scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of soybean aphid

Shaolong Qiu[1,4], Ningning Wu[1,4], Xiaodong Sun[2,4], Yongguo Xue[3] & Jixing Xia[1] ✉

Soybean aphid (*Aphis glycines*) is one of the main pests on soybeans, which causes serious damage to the soybean worldwide. The current genome of the soybean aphid is quite fragmented, which has impeded scientific research to some extent. In this study, we assembled a chromosome-level genome of the soybean aphid using MGI short reads, PacBio HiFi long reads and Hi-C reads. The genome sequence was anchored to four pseudo-chromosomes, with a total genome length of 324 Mb and a scaffold N50 length of 88.85 Mb. We evaluated the genome based on insecta_odb10 and the results show it has a completeness of 97.2%. A total of 20,781 protein-coding genes were predicted in the genome, of which 17,183 genes were annotated in at least one protein database. Our work provides a new genomic resource for the soybean aphid study.

Background & Summary

Soybean aphid (*Aphis glycines*), an oligophagous pest of Hemiptera Aphididae, is a heteroecious and holocyclic insect[1,2]. The whole life cycle of soybean aphid includes eggs, nymphs and adults, which need to be completed on different host plants[3,4]. The soybean aphid reproduce sexually on the primary host genus *Rhamnus*, on which it overwinters with eggs[5,6]. The secondary host, soybean, is the host for parthenogenesis of soybean aphid, on which it causes major economic damage[1]. All insects in the family Aphididae harm plants both directly and indirectly, and soybean aphid is no exception[2]. The nymphs and adults of soybean aphids can feed on the vascular tissue, such as phloem sap, through their piercing-sucking mouthparts, which perturb the plant's nutritional equilibrium and precipitate a decrease in soybean yields[7]. During the ingesting process, the aphid excretes honeydew that covers the plants, inhibits the plants' photosynthesis, and fosters the proliferation of sooty mold[7,8]. In addition, soybean aphid serves as a vector for several phytovirus, including *soybean mosaic virus* (SMV)[9], *alfalfa mosaic virus* (AMV)[10], and *potato leafroll virus* (PLRV)[11]. Soybean aphids can disseminate these viruses to both host and non-host plants, thereby indirectly causing economic losses.

Although soybean naturally contains some Rag (Resistant to *A. glycines*) genes, the existence of different soybean aphid biotypes has considerably constrained the popularization of aphid-resistant soybean cultivars[12,13]. Therefore, the control of soybean aphids is still primarily relied on pesticides, but the long-term use of insecticides may enhance insect adaptation[14–16]. A high-quality genome contains more accurate sequences and a more complete set of genes, facilitating the selection and study of resistance-related genes. However, the current genome of soybean aphid is quite fragmented due to technical limitation (Table S1)[17–34].

In this study, to obtain a high-quality genome with improved continuity, we completed the sequencing and assembly of the soybean aphid genome using a combination of MGI short-read sequencing, PacBio high-fidelity (HiFi) sequencing and chromosome conformation capture (Hi-C) sequencing. We obtained a chromosome-level genome assembly of the soybean aphid with a size of 324 Mb. Our study provides the first chromosome-level genome assembly for soybean aphid, which will contribute to clarifying the molecular mechanisms of adaptation.

Methods

Insect rearing and sample collection. Soybean aphids used in this study were collected from a soybean field in Harbin, Heilongjiang Province, China. A laboratory population was established from an apterous female adult. The insects were reared in 50 × 34 × 50 cm cages under conditions of 26 ± 1 °C, a photoperiod of 16:8 (L: D), and a relative humidity of 65 ± 5%. Approximately 150 apterous adults were selected as samples for MGI, PacBio

[1]State Key Laboratory of Agricultural and Forestry Biosecurity, MOA Key Lab of Pest Monitoring and Green Management, College of Plant Protection, China Agricultural University, Beijing, 100193, China. [2]Heilongjiang Province Agro-technical Extension Station, Harbin, 150090, China. [3]Institute of Soybean Research, Heilongjiang Provincial Academy of Agricultural Sciences, Harbin, 150086, China. [4]These authors contributed equally: Shaolong Qiu, Ningning Wu, Xiaodong Sun. ✉e-mail: jixingxia@cau.edu.cn

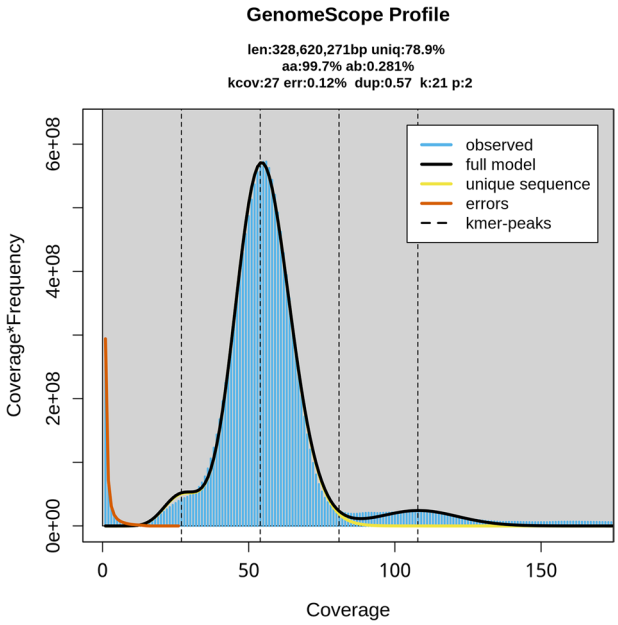

Fig. 1 The characteristics of *A. glycines* genome estimated using *k*-mer distribution (k = 21).

	Summary
Total Length (bp)	324,004,516
Contig N50 (bp)	54,110,121
Scaffold N50 (bp)	88,848,336
The longest length (bp)	88,971,736
The shortest length (bp)	17,878
GC content (%)	27.16
BUSCO genes	C: 97.2% [S: 93.9%, D: 3.3%], F: 0.6%

Table 1. Chromosome level genome assembly statistics of *A. glycines*.

HiFi, and Hi-C sequencing, respectively. The samples were then cleaned twice with 1 × phosphate-buffered saline (PBS) and ultrapure water. After drying with absorbent paper, the samples were placed in 5 mL centrifuge tubes, flash frozen with liquid nitrogen, and stored at −80 °C. Apterous and alate female adults were placed in 1.5 mL nuclease-free centrifuge tubes, frozen in liquid nitrogen, and stored at −80 °C for transcriptome sequencing.

DNA extraction and genome sequencing. Genomic DNA was isolated from the sample using the CTAB method and purified using the Grandomics Genomic kit. A total of 72,326,178 paired-reads were obtained after sequencing the genome short-read library. For PacBio HiFi sequencing, a PacBio HiFi library was constructed using the SMRTbell® prep kit 3.0, which was sequenced on the PacBio Revio device following the operation manual, resulting in 919,364 of high-quality reads. To assemble the genome at the chromosome level, we performed Hi-C sequencing. In short, we cross-linked the cells with 1% formaldehyde for 10 min, followed by cutting the DNA with the restriction endonuclease DpnII. The Hi-C library was constructed according to the NEBNext Ultra II DNA library Prep Kit and sequenced on the MGI 2000 platform, resulting in approximately 159,908,784 paired-reads. The TRIzol method was used to extract total RNA from tissues for transcriptome sequencing. After the samples passed quality control, a sequencing library was constructed, and transcriptome sequencing was completed on the MGI 2000 platform. Finally, a total of 100.6 G of sequencing data was obtained, with an average data volume of 8.4 G.

Genome survey. The PacBio HiFi sequencing data were filtered using Fastp v0.23.4[35]. Jellyfish v2.2.10[36] and GenomeScope v2.0[37] were used to estimate genome size and heterozygosity based on k-mers. When k = 21, the genome size was about 328.62 Mb, and the heterozygosity was 0.281% (Fig. 1).

Genome assembly. Before genome assembly, SeqKit v2.8.1[38] was applied to generate statistics on PacBio HiFi reads, and the N50 length was about 21 kb. HiFi reads were employed as input data for preliminary genome assembly with Hifiasm v0.19.8[39] (with the parameter of -l 2), obtaining a genome containing 58 contigs with an N50 length of 54.11 Mb and a total size of 331.59 Mb. The method of removing symbiotic bacterial contamination from genomes after assembly was adopted, and the contamination sequences were identified in the preliminary assembly results using FCS-GX v0.5.0[40] according to the operation manual. These results showed that the

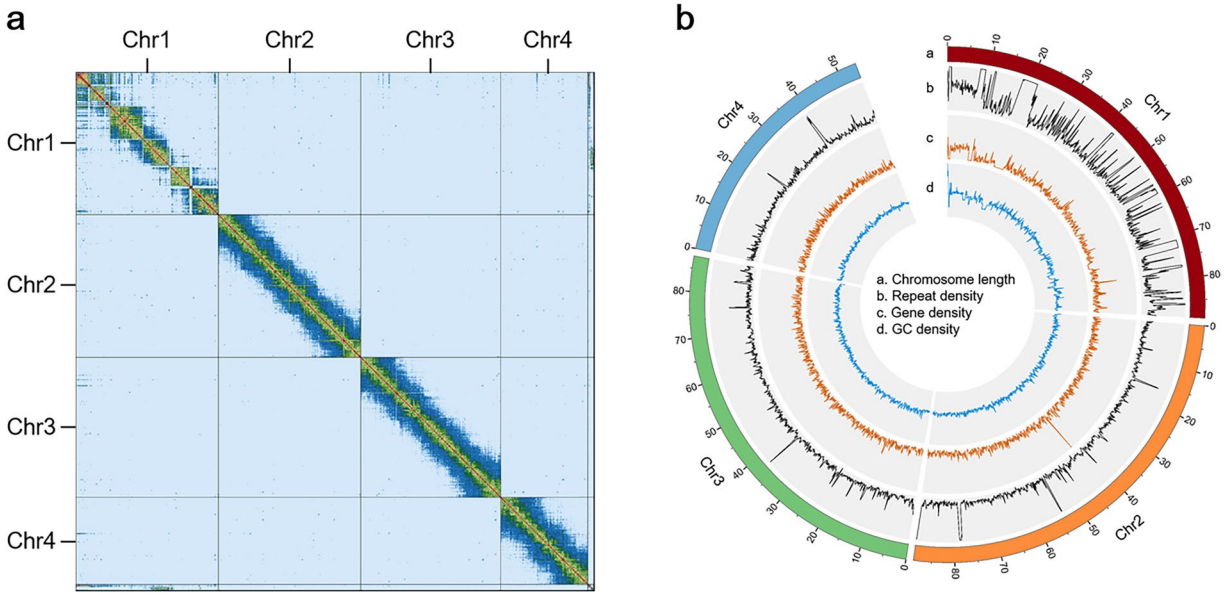

Fig. 2 Genome-wide Hi-C heatmap and circos plot of the *A. glycines* Genome. (**a**) The Hi-C contact heatmap of the *A. glycines* genome. The boundary indicates that the genome contains four chromosomes. (**b**) The circos plot of the *A. glycines* genomic features. The four tracks represent chromosome length, repeat density, gene density and GC density from the outermost to the innermost. The window size was defined as 100 kb.

	Number of elements	Length occupied (bp)	Percentage (%)
Retroelements	28,668	6,712,406	2.07
SINEs	360	46,389	0.01
LINEs	16,566	3,797,719	1.17
L2/CR1/Rex	3,938	527,095	0.16
R1/LOA/Jockey	4,066	1,013,422	0.31
R2/R4/NeSL	266	135,965	0.04
RTE/Bov-B	3,708	670,462	0.21
LTR elements	11,742	2,868,298	0.89
BEL/Pao	1,549	645,867	0.2
Ty1/Copia	511	45,737	0.01
Gypsy/DIRS1	9,603	2,111,781	0.65
DNA transposons	155,128	35,538,877	10.97
hobo-Activator	47,453	8,934,762	2.76
Tc1-IS630-Pogo	10,587	1,590,714	0.49
MULE-MuDR	7,734	1,571,244	0.48
Tourist/Harbinger	1,226	258,780	0.08
Rolling-circles	8,754	2,032,965	0.63
Unclassified	101,684	41,190,113	12.71
Small RNA	1,165	944,693	0.29
Satellites	182	49,842	0.02
Simple repeats	325,384	14,970,350	4.62
Low complexity	48,488	2,423,603	0.75
Total		103,860,962	32.06

Table 2. Classification and statistics of repetitive sequences in *A. glycines* genome.

assembled genome contained 20 contamination sequences, which derived from *Buchnera aphidicola*, *Wolbachia endosymbiont*, *Arsenophonus endosymbiont*, and *Candidatus Blochmannia ocreatus*, respectively.

The assembled contigs were anchored to chromosomes based on Hi-C data using Juicer v1.6[41] and 3D-DNA v201008[42] After manually checking and correcting in Juicebox v2.15[43], 3D-DNA was run again. The pipeline finally generated a chromosome-level genome assembly at 324 Mb, with the longest chromosome length of 88.97 Mb and the shortest chromosome length of 54.11 Mb (Table 1). The Hi-C contact map was visualized with HiGlass v1.13.3[44]. Approximately 319.53 Mb (98.62%) of the sequences were anchored to four chromosomes (Fig. 2b), which is the consistent with the previous observed karyotype[45].

Database	Annotation gene num	Percentage (%)
NR	17,139	82.47
EggNOG	14,752	70.99
GO	6,270	30.17
Swissprot	10,730	51.63
Pfam	10,927	52.58
IPR	10,700	51.49
BUSCO genes	C: 96.7% [S: 93.1%, D:3.6%], F: 1.0%	

Table 3. Functional annotation of *A. glycines* genome.

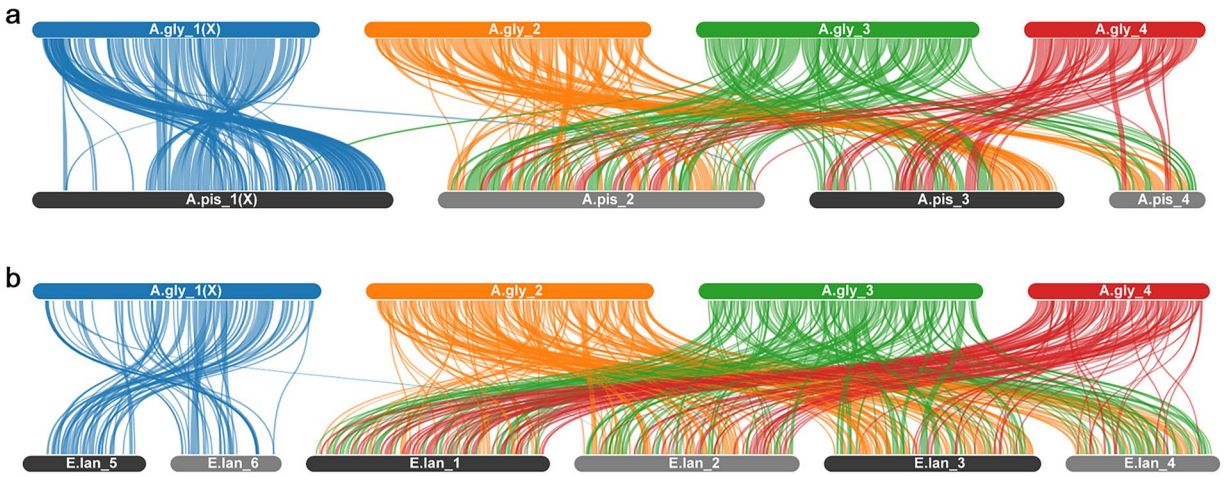

Fig. 3 Genome synteny analyses of *A. glycines* and two aphids. (**a**) Genome synteny analysis between A.gly and A.pis. (**b**) Genome synteny analysis between A.gly and E.lan. A.gly refers to *A. glycines*, A.pis refers to *A. pisum* and E.lan refers to *E. lanigerum*.

Repeat element annotation. The species-specific repeat sequence library was built using RepeatModeler[46]. Based on the arthropod repeat sequence library from Repbase v20181026[47] and the repeat sequence library predicted by RepeatModeler, RepeatMasker v4.1.5[48] was used to soft mask (-xsmall) the repeat sequence. A total of 103.86 Mb repeat sequences were identified, accounting for 32.06% of the entire genome (Table 2). Tandem repeat elements were identified using TRF v4.09.1[49].

Gene prediction and functional annotation. In order to obtain a more accurate gene set, we used the RNA-seq based BRAKER3 pipeline[50] to predict gene structure. In short, *de novo* prediction of genes was mainly performed using GeneMark-ETP v1.02[51] and Augustus v3.5.0[52]. The transcriptome-based prediction was performed by Hisat2 v2.2.1[53] and StringTie v2.2.1[54,55]. BRAKER3 predicted a total of 20,781 protein-coding genes and 25,231 transcripts. The transcriptome data were partially sourced from this study and partially from NCBI SRA database. The downloaded transcriptome data accession numbers are SRP327988[56], SRP442783[57], and SRP442816[58].

Blast v2.15.0[59], Eggnog-Mapper v2.1.2[60,61] and InterproScan v5.66–98.0[62,63] were applied to search NR, Swissprot, Pfam, eggNOG and GO databases to complete functional annotation of predicted genes. A total of 17,183 (82.69%) genes were annotated in at least one database (Table 3).

For the annotation of non-coding RNA tRNA was annotated by tRNAscan-SE v2.0.12[64]. Infernal v1.1.5[65] and Rfam were employed to annotate other ncRNAs.

Genome synteny analysis. BLAST (with the parameters of -evalue 1e-10 -num_alignments 5) was utilized to perform an alignment between the protein sequences annotated in this study and the protein sequences of *A. pisum* and *E. lanigerum*. MCScanX[66] was applied to analyze the genome synteny. These results were visualized with SYNVISIO (https://synvisio.github.io). These results indicate the longest chromosome may be the chromosome X of soybean aphid (Fig. 3).

Data Record

The raw genome and transcriptome sequencing data generated in this study have been deposited in the National Center for Biotechnology Information (NCBI) SRA database. The accession number of DNA-Seq is SRP537912[67], and the accession number of RNA-Seq is SRP538390[68]. The final chromosome level genome assembly data has been submitted to NCBI GenBank and National Genomics Data Center (NGDC) with the accession number of JBJIER000000000[69] and GWHFGPW00000000.1[70]. Genome annotation file is available at the Figshare database[71].

Technical Validation

Benchmarking Universal Single-Copy Orthologs (BUSCO v5.7.1[72]) was used to verify the integrity of the genome and annotation. These results showed that 97.2% of the complete BUSCOs in insecta_odb10 were present in the genome, with 93.9% single-copy genes and 3.3% duplicated genes (Table 1). And the completeness of predicted protein is 96.7% (Table 3).

Code availability

In this study, no custom codes or scripts were used. The software and pipelines mentioned above were executed with default parameters unless specifically indicated.

Received: 14 November 2024; Accepted: 26 February 2025;
Published online: 05 March 2025

References

1. Ragsdale, D. W., Landis, D. A., Brodeur, J., Heimpel, G. E. & Desneux, N. Ecology and management of the soybean aphid in North America. *Annu Rev Entomol.* **56**, 375–399 (2011).
2. Shih, P. Y., Sugio, A. & Simon, J. C. Molecular mechanisms underlying host plant specificity in aphids. *Annu Rev Entomol.* **68**, 431–450 (2023).
3. Ragsdale, D. W., Voegtlin, D. J. & O'neil, R. J. Soybean aphid biology in North America. *Ann Entomol Soc Am.* **97**, 204–208 (2004).
4. Wu, Z., Schenk-Hamlin, D., Zhan, W., Ragsdale, D. W. & Heimpel, G. E. The soybean aphid in China: a historical review. *Ann Entomol Soc Am.* **97**, 209–218 (2004).
5. Wang, C. L., Xiang, L. Y., Zhang, G. X. & Zhu, H. F. Studies on the soybean aphid, *Aphid glycines* Matsumura. *Acta Entomol. Sinica.* **11**, 31–44 (1962).
6. Hill, C. B., Chirumamilla, A. & Hartman, G. L. Resistance and virulence in the soybean-*Aphis glycines* interaction. *Euphytica.* **186**, 635–646 (2012).
7. Beckendorf, E. A., Catangui, M. A. & Riedell, W. E. Soybean aphid feeding injury and soybean yield, yield components, and seed composition. *Agron. J.* **100**, 237–246 (2008).
8. He, F. G. *et al*. Optimal spraying time and economic threshold of the soybean aphid. *Acta Phytopathol. Sin.* **18**, 155–159 (1991).
9. Clark, A. J. & Perry, K. L. Transmissibility of field isolates of soybean viruses by *Aphis glycines*. *Plant Dis.* **86**, 1219–1222 (2002).
10. Davis, J. A. & Radcliffe, E. B. The importance of an invasive aphid species in vectoring a persistently transmitted potato virus: *Aphis glycines* is a vector of *potato leafroll virus*. *Plant Dis.* **92**, 1515–1523 (2008).
11. Guo, H. *et al*. Salivary carbonic anhydrase II in winged aphid morph facilitates plant infection by viruses. *Proc Natl Acad Sci USA.* **120**, e2222040120 (2023).
12. Natukunda, M. I. *et al*. Interaction between Rag genes results in a unique synergistic transcriptional response that enhances soybean resistance to soybean aphids. *BMC genomics.* **22**, 887 (2021).
13. Natukunda, M. I. & MacIntosh, G. C. The resistant soybean-*Aphis glycines* interaction: current knowledge and prospects. *Front Plant Sci.* **11**, 1223 (2020).
14. Koch, R. L., Hodgson, E. W., Knodel, J. J., Varenhorst, A. J. & Potter, B. D. Management of insecticide-resistant soybean aphids in the Upper Midwest of the United States. *J Integr Pest Manag.* **9**, 23 (2018).
15. Menger, J. P. *et al*. Lack of evidence for fitness costs in soybean aphid (Hemiptera: Aphididae) with resistance to pyrethroid insecticides in the upper midwest region of the United States. *J Econ Entomol.* **115**, 1191–1202 (2022).
16. Panini, M. *et al*. Transposon-mediated insertional mutagenesis unmasks recessive insecticide resistance in the aphid *Myzus persicae*. *Proc Natl Acad Sci USA*. **118**, e2100559118 (2021).
17. Mathers, T. C. Improved genome assembly and annotation of the soybean aphid (*Aphis glycines* Matsumura). *G3 (Bethesda).* **10**, 899–906 (2020).
18. Zhang, S. *et al*. Chromosome-level genome assemblies of two cotton-melon aphid *Aphis gossypii* biotypes unveil mechanisms of host adaption. *Mol Ecol Resour.* **22**, 1120–1134 (2022).
19. Mathers, T. C., Mugford, S. T., Hogenhout, S. A. & Tripathi, L. Genome sequence of the banana aphid, *Pentalonia nigronervosa* Coquerel (Hemiptera: Aphididae) and its symbionts. *G3 (Bethesda).* **10**, 4315–4321 (2020).
20. Chen, W. *et al*. Genome sequence of the corn leaf aphid (*Rhopalosiphum maidis* Fitch). *Gigascience.* **8**, giz033 (2019).
21. Wang, Y. & Xu, S. A high-quality genome assembly of the waterlily aphid *Rhopalosiphum nymphaeae*. *Sci. Data.* **11**, 194 (2024).
22. Mathers, T. C. *et al*. Chromosome-scale genome assemblies of aphids reveal extensively rearranged autosomes and long-term conservation of the X chromosome. *Mol. Biol. Evol.* **38**, 856–875 (2021).
23. Wu, J. *et al*. A chromosome-level genome assembly of the cabbage aphid *Brevicoryne brassicae*. *Sci. Data*. **12**, 167 (2025).
24. Nicholson, S. J. *et al*. The genome of *Diuraphis noxia*, a global aphid pest of small grains. *BMC genomics.* **16**, 1–16 (2015).
25. Ye, S. *et al*. A chromosome-level genome assembly of *Neotoxoptera formosana* (Takahashi, 1921)(Hemiptera: Aphididae)[J]. *G3 (Bethesda).* **12**, jkac164 (2022).
26. Byrne, S. *et al*. Genome sequence of the English grain aphid, *Sitobion avenae* and its endosymbiont *Buchnera aphidicola*. *G3 (Bethesda).* **12**, jkab418 (2022).
27. Jiang, X. *et al*. A chromosome-level draft genome of the grain aphid *Sitobion miscanthi*. *Gigascience.* **8**, giz101 (2019).
28. Wei, H. Y. *et al*. Chromosome-level genome assembly for the horned-gall aphid provides insights into interactions between gall-making insect and its host plant. *Ecol Evol.* **12**, e8815 (2022).
29. Xu, S. *et al*. Two chromosome-level genome assemblies of galling aphids *Slavum lentiscoides* and *Chaetogeoica ovagalla*. *Sci. Data.* **11**, 803 (2024).
30. Julca, I. *et al*. Phylogenomics identifies an ancestral burst of gene duplications predating the diversification of Aphidomorpha. *Mol Biol Evol.* **37**, 730–756 (2020).
31. Crowley, L. M. & James, R. Darwin Tree of Life Consortium. The genome sequence of the Common Sycamore Aphid, *Drepanosiphum platanoidis* (Schrank, 1801). *Wellcome Open Res.* **8**, 481 (2023).
32. Biello, R. *et al*. A chromosome-level genome assembly of the woolly apple aphid, *Eriosoma lanigerum* Hausmann (Hemiptera: Aphididae). *Mol. Ecol. Resour.* **21**, 316–326 (2021).
33. Renoz, F. *et al*. PacBio Hi-Fi genome assembly of *Sipha maydis*, a model for the study of multipartite mutualism in insects. *Sci. Data.* **11**, 450 (2024).
34. Huang, T. *et al*. Chromosome-level genome assembly of the spotted alfalfa aphid *Therioaphis trifolii*. *Sci. Data.* **10**, 274 (2023).
35. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics.* **34**, i884–i890 (2018).
36. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics.* **27**, 764–770 (2011).
37. 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).

38. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PloS one.* **11**, e0163962 (2016).
39. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat Methods.* **18**, 170–175 (2021).
40. Astashyn, A. *et al*. Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol.* **25**, 60 (2024).
41. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
42. Dudchenko, O. *et al*. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science.* **356**, 92–95 (2017).
43. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell syst.* **3**, 99–101 (2016).
44. Kerpedjiev, P. *et al*. HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol.* **19**, 125 (2018).
45. Mandrioli, M. *et al*. Analysis of the extent of synteny and conservation in the gene order in aphids: A first glimpse from the *Aphis glycines* genome[J]. *Insect Biochem Mol Biol.* **113**, 103228 (2019).
46. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA.* **117**, 9451–9457 (2020).
47. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA.* **6**, 1–6 (2015).
48. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics.* **25**, 4–10 (2009).
49. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
50. Gabriel, L. *et al*. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* **34**, 769–777 (2024).
51. Bruna, T., Lomsadze, A. & Borodovsky, M. GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes. *Genome Res.* **34**, 757–768 (2024).
52. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
53. 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).
54. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol.* **33**, 290–295 (2015).
55. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 1–13 (2019).
56. *NCBI Sequence Read Archive* https://identifiers.org/insdc.sra:SRP327988 (2022).
57. *NCBI Sequence Read Archive* https://identifiers.org/insdc.sra:SRP442783 (2023).
58. *NCBI Sequence Read Archive* https://identifiers.org/insdc.sra:SRP442816 (2023).
59. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC bioinformatics.* **10**, 1–9 (2009).
60. 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).
61. Huerta-Cepas, J. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
62. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics.* **30**, 1236–1240 (2014).
63. Blum, M. *et al*. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354 (2021).
64. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* **49**, 9077–9096 (2021).
65. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics.* **29**, 2933–2935 (2013).
66. Wang, Y. *et al*. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
67. *NCBI Sequence Read Archive* https://identifiers.org/insdc.sra:SRP537912 (2024).
68. *NCBI Sequence Read Archive* https://identifiers.org/insdc.sra:SRP538390 (2024).
69. *NCBI GenBank* http://identifiers.org/insdc:JBJIER010000000 (2024).
70. National Genomics Data Center (NGDC). *Genome Warehouse* https://ngdc.cncb.ac.cn/gwh/Assembly/86284/show (2024).
71. Qiu, S.-L., Wu, N.-N., Sun, X.-D., Xue, Y.-G. & Xia, J.-X. Chromosome-level genome assembly of soybean aphid, *Aphis glycines*. *figshare.* https://doi.org/10.6084/m9.figshare.27221433.v3 (2024).
72. Simão, F. A. *et al*. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* **31**, 3210–3212 (2015).

Acknowledgements

This study was supported by the Pinduoduo-China Agricultural University Research Fund (PC2024B01010 to J. X.), the National Top Young Talents Program of China, and the National Natural Science Foundation of China (32100330 to N. W.).

Author contributions

J.X., N.W. and S.Q. conceived this study. S.Q., X.S. and Y.X. collected the samples and prepared DNA and RNA for sequencing. S.Q., N.W. and J.X. analyzed the data. S.Q. wrote the manuscript. J.X. and N.W. revised the manuscript. All authors have reviewed and approved the 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/s41597-025-04711-8.

Correspondence and requests for materials should be addressed to J.X.