scientific data

OPEN

DATA DESCRIPTOR

A chromosome-level genome assembly of two-spotted cricket, *Gryllus bimaculatus* (Orthoptera: Grylloidea)

Xuekuo Li[1,4], Yuanyuan Wang[2,4], Chengwei Lu[1], Tingting Lu[3], Hetan Chang[3], Guirong Wang[3], Shanchun Yan[1] ✉, Hongran Li[3] ✉ & Weikang Han[3] ✉

The crickets, *Gryllus bimaculatus* is a nutritionally valuable insect widely utilized in agricultural production as a high-quality protein source and an emerging resource species. However, the absence of a sequenced genome has limited exploration of its biological characteristics and population genetics. In this study, we generated a chromosome-level genome assembly for *G. bimaculatus* using PacBio HiFi and Hi-C technologies, yielding a 1,655.48 Mb genome with scaffold N50 of 111.16 Mb and contig N50 of 4.08 Mb. BUSCO analysis demonstrated 98.7% completeness, with repetitive sequences constituting 41.80% of the assembly. A total of 14,457 protein-coding genes were annotated. This high-quality genome assembly provides a valuable resource for investigating the genetic basis of local adaptation and developing effective management strategies.

Background & Summary

Crickets belong to the Orthoptera, and are typically characterized by the chirping of males during mating[1]. Besides, the males also displayed aggressive interactions among conspecific males[2]. This distinctive behavior underpinned the emergence of cricket fighting as a cultural practice in China. The crickets, such as *Gryllus bimaculatus*, *Gryllus campestris*, also served as the representative model for the behavioral, acoustic, and neurophysiological study[3–6].

Insects have been a human dietary tradition for thousands of years, and approximately 2 billion people around the world consume more than 2000 species of insects[7]. Climate change, extreme weather events, and unpredictable geopolitical conflicts can lead to food security issues[8]. Compared to traditional meat production, raising insects as food requires few resources. Therefore, the insects are considered as new sources of food for human to face the ever-increasing human population growth[9]. Crickets are also frequently reared on a large scale to satisfy the demands of livestock and pets[10].

High-quality reference genome provides a foundational resource in evolutionary biology, regeneration, neuropgysiology and behavior research[11]. Several cricket genomes have been sequenced and are available from NCBI, including *Acheta domesticus*[12], *Gryllus Longicercus*[13], *Teleogryllus oceanicus*[14] and *Laupala kohalensis*[15]. Though the genome of *Gryllus bimaculatus* has been sequenced and assembled[16], the quality of the genome is not enough to meet the experimental requirements, especially for the research on chemical perception. Here, we generate a high-quality chromosome-level genome assembly for *G. bimaculatus* using PacBio HiFi and Hi-C technologies. Comparative genomics revealed the expansion or contraction of the chemoreception-associated

[1]Key Laboratory of Sustainable Management of Forest Ecosystem, Ministry of Education, Northeast Forestry University, Harbin, 150040, China. [2]Key Laboratory of Biohazard Monitoring, Green Prevention and Control for Artificial Grassland, Ministry of Agriculture and Rural Affairs, Institute of Grassland Research of Chinese Academy of Agricultural Sciences, Hohhot, 010010, P. R. China. [3]Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Key Laboratory of Gene Editing Technologies (Hainan), Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, 518116, China. [4]These authors contributed equally: Xuekuo Li, Yuanyuan Wang. ✉e-mail: yanshanchun@126.com; lihongran@caas.cn; hanweikang@caas.cn

Libraries	Insert sizes (bp)	Clean data (Gb)	Sequencing coverage (×)
Illumina	350	29.61	17.89
PacBio HiFi	20 Kb	86.90	52.49
Hi-C	350	360.09	217.52
RNA	350	20.35	—

Table 1. Statistics of the sequencing data used for genome assembly.

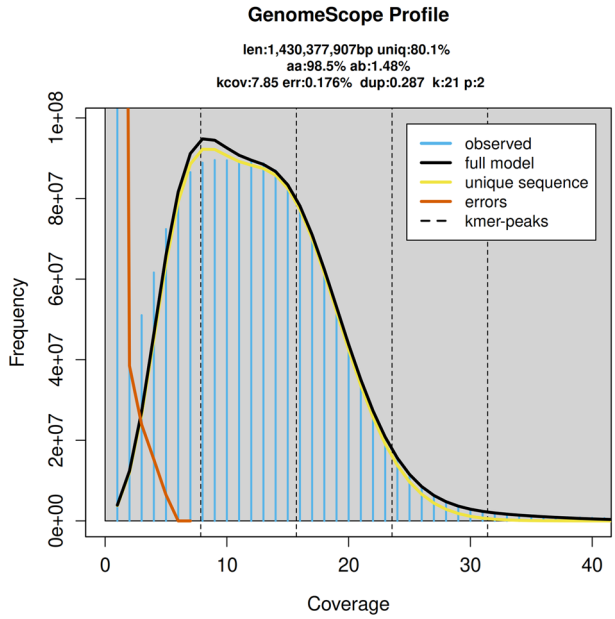

Fig. 1 Genome size, heterozygosity, and repeat content of the *Gryllus bimaculatus* genome were estimated using GenomeScope based on male *G. bimaculatus* Illumina short reads.

gene families, such as the ionotropic receptors. This high-quality genomic of *G. bimaculatus* provide a valuable resource not only for the fields of fundamental research, but also for the research of crickets as food.

Methods

Sample collection and sequencing. The *G. bimaculatus* samples were obtained from a local cricket climate chamber in Guangdong Province, China in November 24, 2023. The crickets were reared under the controlled conditions (27 ± 2 °C, 16 L: 8D and RH 50 ± 5%). The cricket population was maintained without introduction of outside specimens. The commercial poultry feed and cabbage leaves were used to fed the cricket. To eliminate the microbial contaminants, present on the surface of the cricket, the cricket sample underwent surface sterilization through a triple treatment process involving 75% alcohol and sodium hypochlorite solution. The PacBio HiFi library was constructed using the SMRTbell® Express Template Prep Kit 2.0 (Pacific Biosciences) targeting ~15 kb insert fragments. This involved fragmentation of quality-controlled genomic DNA using the Megaruptor system, followed by purification and concentration with AMPure® PB Beads. Subsequent steps included end repair, A-tailing, adapter ligation, and enzymatic digestion, with size selection performed using either the SageELF or BluePippin system. Sequencing of this library was conducted on the PacBio Sequel IIe platform. For BGISEQ-500 genomic sequencing (used for genome survey), DNA extracted via the CTAB method was used to construct libraries with the Agencourt AMPure XP kit, yielding ~350 bp inserts; these libraries were sequenced as 150 bp paired-end (PE) reads on the BGISEQ-500 platform. RNA sequencing libraries were prepared from total RNA isolated with TRIzol™ Reagent using the TruSeq RNA v2 kit and sequenced as 150 bp PE reads on the Illumina NovaSeq 6000 platform. The Hi-C library construction encompassed formaldehyde crosslinking, MboI restriction enzyme digestion, end repair, DNA circularization, and purification steps, followed by 150 bp PE sequencing on the Illumina NovaSeq 6000 platform. All library construction procedures were performed by Berry Genomics (Beijing, China). In total, 496.95 Gb of sequencing data were generated, comprising 86.90 Gb of PacBio HiFi reads (52.49x), 29.61 Gb of Illumina reads (17.89x), and 360.09 Gb of Hi-C data (217.52x) (Table 1).

Genome survey analysis. A k-mer analysis was performed using second-generation illumina sequencing data (350 bp insert size) to assess genome characteristics (size, heterozygosity, and repetitive sequence proportion) for guiding subsequent assembly strategies. Raw data underwent quality control with BBTools v38.82[17], deduplication using 'clumpify.sh', followed by filtering of low-quality sequences ($Q < 20$, length < 15 bp, terminal

Assembly	Total length (Mb)	Number scaffolds/ contigs (chromosomes)	Scaffold/contig N50 length (Mb)	GC (%)	BUSCO (n = 1,367) (%)				
					C	S	D	F	M
Hifiasm	1857.05	1089/1089	3.77/3.77	39.82	98.9	84.9	14.0	0.7	0.4
Purge_Dups	1664.89	688/688	4.11/4.11	39.97	98.7	95.8	2.9	0.7	0.6
3D-DNA	1664.96	397/1056 (15)	111.16/4.07	39.97	98.7	95.8	2.9	0.7	0.6
Final	1655.48	144/802 (15)	111.16/4.08	40.00	98.7	95.8	2.9	0.7	0.6

Table 2. Genome assembly statistics of *G. bimaculatus*. C: complete BUSCOs; S: Complete and single-copy BUSCOs; D: complete and duplicated BUSCOs; F: fragmented BUSCOs; M: missing BUSCOs.

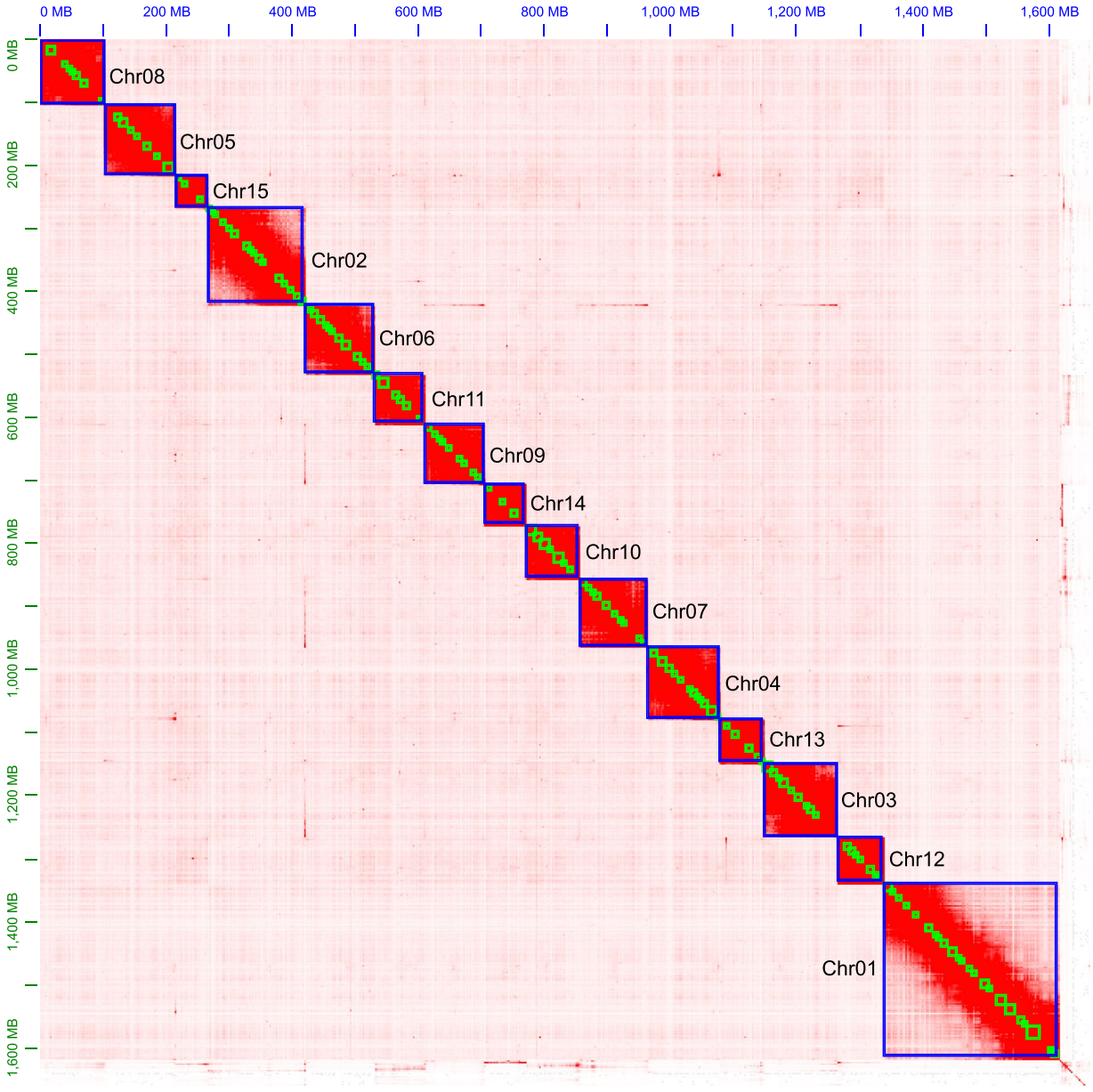

Fig. 2 Hi-C heatmap of *G. bimaculatus* genome containing 15 chromosomes. The genome information landscape map, from the outer ring, I: chromosome ideograms (Mb scale), II: Density of protein-coding genes; III: Density of G + C content; IV: Density of genome-wide single nucleotide polymorphism (SNP) sites; V: Density of long interspersed nuclear elements (LINEs); VI: Density of long terminal repeats (LTR); VII: Density of DNA transposons.

homopolymers ≥ 10 bp) and overlapping read correction using 'bbduk.sh' (parameters: 'qtrim = rl, trimq = 20, minlen = 15, ecco = t'). Frequency statistics were generated for k-mer = 21 ('khist.sh'), and the k-mer distribution was analyzed using GenomeScope v2.0[18] (parameters: '-k 21 -p 2 -m 1000'; maximum depth threshold: 10,000).

Species	Order	Family	Source
Tribolium castaneum	Coleoptera	Tenebrionidae	NCBI (GCA_031307605.1)
Eucriotettix oculatus	Orthoptera	Tetrigidae	NCBI (GCA_034510155.1)
Apis mellifera	Hymenoptera	Apidae	NCBI (GCA_003254395.2)
Bombyx mori	Lepidoptera	Bombycidae	NCBI (GCA_030269925.2)
Drosophila melanogaster	Diptera	Drosophilidae	NCBI (GCA_000001215.4)

Table 3. Species taxonomic information and accession code of all samples used in this study.

	G. bimaculatus
Structure annotation	
Number of protein-coding genes	14,457
Number of predicted protein sequences	19,058
Mean protein length (aa)	516.8
Mean gene length (bp)	25,982.0
Gene ratio	22.69%
Number of exons per gene	8.7
Mean exon length (bp)	289.2
Exon ratio	2.22%
Number of CDSs per gene	8.2
Mean CDS length (bp)	196.4
CDS ratio	1.42%
Number of introns per gene	7.6
Mean intron length (bp)	3,133.8
Intron ratio	20.47%
Function annotation	
Number of genes matching Uniprot records	11,878
Number of genes labeled as "Uncharacterized protein"	118
Number of genes labeled as "unknown function"	2,652
Number of genes with InterProScan annotations	11,240
Number of genes with GO items from InterProScan annotations	7,074
Number of genes with eggNOG annotations	11,944
Number of genes with GO items from eggNOG annotations	9,200
Number of genes with Enzyme Codes (EC) from eggNOG annotations	2,910
Number of genes with KEGG ko terms from eggNOG annotations	8,228
Number of genes with KEGG pathway terms from eggNOG annotations	5,074

Table 4. Genome annotation and functional annotation of *G. bimaculatus*.

This analysis predicted a genome size of 1,423.92 Mb, with a heterozygosity rate of 1.43% and a repetitive sequence proportion of 19.91%, while anomalous k-mer peaks indicated potential contamination risks (Fig. 1).

Genome assembly. The initial genome assembly was generated from high-quality HiFi reads using Hifiasm v0.16.1[19] with the parameter -l 2, and contigs with sequencing depth exceeding 10-fold coverage were retained to filter potential contaminant sequences. Subsequent removal of heterozygous sequences was performed using Purge_Dups v1.2.5[20] based on alignment results from Minimap2[21]. Following quality control of Hi-C data with Juicer, chromosomal scaffolding was accomplished using 3D-DNA combined with manual error correction in Juicebox[22], resulting in 15 pseudochromosomes with an anchoring rate of 97.66% and spanning 1,616.67 Mb. Potential contaminants were screened by aligning the assembly against the NCBI nt/UniVec databases using MMseqs2[23]. The final assembled genome size reached 1,655.48 Mb, with Scaffold/Contig N50 values of 111.16 Mb and 4.08 Mb, respectively, and a GC content of 40.00%. To assess the quality of the *de novo* assembly, Merqury v1.3[24] was used to calculate the QV score, which was estimated to be 36.69. BUSCO assessment (n = 1,367) indicated a completeness of 98.7%, including 95.8% single-copy orthologs, 2.9% duplicated genes, 0.7% fragmented genes, and 0.6% missing genes. Mapping rates for second-generation survey, RNA-seq, and HiFi data were 98.57%, 92.78%, and 99.95%, respectively, collectively demonstrating high continuity and completeness of the genome assembly (Table 2, Fig. 2).

Genome annotation. Repetitive sequences were annotated by first constructing a *de novo* repeat library for this species using RepeatModeler v2.0.4[25] with the additional '-LTRStruct' parameter to enable LTR structural detection based on repeat-specific architectures. Subsequently, the Dfam 3.5[26] and RepBase-20181021[27] databases were merged to generate the final repeat reference database. This composite database was then utilized by

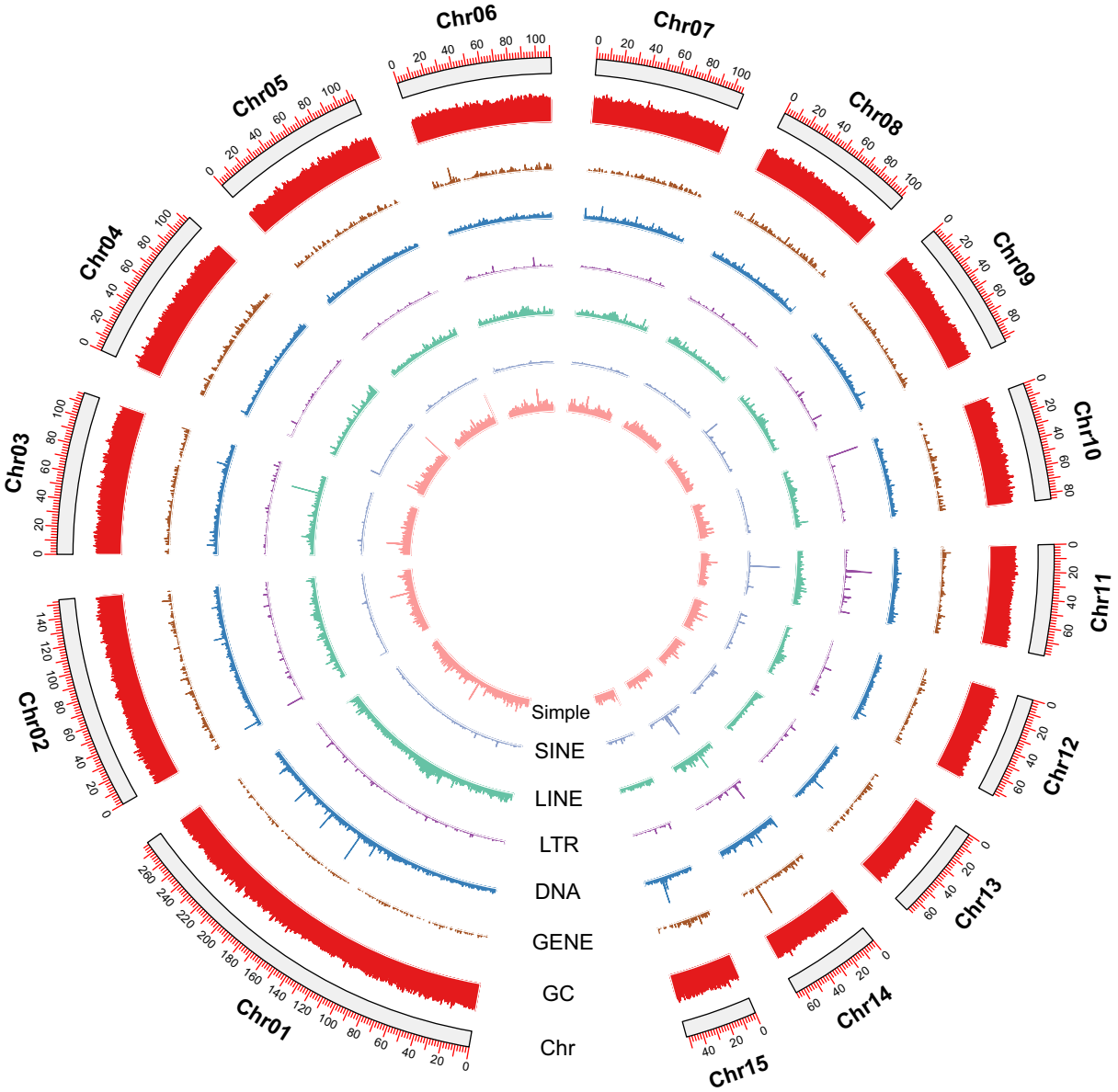

Fig. 3 The genomic features of *G. bimaculatus* are displayed in a circular layout. Moving inward from the outermost ring, the visualization depicts (1) chromosome length, (2) GC content, (3) gene density, and (4) various repetitive elements, including transposable elements (DNA, SINEs, LINEs, and LTRs), along with simple repeat sequences.

RepeatMasker v4.1.4[28] to identify repetitive elements across the genome. The analysis revealed 3,420,352 repetitive elements spanning 691,911,466 bp, accounting for 41.80% of the genome. The five predominant repeat classes were: Unknown (25.79%), LINE elements (5.57%), DNA transposons (4.14%), LTR elements (2.09%), and Simple repeats (3.30%).

Protein-coding gene structures were predicted using the MAKER v3.01.03[29] pipeline, which integrated multiple lines of evidence. Initially, ab initio predictions were generated by BRAKER v2.1.6[30], which automated the training of Augustus[31] and GeneMark-ES/ET[32] using RNA-seq data (aligned via HISAT2 v2.2.0[33] and conserved arthropod protein sequences from the OrthoDBv10[34] database. Concurrently, transcriptomic data underwent reference-based assembly with StringTie[35], while protein sequences from five closely related species (*Apis mellifera* (GCA_003254395.2), *Bombyx mori* (GCA_030269925.2), *Eucriotettix oculatus* (GCA_034510155.1), *Drosophila melanogaster* (GCA_000001215.4), and *Tribolium castaneum* (GCA_031307605.1)) were aligned to the genome using GeMoMa v1.8[36] with parameters 'c = 0.4' and 'p = 10' to refine predictions by leveraging homology and intron position information (Table 3). This integrated approach yielded 14,457 protein-coding genes characterized by an average gene length of 25,982.0 bp, containing 8.7 exons (mean 289.2 bp), 7.6 introns (mean 3,133.8 bp), and 8.2 CDS regions (mean 196.4 bp) (Table 4). BUSCO assessment (n = 1,367) confirmed a gene set completeness of 95.9%.

Characteristics	*Gryllus bimaculatus*
Genome assembly	
Genome Size (Mb)	1655.48
Number of scaffolds	144
Number of chromosomes	15
Scaffold N50 length (Mb)	111.16
Contig N50 length (Mb)	4.08
GC (%)	40.00
BUSCO completeness (%)	98.7%
Protein-coding genes	
Number	14457
Mean gene length (bp)	25982.0
BUSCO completeness (%)	95.9
Repetitive elements	
Size (Mb)	691.91 (41.80%)
DNA transposons (Mb)	68.44 (4.14%)
LINEs (Mb)	92.18 (5.57%)
LTRs (Mb)	34.59 (2.09%)
SINEs (Mb)	3.79 (0.22%)
Unclassified (Mb)	426.89 (25.79%)
Simple repeats (Mb)	54.62 (3.30%)
Other repeats (Mb)	11.4 (0.69%)
ncRNA	
Number of ncRNA	8389
tRNA	6834
rRNA	235
miRNA	82
snRNA	567
ribozyme	4
Histone3	66
R2_retro_el	535
Other ncRNA	66
Function annotation	
Number of genes matching Uniprot records	11878
Number of genes with InterProScan annotations	11240
Number of genes with eggNOG annotations	11944
Number of genes with COG Functional Categories combining InterProScan and eggNOG results	11426
Number of genes with GO items combining InterProScan and eggNOG results	10515
Number of genes with KEGG pathways items combining InterProScan and eggNOG results	5074

Table 5. Genome assembly and annotation statistics of *G. bimaculatus*.

Functional annotation of protein-coding genes was performed using an integrated multi-database strategy. Initial annotations were generated through Diamond v2.0.11.1[37] with sensitive parameters (–very-sensitive -e 1e-5) against the UniProtKB[38] database. Concurrently, protein domain analysis was conducted using InterProScan[39], which scanned Pfam[40], SMART[41], Superfamily[42], and CDD databases[43]. Functional predictions, including Gene Ontology (GO) terms and KEGG/Reactome pathways, were further enriched via eggNOG-mapper v2.1.5[44] leveraging the eggNOG v5.0[45] database. The integrated annotation pipeline successfully covered 11,878 genes (82.16% of the total), with 11,240 proteins assigned structural domains and 10,515 genes annotated with GO terms. Additionally, 5,074 genes were mapped to KEGG pathways, while 8,228 KEGG orthologs (KO), 2,910 Enzyme Commission (EC) numbers, and 11,426 Clusters of Orthologous Groups (COG) functional categories were identified (Table 4). Finally, a genomic annotation circle plot was generated using TBtools v1.098769[46] to visualize the comprehensive functional landscape (Fig. 3).

Non-coding RNA annotation was performed using a dual-strategy approach. First, Infernal v1.1.4[47] was employed to identify conserved non-coding RNAs (including rRNAs, miRNAs, and snRNAs) through alignment against the Rfam database[48]. Second, tRNAscan-SE v2.0.9[49] was utilized for tRNA prediction, with high-confidence results retained after stringent filtering based on the default covariance model (CM) cutoff scores. Collectively, this integrated workflow annotated 8,389 non-coding RNAs, comprising 6,834 tRNAs (accounting for 81.46% of the total), 567 snRNAs (including 505 spliceosomal RNAs such as U1, U2, U4, U5, U6, U11; 17 minor spliceosomal RNAs including U4atac, U6atac, U12; and 223 snoRNAs), 235 rRNAs, 82 miRNAs, 4 ribozymes, and 2 lncRNAs (Table 5).

Data Records

The sequencing data generated in this study are publicly available under the following National Center for Biotechnology Information (NCBI) SRA accession numbers: transcriptome reads (SRR35113091)[50], Hi-C data (SRR35113090)[51], Illumina short reads (SRR35113092)[52], and PacBio HiFi long reads (SRR35113093)[53]. The final genome assembly is avaiable under NCBI accession GCA_054507375.1[54]. Genome annotation data have been deposited in the figshare database (https://doi.org/10.6084/m9.figshare.30225808.v1)[55].

Technical Validation

Three validation methods were employed to assess the contiguity, accuracy and completeness of the genome assembly. Firstly, a total of 802 contigs were assembled, with a contig N50 size measuring 4.08 Mb. We then used the Hi-C data combined with the contig-level assembly to generate a chromosome-level assemble that spanned 1655.48 Mb, characterized by a scaffold N50 value of 111.16 Mb (Table 1). Secondly, the Hi-C heatmap unveiled a discernible and well-structured pattern of interaction contacts along the diagonals within and around the chromosome inversion region, resulted in the encapsulation of over 99.95% of the assembled sequences within these chromosomes (Fig. 2). Finally, we employed Benchmarking Universal Single-Copy Orthologs (BUSCOv5.1.2)[56], using homologous genes from insecta_odb10. The results indicated that 98.7% of BUSCO genes (insecta_odb10) were successfully detected within the genome assembly (Table 1). Among them, 95.8% were identified as single-copy genes and 2.9% were categorized as duplicated. These observations indirectly validated the precision and accuracy of the chromosome assembly.

To ensure the completeness and accuracy of the annotated gene set, the forecasted gene models underwent a comparison with multiple protein databases including NR, eggNOG, Trembl, and KEGG. The outcome demonstrated that 14,457 (95.9%) of the projected gene models exhibited notable homology with proteins found within at least one of these databases. Moreover, clean reads from three transcriptomes of larvae, pupa and adults were mapped onto the genome assembly, more than 92.78% of the RNA-Seq reads can be aligned to the coding regions of the reference genome.

Data availability

This dataset has not been previously published or presented, and no conflicts of interest are declared. The raw sequencing reads of *G. bimaculatus* have been deposited in the NCBI as described in Data Records. Additionally, final genome assembly and annotation data have been deposited in the figshare database (https://doi.org/10.6084/m9.figshare.30225808.v1)[54].

Code availability

All bioinformatic tools and software for data analysis in this study were used according to the manuals, and the versions and code/parameters of the software have been included in the Methods section. No custom code was used.

Received: 8 October 2025; Accepted: 28 February 2026;
Published online: 17 March 2026

References

1. Clark, H. Songs and the Physics of Sound Production. *Cricket Behav. Neurobiol*. 227–261 (1989).
2. Alexander, R. D. Aggressiveness, Territoriality, and Sexual Behavior in Field Crickets (Orthoptera: Gryllidae). *Behaviour* **17**, 130–223 (1961).
3. Elliott, C. J. H. & Koch, U. T. Sensory feedback stabilizing reliable stridulation in the field cricket *Glyllus campestris L. Anim. Behav*. **31** (1983).
4. Wendler, G., Dambach, M., Schmitz, B. & Scharstein, H. Analysis of the acoustic orientation behavior in crickets (*Gryllus campestris L.*). *Naturwissenschaften* **67**, 99–101 (1980).
5. Rillich, J., Schildberger, K. & Stevenson, P. A. Octopamine and occupancy: An aminergic mechanism for intruder-resident aggression in crickets. *Proc. R. Soc. B Biol. Sci.* **278**, 1873–1880 (2011).
6. Kostarakos, K. & Hedwig, B. Calling song recognition in female crickets: Temporal tuning of identified brain neurons matches behavior. *J. Neurosci.* **32**, 9601–9612 (2012).
7. Imathiu, S. Benefits and food safety concerns associated with consumption of edible insects. *NFS J.* **18**, 1–11 (2020).
8. Hazarika, A. K. & Kalita, U. Human consumption of insects. *Science* **379**, 140–141 (2023).
9. Deroy, O. Eat insects for fun, not to help the environment. *Nature* **521**, 395 (2015).
10. Van Huis, A. & Gasco, L. Insects as feed for livestock production. *Science* **379**, 138–139 (2023).
11. Wilson Horch, H., Mito, T., Popadić, A., Ohuchi, H. & Noji, S. *The cricket as a model organism: Development, regeneration, and behavior*. 1–376 (Springer Nature Press, 2017).
12. Dossey, A. T. *et al*. Genome and Genetic Engineering of the House Cricket (*Acheta domesticus*): A Resource for Sustainable Agriculture. *Biomolecules*. **13** (2023).
13. Szrajer, S., Gray, D. & Ylla, G. The genome assembly and annotation of the cricket *Gryllus longicercus*. *Sci. Data*. **11** (2024).
14. Pascoal, S. *et al*. Field cricket genome reveals the footprint of recent, abrupt adaptation in the wild. *Evol. Lett*. **4**, 19–33 (2020).
15. Blankers, T., Oh, K. P., Bombarely, A. & Shaw, K. L. The genomic architecture of a rapid Island radiation: Recombination rate variation, chromosome structure, and genome assembly of the hawaiian cricket Laupala. *Genetics*. **209**, 1329–1344 (2018).
16. Ylla, G. *et al*. Insights into the genomic evolution of insects from cricket genomes. *Commun. Biol*. **4** (2021).
17. Bushnell, B. BBtools. Available online: https://sourceforge.net/projects/bbmap/ (accessed on 1 October 2022) (2014).
18. 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** (2020).
19. 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).
20. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
21. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

22. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3**, 95–98 (2016).
23. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
24. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21** (2020).
25. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA.* **117**, 9451–9457 (2020).
26. Hubley, R. *et al*. The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**, D81–D89 (2016).
27. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA*. **6** (2015).
28. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. Available online: http://www.repeatmasker.org (accessed on 1 October 2022) (2013–2015).
29. Holt, C. & Yandell, M. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*. **12** (2011).
30. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*. **32**, 767–769 (2016).
31. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: A web server for gene finding in eukaryotes. *Nucleic Acids Res*. **32** (2004).
32. Bruna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics Bioinforma.* **3**, 1–11 (2021).
33. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods.* **12**, 357–360 (2015).
34. Kriventseva, E. V. *et al*. OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **47**, D807–D811 (2019).
35. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol*. **20** (2019).
36. Keilwagen, J., Hartung, F. & Grau, J. GeMoMa: Homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods Mol. Biol.* **1962**, 161–177 (2019).
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. Bateman, A. *et al*. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
39. Finn, R. D. *et al*. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.* **45**, D190–D199 (2017).
40. El-Gebali, S. *et al*. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
41. Ivica, L. & Peer, B. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* **46**, D493–D496 (2018).
42. Wilson, D. *et al*. SUPERFAMILY—Sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.* **37**, D380–D386 (2009).
43. Marchler-Bauer, A. *et al*. CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* **45**, D200–D203 (2017).
44. Huerta-Cepas, J. *et al*. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
45. 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).
46. Chen, C. *et al*. TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Mol. Plant.* **13**, 1194–1202 (2020).
47. Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA genes in genomic sequences. *Methods Mol. Biol.* **1962**, 1–14 (2019).
48. Griffiths-Jones, S. *et al*. Rfam: Annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*. **33** (2005).
49. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol*. **20** (2019).
50. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35113091 (2025).
51. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35113090 (2025).
52. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35113092 (2025).
53. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35113093 (2025).
54. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_054507375.1 (2026).
55. Han, W. A chromosome-level genome assembly of two-spotted cricket, *Gryllus bimaculatus* (Orthoptera: Grylloidea). *figshare. Dataset.* https://doi.org/10.6084/m9.figshare.30225808.v1 (2025).
56. Waterhouse, R. M. *et al*. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (32402373) and China Postdoctoral Science Foundation (2024M763625).

Author contributions

H.C. and W.H. conceived of the study. W.H., X.L., T.L., and C.L. prepared samples for genome sequencing and conducted bioinformatics analysis. The manuscript was written by X.L. and Y.W., finalized by G.W. with contributions from S.Y. and H.L. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.Y., H.L. or W.H.