scientific data

Check for updates

OPEN

DATA DESCRIPTOR

The chromosome-level genome assembly of *Sitobion avenae* Fabricius (Hemiptera: Aphididae)

Jingting Wang[1], Jiaqi Zhang[1], Chuting Shi[1], Hanlin Wang[2], Yucheng Zhang[2], Yufan Li[2], Tianye Hu[2], Huihui Zhang[1] & Zewen Liu[1] ✉

Sitobion avenae **(Fabricius), a major pest of cereal crops worldwide, causes significant damage on agricultural production. However, the lack of high-quality genomes has limited in-depth studies on its biology and effective management strategies. Here, we report a chromosome-level genome assembly of *S. avenae*, generated by combining PacBio long-read sequencing, Illumina short-read sequencing, and Hi-C scaffolding. The assembled genome has a size of 421 Mb, with 99.97% of sequences anchored to nine pseudochromosomes. It comprises only 11 scaffolds, exhibits the scaffold N50 and contig N50 of 39.84 Mb and 28.00 MB, and achieves a BUSCO completeness of 97.30%. Repetitive elements account for 37.40% (157.66 Mb) of the assembly. We annotated 15,369 protein-coding genes and 1,188 non-coding RNAs. This high-quality genome of *S. avenae* serves as a valuable genomic resource for advancing research on pest management and host adaptation mechanisms in this aphid species.**

Background & Summary

The English green aphid, *Sitobion avenae* (Fabricius) (Hemiptera: Aphididae), is one of the most widespread and economically significant cereal aphids worldwide. It exhibits a non-host-alternating holocyclic life cycle characterized by seasonal shifts in reproductive mode[1,2]. Throughout most of the growing season, populations reproduce parthenogenetically over multiple successive generations. As temperatures decline, virginoparae produce sexuparae, which subsequently generate oviparae and males. *S. avenae* infests a wide range of cereal crops and grasses, including barley, wheat, oats, rice, and various grasses[3]. Both nymphs and adults impair nutrient assimilation and allocation in cereal crops through feeding on the phloem sap, which leads to substantial yield losses. During the probing and feeding process, honeydew secreted by the green aphids accumulates on the plant surface, which inhibits photosynthesis and promoting the growth of sooty mold. Furthermore, *S. avenae* serves as an important vector of barley yellow dwarf virus (BYDV) (*Luteoviridae: Luteovirus*)[4–6], a major threat to global cereal production.

In cereal cropping systems, the control of *S. avenae* relies predominantly on chemical pesticides. However, intensive and prolonged application has led to the development of varying levels of resistance to several classes of insecticides, such as neonicotinoids and pyrethroids[7–9]. The lack of high-quality reference genome has significantly hindered the identification and functional characterization of resistance-associated genes in this species. A high-quality genome assembly would not only provide more accurate gene models and a complete gene set, but also serve as a valuable foundation for elucidating the genetic mechanisms underlying virus transmission and for identifying key genomic loci involved in overcoming host defenses. Despite recent advances in sequencing technologies, chromosome-level genomes of aphids are still limited. To date, only a limited number of species have been resolved at this resolution, including *Myzus persicae*, *Acyrthosiphon pisum*, *Rhopalosiphum maidis*, *Aphis gossypii*, and *Aphis glycines*[10–12]. Moreover, the current available genome assembly for *S.avenae* remains highly fragmented, constraining in-depth molecular and evolutionary studies[13,14].

In this study, we generated a chromosome-level genome assembly of the English green aphid, Sitobion avenae, through a combination of Illumina short-read sequencing, PacBio high-fidelity (HiFi) sequencing and chromosome conformation capture (Hi-C) technologies. As the first chromosome-level reference genome of

[1]Key Laboratory of Biohazard Monitoring and Green Prevention and Control for Artificial Grassland, Ministry of Agriculture and Rural Affairs, Institute of Grassland Research, Chinese Academy of Agricultural Sciences, Hohhot, 010010, China. [2]State Key Laboratory of Agricultural and Forestry Biosecurity, Nanjing Agricultural University, Nanjing, 210095, China. ✉e-mail: liuzewen@caas.cn

Libraries	Clean data (Gb)	Insert sizes (bp)	Sequencing coverage (X)
WGS	41.16	150	97.62
HiFi	41.92	16,380.32	99.42
Hi-C	46.56	150	110.42
RNA-seq (short-read)	44.81	150	—
RNA-ONT	10.09	1,451.52	—

Table 1. Summary of the sequencing data generated for *Sitobion avenae* genome assembly and annotation.

S. avenae, this resource provides a critical foundation for future investigations into the host adaptation mechanisms and the identification of key genes associated with insecticide resistance.

Methods

Insect rearing and sample collection. *S. avenae* were collected from a cereal crop field in Yangling, Shannxi Province, China (34°27′27″N, 108°08′11″E) in 2016. The laboratory population was established from one apterous female adult. The aphids were then fed on wheat (*Triticum aestivum*) under standard conditions of 22 ± 1 °C, 65 ± 5% relative humidity with a 16 h:8 h (light: dark) photoperiod, without exposure to any pesticides.

The adult *S. avenae* were first placed into culture dishes and starved for 24 h. Approximately 200 adults were collected as samples for Illumina, PacBio HiFi and Hi-C sequence, respectively. In addition, 60 winged and wingless adult aphids were selected fort transcriptome sequencing.

Sample extraction and sequencing. Genomic DNA was extracted from the sample using the CTAB (cetyltrimethylammonium bromide) method and further purified with the Grandomics Genomic kit (Beijing, China). Illumina genome-sequencing was sequenced on an Illumina NovaSeq X Plus platform and was prepared through a 350bp-insert fragment library (150 bp paired-end) by Truseq DNA PCR-free kit (San Diego, CA, USA). For Pacbio HiFi sequencing, the HiFi library of 20 kb insert size was constructed using the SMRTbell® Express Template Prep Kit 3.0 (Pacific Biosciences, CA, USA). Then, the library was sequenced on the PacBio Sequel Revio following the operation manual.

High-throughput chromosome conformation capture (Hi-C) sequencing was performed to achieve chromosome-level genome assembly The Hi-C library preparation procedure involved cross-linking of chromatin, restriction enzyme digestion, end repair, DNA cyclization and purification. The resulting library was sequenced on the NovaSeq X Plus platform (Illumina Inc., San Diego, CA, USA). The total RNA was extracted from the sample using the TRIzol reagent (Invitrogen, Carlsbad, CA, USA). The RNA library was constructed using VAHTS mRNA-seq v2 Library Prep Kit (Vazyme, Nanjing, China) and sequenced through the Illumina NovaSeq X Plus platform. Full-length transcriptome sequencing library was constructed using the SQK-PCS109 and SQK-PBK004 kits (Oxford Nanopore, Oxford, UK), and subsequently sequenced on the Oxford Nanopore PromethION platform. Finally, we generated 41.16 Gb (97×) of Illumina short reads, 41.92 Gb (99×) of PacBio long reads, 46.59 Gb (110×) of Hi-C reads, and 54.89 Gb of transcriptome data, including 44.81 Gb from Illumina and 10.09 Gb from Oxford Nanopore Technology (ONT), as shown in Table 1.

Genome survey. Fastp v0.23.4[15] was employed to perform quality control and preprocessing of the Illumina raw reads. The processing included trimming of low-quality raw reads by trimming low-quality bases, removal of duplicates, clipping of poly-G and adapter sequences, and exclusion of reads containing more than 10% of bases with quality scores ≤ 5 (Q ≤ 5). Additionally, base quality correction was conducted using overlapping read information (parameters: -q 20 -D -g -x -u 10 -c). GenomeScope v2.0[16] was used to estimate genome size and heterozygosity based on k-mer frequency distribution. Under a maximum k-mer coverage cutoff of 1000 and a k-mer size of 21, the estimated genome size was approximately 401.05 Mb with a heterozygosity of 1.31%, indicating a highly heterozygous genome (Fig. 1).

Pbccs v6.4.0 (https://github.com/PacificBiosciences/ccs) was used to generate high-quality HiFi reads, retaining sequences with a base quality of Q30 or higher. The primary assembly was conducted with Hifiasm v0.24.0[17] using the parameter "-l 3". To minimize contamination and assembly errors, only contigs with sequencing depth greater than 10 × (Save) were retained, while those below one-tenth of the average depth were discarded. Owing to the exceptional base accuracy of the assembled contigs, with QV values consistently around 60 (Table S2), additional polishing steps were deemed unnecessary. K-mer frequency analysis indicated that the aphid genome is highly heterozygous due to its diploid nature. To remove redundant heterozygous sequences from the Hifiasm assembly, Purge_dups v1.2.5[18] was run with default parameters based on contig similarity and sequencing depth. Minimap2 v2.29[19,20] was used for sequence alignment, with the options '-x map-hifi' for mapping HiFi reads to the genome and '-x asm5 -DP' for genome self-alignment.

Hi-C sequence data were first processed using Chromap v0.2.69[21] for quality control, including read alignment, duplicate removal, and Hi-C contact extraction. Then, scaffolding was performed with YAHS v1.2[22] under default parameters with two rounds. Hi-C heatmaps were manually inspected and corrected using Juicebox v1.11.08[23] to identify potential errors. MMseqs. 2 v13[24] was used to detect possible contamination (parameters: -min-seq-id 0.8) compared with NCBI nucleotide and UniVec databases. Sequences over 90% alignments were removed. The completeness of genome assembly was evaluated with BUSCO v5.8.2[25] using insecta_odb 10 database. The single base mass score (QV value) and genomic k-mer profile were evaluated by merqury v1.3[26].

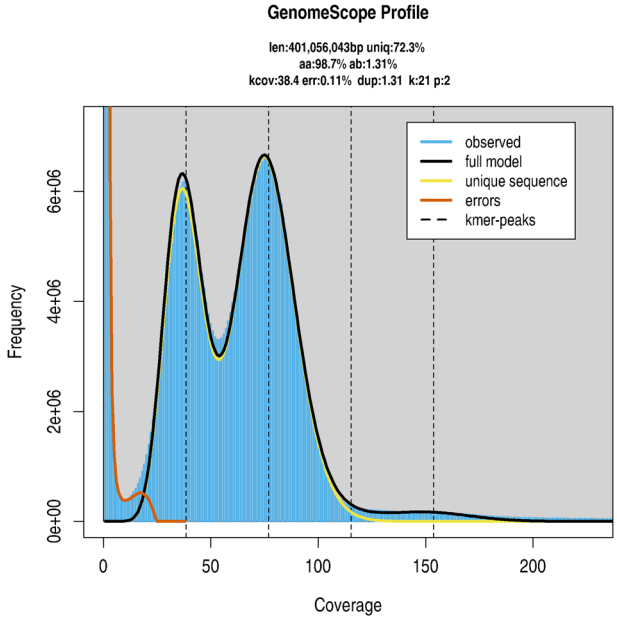

Fig. 1 The genome survey of *Sitobion avenae* at 21-mer estimated by GenomeScope.

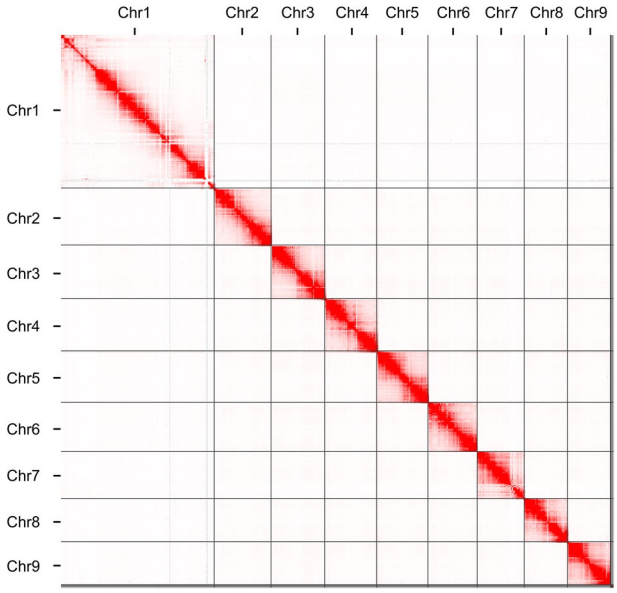

Fig. 2 Genome-wide Hi-C heatmap of the *Sitobion avenae* genome. The boundary indicates that the genome contains 9 chromosomes.

Finally, the chromosome-level genome assembly of *S. avenae* had a size of 421.65 Mb, comprising 11 scaffolds and 30 contigs, with the scaffold and contig N50 of 39.84 Mb and 28.00 Mb, respectively. A remarkable 99.97% (421.51 Mb) of the genome was anchored into 9 pseudochromosomes, with the chromosome length ranging from 32.49 Mb to 118.46 Mb and the GC level of 30.06% (Fig. 2, Table 1, Table S1). This chromosome number is also similar to that reported for its closely related species *Sitobion miscanthi*, which likewise possesses nine chromosomes[27]. BUSCO assessment of *S. avenae* genome assembly indicated a completeness of 97.3%, with 4.0% duplicated BUSCOs, suggesting no obvious redundancy. All these statistics all indicate that our assemblies are remarkably contiguous and comprehensive.

Genome annotation. The species-specific repeat library was built using RepeatModeler v2.0.7[28] with parameters '-LTRStruct'. Then, this species-specific library was merged with Repbase-20181026[29] and Dfam 3.1[30] into a custom library. RepeatMasker v 4.1.7-p1[31] was used to identify the repetitive elements in *S. avenae* genome and soft mask the repeat sequences of the genome. RepeatMasker result reveled that *S. avenae* genome contains

Characteristics	Values
Genome assembly	
Assembly size (Mb)	421.65
Number of chromosomes	9
Number of scaffolds/contigs	11/30
Longest scaffold/contig (Mb)	118.46/64.66
N50 scaffold/contig length (Mb)	39.84/28.00
GC content (%)	30.06
BUSCO completeness (%)	C: 97.3 [S: 93.3, D: 4], F: 0.4, M: 2.3
Repetitive elements Size (Mb)	157.66 (37.40%)
Repetitive elements	
DNA transposons (Mb)	59.99 (14.23%)
LINEs (Mb)	9.85 (2.34%)
LTRs (Mb)	9.63 (2.28%)
SINEs (kb)	35.09 (0.01%)
Simple repeats (Mb)	11.57 (2.74%)
Unclassified (Mb)	39.26 (9.31%)
Protein-coding genes	
Number	15,369
Mean gene length (bp)	11,662.8
BUSCO completeness (%)	C: 97.0 [S: 84.2, D: 12.8], F: 0.5, M: 2.5

Table 2. Genome and annotation statistics for the chromosome-level assembly of *Sitobion avenae*. C represents complete BUSCOs, S represents single BUSCOs, D represents duplicated BUSCOs, F represents fragile BUSCOs, M represents missing BUSCOs.

37.40% repetitive elements, including DNA transposons (14.23%), unclassified elements (14.20%), simple repeats (2.74%), LINEs (2.34%), LTRs (2.28%) and other elements (Table 2).

Infernal v1.1.423[32] and tRNAscan-SE v2.0.9[33] were used to identify ncRNAs and tRNAs, respectively. Low-confidence of tRNAs was filtered with 'EukHighConfidenceFilter' script in tRNAscan-SE software. Finally, we obtained 1,188 ncRNAs, including 77 rRNAs, 112 miRNAs, 110 snRNAs, 271 tRNAs, 372 ribozymes and 4 lncRNAs (Table S3).

Maker v3.01.03[34] was used to annotate protein-coding genes, by combining results from *ab initio* gene predictions, transcriptome sequencing data and protein homology searches. For *ab initio* gene predictions, BRAKER v3.0.6[35] was utilized, employing GeneMark-ES/ET/EP 4.68_lic[36] and Augustus v3.4.0[37]. Moreover, this software automatically trained them based on transcriptome database and OrthoDB12database[38]. Illumina short-read transcriptome data were aligned to *S. avenae* reference genome using Hisat2 v2.2.1[39], while ONT reads were mapped with Minimap2 v2.29[19,20], and BAM files were subsequently generated with SAMtools v1.8[40]. Transcript-based gene prediction was then conducted by integrating RNA-seq and ONT-seq assemblies with StringTie v2.1.5[41] in "–merge" mode. Subsequently, homology proteins from the five closed species, including *Thrips palmi* (GCF_012932325.1), *Drosophila melanogaster* (GCF_000001215.4), *Rhopalosiphum maidis* (GCF_003676215.2), *Rhodnius prolixus* (GCF_049639745.1) and *Metopolophium dirhodum* (GCF_019925205.1) were incorporated using GeMoMa v1.9[42] with the parameters "GeMoMa.c = 0.4 GeMoMa.m = 72000". Finally, we predicted 15,369 protein-coding genes in *S. avenae* genome, with an average length 11,662.8 bp. The average number of exons, introns, and CDS of each gene were 7.2, 6.2, 6.9, respectively, and their mean lengths were 306.8 bp, 1595.5 bp, 210.6 bp, correspondingly. The completeness of these protein-coding genes from BUSCO was 97.0%, which is comparable to the genome assembly assessment (Table 2, Table S4). TBtools[43] was used produce the visual diagram of S. avenae genomic characteristics (chromosome length, GC content, gene density, DNA transposons, LINE, SINE, LTR and simple repeat density) (Fig. 3).

Gene functional annotation search of protein-coding genes was performed using two methods. First, Diamond v2.1.7.161[44] was performed to search the SwissProt and TrEMBL databases with the parameters "-very-sensitive, -e 1e-5". Then, eggNOGmapper v2.1.12[45] and InterProScan 5.70–102.0[46] were utilized to assign Gene Ontology (GO) and (KEGG, Reactome) pathway annotations and to identify protein domains. The InterProScan analysis included five databases: Pfam[47], SMART[48], Superfamily[49], Gene3D[50], and CDD[51]. Subsequently, the results predicted by the above tools were integrated to obtain the final gene function prediction. Finally, 9,390 GO terms, 4,668 KEGG pathways, 2,533 enzyme codes and 11,431 COG categories were assigned by integrating the InterProScan and eggNOG annotation results (Table S5).

From inner to outer circles: simple sequence repeats (SSRs), long terminal repeats (LTRs), long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), DNA transposon density, protein-coding gene density, GC content, and chromosome-scale length.

Data Records

The genome sequencing data generated in this study have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession number PRJNA1282322. The NCBI accession number of the assembled

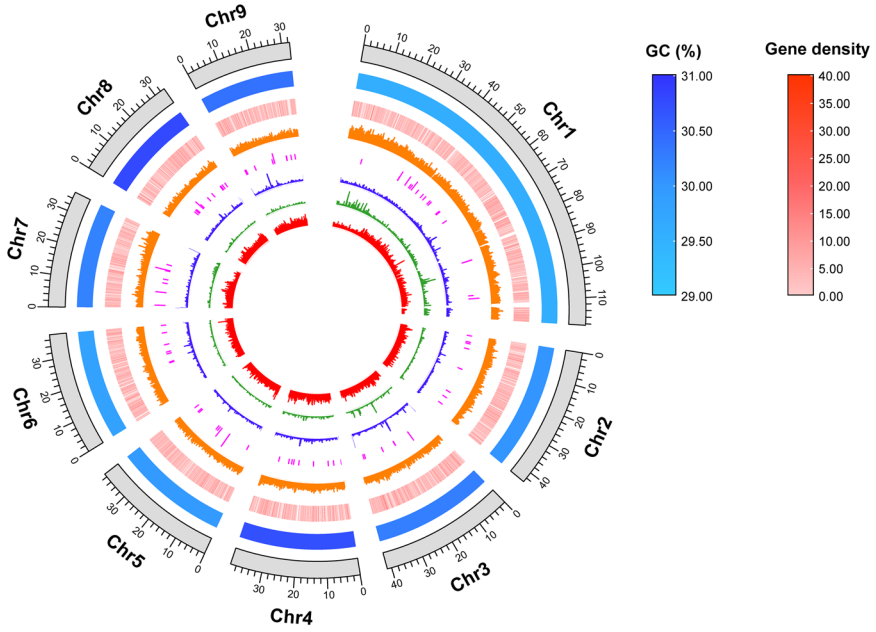

Fig. 3 Circos plot depicting genomic features of *S. avenae*.

genome is GCA_053477475.1[52]. The Illumina, Hi-C, HiFi, and transcriptome data can be found under identification numbers SRR35536597[53], SRR35536595[54], SRR35536596[55], SRR35536594[56]. Genome annotation files have been deposited in Figshare and can be accessed at *Figshare*[57].

Technical Validation

Two methods were used to evaluate the quality of the genome assembly. BUSCO v5.82 was used to assess the completeness of *S. avenae* genome with the insect_odb10 database (n = 1,367). The BUSCO completeness of *S. avenae* genome was 97.3%, comprising single-copy of 93.3%, duplicated BUSCOs of 4.0%, fragmented BUSCOs of 0.4%, and missing BUSCOs of 2.3%. The mapping rate from aligning PacBio, Illumina, and RNA reads were used to evaluate the genome assembly accuracy. The mapping rates for PacBio, Illumina and RNA-seq were 98.35%, 95.54%, and 98.58%, respectively. These comprehensive evaluations demonstrate the high contiguity, completeness, and accuracy of the genome assembly presented in this study.

Data availability

All raw sequencing data generated in this study, including Illumina (https://identifiers.org/ncbi/insdc.sra:SRR35536597), Hi-C (https://identifiers.org/ncbi/insdc.sra:SRR35536595), PacBio HiFi (https://identifiers.org/ncbi/insdc.sra:SRR35536596) and transcriptome reads (https://identifiers.org/ncbi/insdc.sra:SRR35536594), are available in the NCBI Sequence Read Archive (SRA) under BioProject PRJNA1282322. The assembled genome is available in NCBI GenBank under accession number GCA_053477475.1 (https://identifiers.org/ncbi/insdc.gca:GCA_053477475.1). Genome annotation files have been deposited in the Figshare repository (https://doi.org/10.6084/m9.figshare.30229090).

Code availability

No custom scripts were developed for this work. All data processing and analyses were performed using standard commands and pipelines according to the manual and protocols of the corresponding bioinformatic tools.

Received: 23 December 2025; Accepted: 27 March 2026;
Published online: 14 April 2026

References

1. Helden, A. J. & Dixon, A. Life-cycle variation in the aphid *Sitobion avenae*:: costs and benefits of male production. *Ecol Entomol.* **27**, 692–701 (2002).
2. Simon, J. C., Rispe, C. & Sunnucks, P. Ecology and evolution of sex in aphids. *Trends Ecol Evol.* **17**, 34–39 (2002).
3. Luo, C. *et al*. Comparison of life-history traits and resistance for *Sitobion avenae* (Fabricius) harboring a facultative symbiont. *Entomol Gen.* **40**, 39–47 (2020).
4. Li, D. *et al*. Virus-Dependent and -Independent Responses of *Sitobion avenae* (Homoptera: Aphididae) Feeding on Wheat Infected by Transmitted and Nontransmitted Viruses at Transcriptomic Level. *J Econ Entomol.* **112**, 2067–2076 (2019).
5. Schughart, M. *et al*. Insecticide Resistance in the English Grain Aphid *Sitobion avenae* has No Impact on Luteovirus Transmission Efficiency. *Plant Pathol.* **74**, 2137–2149 (2025).
6. Yu, W. *et al*. Variation in the transmission of barley yellow dwarf virus-PAV by different *Sitobion avenae* clones in China. *J Virol Methods.* **194**, 1–6 (2013).
7. Fontaine, S., Caddoux, L. & Barres, B. First report of the kdr pyrethroid resistance mutation in a French population of the English grain aphid, *Sitobion avenae*. *Crop Prot.* **165**, 106153 (2023).

8. Gong P. *et al*. Field evolved resistance to pyrethroids, neonicotinoids, organophosphates and macrolides in *Rhopalosiphum padi* (Linnaeus) and *Sitobion avenae* (Fabricius) from China. *Chemosphere*. **269** (2021).
9. Zhang, B. *et al*. Multiple detoxification genes confer imidacloprid resistance to *Sitobion avenae* Fabricius. *Crop Prot*. **128**, 105014 (2020).
10. 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).
11. Chen W. *et al*. Genome sequence of the corn leaf aphid (Rhopalosiphum maidis Fitch). *Gigascience*. **8** (2019).
12. Qiu S. *et al*. Chromosome-level genome assembly of soybean aphid. *Sci Data*. **12** (2025).
13. Villarroel, C. A. *et al*. Genome sequencing of a predominant clonal lineage of the grain aphid *Sitobion avenae*. *Insect Biochem Molec*. **143**, 103742 (2022).
14. Yan, Y. *et al*. Comprehensive full-Length transcriptomic resource for wing dimorphism in English green aphid, *Sitobion avenae*. *Sci Data*. **12**, 1034 (2025).
15. Chen, S. *et al*. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. **34**, 884–890 (2018).
16. 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).
17. Cheng, H. *et al*. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat Methods*. **18**, 170 (2021).
18. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*. **36**, 2896–2898 (2020).
19. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. **34**, 3094–3100 (2018).
20. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics*. **37**, 4572–4574 (2021).
21. Zhang, H. *et al*. Fast alignment and preprocessing of chromatin profiles with Chromap. *Nat Commun*. **12**, 6566 (2021).
22. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics*. **39**, c808 (2023).
23. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst*. **3**, 95–98 (2016).
24. Steinegger, M. Soeding J. MMseqs. 2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*. **35**, 1026–1028 (2017).
25. Manni, M., Berkeley, M. R. & Seppey, M. *et al*. 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).
26. Rhie, A. *et al*. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. **21**, 245 (2020).
27. Jiang X. *et al*. A chromosome-level draft genome of the grain caphid Sitobion miscanthi. *Gigascience*. **8** (2019).
28. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *P Natl Acad Sci Usa*. **117**, 9451–9457 (2020).
29. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile Dna-Uk*. **6**, 11 (2015).
30. Storer, J. *et al*. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile Dna-Uk*. **12**, 2 (2021).
31. Smit A. F. A., Hubley R., Green P. RepeatMasker Open-4.0. http://www.repeatmasker.org (2013).
32. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*. **29**, 2933–2935 (2013).
33. Chan P. P., Lowe T. M. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. In. Edited by Kollmar M, vol. **2019** 1–14 (1962).
34. Holt C., Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *Bmc Bioinformatics*. **12** (2011).
35. Bruna, T. *et al*. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP plus and AUGUSTUS supported by a protein database. *Nar Genom Bioinform*. **3**, a108 (2021).
36. Bruna, T., Lomsadze, A. & Borodovsky, M. GeneMark-EP plus: eukaryotic gene prediction with self-training in the space of genes and proteins. *Nar Genom Bioinform*. **2**, a26 (2020).
37. Stanke, M. *et al*. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res*. **32**, W309–W312 (2004).
38. Kuznetsov, D. *et al*. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res*. **51**, D445–D451 (2023).
39. Kim, D. *et al*. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. **37**, 907 (2019).
40. Danecek, P. *et al*. Twelve years of SAMtools and BCFtools. *Gigascience*. **10**, b8 (2021).
41. Kovaka S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol*. **20** (2019).
42. Keilwagen J. *et al*. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *Bmc Bioinformatics*. **19** (2018).
43. Chen, C. *et al*. TBtools-II: A "one for all, all for one" bioinformatics platform for biological big-data mining. *Mol Plant*. **16**, 1733–1742 (2023).
44. Buchfink, B., Reuter, K. & Drost, H. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods*. **18**, 366 (2021).
45. Cantalapiedra, C. P., Hernandez-Plaza, A. & Letunic, I. *et al*. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol*. **38**, 5825–5829 (2021).
46. Paysan-Lafosse, T. *et al*. InterPro in 2022. *Nucleic Acids Res*. **51**, D418–D427 (2023).
47. Mistry, J. *et al*. Pfam: The protein families database in 2021. *Nucleic Acids Res*. **49**, D412–D419 (2021).
48. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res*. **46**, D493–D496 (2018).
49. Wilson, D. *et al*. SUPERFAMILY-sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res*. **37**, D380–D386 (2009).
50. Lewis, T. E. *et al*. Gene3D: Extensive prediction of globular domains in proteins. *Nucleic Acids Res*. **46**, D435–D439 (2018).
51. Marchler-Bauer, A. *et al*. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res*. **45**, D200–D203 (2017).
52. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_053477475.1 (2025).
53. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:SRR35536597 (2025).
54. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:SRR35536595 (2025).
55. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:SRR35536596 (2025).
56. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:SRR35536594 (2025).
57. Wang, J. T. The chromosome-level genome assembly of *Sitobion avenae* Fabricius (Hemiptera: Aphididae). *Figshare*. https://doi.org/10.6084/m9.figshare.30229090 (2025).

Acknowledgements

We acknowledge the technical support of Nanjing Agricultural University Bioinformatics.

Author contributions

Zewen Liu and Jingting Wang. designed the study. Jingting Wang, Jiaqi Zhang, Yufan Li collected and reared *S. avenae*. Jingting Wang conducted the bioinformatics analyses. Jingting Wang, Huihui Zhang, Jiaqi Zhang, Yufan Li, Tianye Hu, Chuting Shi, Zewen Liu wrote and revised the draft manuscript. All authors read and approved the final version of 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/s41597-026-07147-w.

Correspondence and requests for materials should be addressed to Z.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.