted parameters were used, the software version and method used are described in the Methods.

Received: 25 December 2024; Accepted: 21 March 2025;
Published online: 01 April 2025

References

1. Marin, J., Crouau-Roy, B., Hemptinne, J. L., Lecompte, E. & Magro, A. *Coccinella septempunctata* (Coleoptera: Coccinellidae): a species complex? *Zool. Scr.* **39**, 591–602 (2010).
2. Nedvěd, O. & Kovář, I. in *Ecology and Behaviour of the Ladybird Beetles (Coccinellidae)* (eds. Hodek, I., van Emden, H. F. & Honěk, A.) (Blackwell Publishing Ltd., 2012)

3. Hodek, I. & Honěk, A. Scale insects, mealybugs, whiteflies and psyllids (Hemiptera, Sternorrhyncha) as prey of ladybirds. *Biol. Control* **51**, 232–243 (2009).
4. Obrycki, J. J. & Kring, T. J. Predaceous Coccinellidae in biological control. *Annu. Rev. Entomol.* **43**, 295–321 (1998).
5. Foster, S. P., Devine, G. & Devonshire, A. In *Aphids as Crop Pests* (eds. van Emden H. F. & Harrington R.) (CAB International, 2007).
6. Jalali, M. A., Leeuwen, T. V., Tirry, L. & De Clercq, P. Toxicity of selected insecticides to the two-spot ladybird *Adalia bipunctata*. *Phytoparasitica* **37**, 323–326 (2009).
7. Goeze, J. A. E. *Entomologische Beyträge zu des Ritter Linné zwölften Ausgabe des Natursystems. Vol. 1*. (Bey Weidmanns Erben und Reich, 1777).
8. Rondoni, G., Ielo, F., Ricci, C. & Conti, E. Intraguild predation responses in two aphidophagous Coccinellids identify differences among juvenile stages and aphid densities. *Insects* **5**, 974–983 (2014).
9. Skouras, P. J. & Stathas, G. J. Development, growth and body weight of *Hippodamia variegata* fed *Aphis fabae* in the laboratory. *Bull. Insect.* **68**, 193–198 (2015).
10. Gordon, R. D. The first North American records of *Hippodamia variegata* (Goeze) (Coleoptera: Coccinellidae). *J. N. Y. Entomol. Soc.* **95**, 307–309 (1987).
11. Franzmann, B. A. *Hippodamia variegata* (Goeze) (Coleoptera: Coccinellidae), a predacious ladybird new in Australia. *Aust. J. Entomol.* **41**, 375–377 (2002).
12. Hammed, S. F., Sud, V. K. & Kashyap, N. P. *Adonia variegata* (Goeze) Coleoptera: Coccinellidae), an important predator of the Indian grain aphid, *Macrosiphum (Sitobion) miscanthi* Tak. In Kulu Valley (Himachal Pradesh). *Indian J. Entomol.* **37**, 209–210 (1975).
13. Wang, Y. H., Liu, B. S., Fu, Z. H. & Gu, L. N. Studies on the life history and occurrence of *Adonia variegata* (Goeze). *Entomol. Knowl.* **21**, 19–22 (1984).
14. Belikova, E. V. & Kosaev, E. M. The biology of the most important species of Coccinellidae and their role in controlling aphids in a cotton-lucerne rotation. *Biologicheskikh Nauk.* **5**, 61–63 (1985).
15. Gumovskaya, G. N. The coccinellid fauna. *Zashch. Rast.* **11**, 43 (1985).
16. Pang, B. P. A structure of insect community in wheat fields and its diversity. *Entomol. Knowl.* **30**, 263–266 (1993).
17. Nicoli, G., Limonta, L., Gavazzuti, C. & Pozzati, M. The role of hedges in the agroecosystem. Initial studies on the coccinellid predators of aphids. *Informatore Fitopathologico* **45**, 7–8 (1995).
18. Yang, C. J. *et al*. Spatial distribution patterns and sampling techniques of *Hippodarnia variegata* (Goeze) on the tobacco fields in northern Shaanxi. *Entomol. Knowl.* **34**, 283–288 (1997).
19. Natskova, V. The effect of aphid predators on the abundance of aphids on peppers. *Rastitelna Zashchita* **21**, 20–22 (1973).
20. Kontodimas, D. C. & Stathas, G. J. Phenology, fecundity and life table parameters of the predator *Hippodamia variegata* reared on *Dysaphis crataegi*. *Biocontrol* **50**, 223–233 (2005).
21. Fan, G. H., Liu, B. X., Song, Q. B. & Ma, G. R. Studies on biology of *Adonia variegata* Goeze. *Entomol. J. East China* **4**, 70–74 (1995).
22. Khan, A. A. & Mir, R. A. Functional response of four predaceous coccinellids, *Adalia tetraspilota* (Hope), *Coccinella septempunctata* L., *Calvia punctata* (Mulsant) and *Hippodamia variegata* (Goeze) feeding on green apple aphid, *Aphis pomi* De Geer (Homoptera: Aphididae). *J. Biol. Control* **22**, 291–298 (2008).
23. Farhadi, R., Allahyari, H. & Juliano, S. A. Functional response of larval and adult stages of *Hippodamia variegata* (Coleoptera: Coccinellidae) to different densities of *Aphis fabae* (Hemiptera: Aphididae). *Environ. Entomol.* **39**, 1586–1592 (2010).
24. Wu, X. H., Zhou, X. R. & Pang, B. P. Influence of five host plants of *Aphis gossypii* Glover on some population parameters of *Hippodamia variegata* (Goeze). *J. Pest Sci.* **83**, 77–83 (2010).
25. Mora, P. *et al*. Satellitome analysis in the ladybird beetle *Hippodamia variegata* (Coleoptera, Coccinellidae). *Genes* **11**, 783 (2020).
26. Pu, D. Q. *et al*. Chromosome-level genome assembly of the giant ladybug *Megalocaria dilatata*. *Sci. Data* **11**, 117 (2024).
27. Li, F. *et al*. Insect genomes: progress and challenges. *Insect Mol. Biol.* **28**, 739–758 (2019).
28. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
29. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
30. Vurture, G. W. *et al*. GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
31. 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).
32. Servant, N. *et al*. HIC-PRO: An optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
33. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
34. Dudchenko, O. *et al*. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
35. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Racon – Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
36. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **25**, 4–10 (2009).
37. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
38. Lagesen, K. *et al*. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
39. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
40. Griffiths-Jones, S. *et al*. Rfam: Annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **3**, 121–124 (2005).
41. 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).
42. Blanco, E., Parra, G. & Guigó, R. Using geneid to identify genes. *Curr. Protoc. Bioinformatics* **18**, 4–30 (2007).
43. Borodovsky, M. & Lomsadze, A. Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. *Curr. Protoc. Bioinformatics* **35**, 4.6.1–4.6.10 (2011).
44. Keilwagen, J. *et al*. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods Mol. Biol.* **1962**, 161–177 (2019).
45. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
46. Haas, B. J. *et al*. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
47. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
48. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
49. Jones, P. *et al*. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
50. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, 182–185 (2007).
51. Lombard, V. *et al*. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, 490–495 (2014).

52. Meng, X. & Ji, Y. Modern computational techniques for the HMMER sequence analysis. *ISRN Bioinform.* **2013**, 252183 (2013).
53. *NCBI BioProject* https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1172969 (2024).
54. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31065962 (2024).
55. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31065963 (2024).
56. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31065964 (2024).
57. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31065965 (2024).
58. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBJYYB000000000 (2024).
59. 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 (2015).

Acknowledgements

This work was supported by the National Key R & D Program of China (2023YFD1400600), Natural Science Foundation Project of Sichuan Province (24NSFSC0390), and the Tea Innovation Team of National Modern Agricultural Industry Technology System (sccxtd-2024-10).

Author contributions

D.Q.P. and H.L.L. conceived and designed the study. X.L.W., Z.T.C. and S.J.W. performed the analyses. H.L.L. and Y.P.Y. drafted the manuscript. P.C. and H.L.L. revised the manuscript. All authors have 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/s41597-025-04882-4.

Correspondence and requests for materials should be addressed to H.-L.L. or D.-Q.P.