

# Chromosome-scale genome assembly and annotation of the two-spotted cricket *Gryllus bimaculatus* (Orthoptera: Gryllidae)

Kosuke Kataoka <sup>1,2,\*</sup> Ryuto Sanno,<sup>3,4</sup> Tomasz Gaczorek <sup>3</sup> Upendra Raj Bhattarai,<sup>5,6</sup> Yuki Ito,<sup>4</sup> Shintaro Inoue <sup>7</sup> Kei Yura <sup>2,4,8</sup> Toru Asahi,<sup>2,4</sup> Guillem Ylla,<sup>3,\*</sup> Taro Mito,<sup>7,\*</sup> Cassandra G. Extavour <sup>5,9,10,\*</sup>

<sup>1</sup>Institute of Engineering, Tokyo University of Agriculture and Technology, 2-24-16 Naka-cho, Koganei-shi, Tokyo 184-8588, Japan

<sup>2</sup>Comprehensive Research Organization, Waseda University, 2-2 Wakamatsu-cho, Shinjuku-ku, Tokyo 162-8480, Japan

<sup>3</sup>Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, 7 Gronostajowa, Kraków 30-387, Poland

<sup>4</sup>Graduate School of Advanced Science and Engineering, Waseda University, 2-2 Wakamatsu-cho, Shinjuku-ku, Tokyo 162-8480, Japan

<sup>5</sup>Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, United States

<sup>6</sup>Harvard T.H. Chan School of Public Health, Harvard University, 677 Huntington Avenue, Boston, MA 02115, United States

<sup>7</sup>Bio-Innovation Research Center, Tokushima University, 2272-2 Ishii, Ishii, Ishii-cho, Myozai-gun, Tokushima 779-3233, Japan

<sup>8</sup>Graduate School of Humanities and Sciences, Ochanomizu University, 2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan

<sup>9</sup>Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, United States

<sup>10</sup>Howard Hughes Medical Institute, 4000 Jones Bridge Road, Chevy Chase, MD 20815, United States

\*Corresponding authors: Kosuke Kataoka, Institute of Engineering, Tokyo University of Agriculture and Technology, 2-24-16 Naka-cho, Koganei-shi, Tokyo 184-8588, Japan. Email: kataokak@go.tuat.ac.jp; Guillem Ylla, Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, 7 Gronostajowa, Kraków 30-387, Poland. Email: guillem.ylla@uj.edu.pl; Taro Mito, Bio-Innovation Research Center, Tokushima University, 2272-2 Ishii, Ishii, Ishii-cho, Myozai-gun, Tokushima 779-3233, Japan. Email: mito.taro@tokushima-u.ac.jp; Cassandra G. Extavour, Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, United States. Email: extavour@oeb.harvard.edu

The two-spotted cricket, *Gryllus bimaculatus*, is a key hemimetabolous model organism for developmental biology, neuroscience, and regeneration. The existing reference genome is, however, highly fragmented into 47,877 scaffolds, hampering chromosome-scale analyses for these fields. Here, we report a high-quality, chromosome-scale genome assembly for the white-eyed mutant strain of this cricket, generated using a combination of Nanopore and PacBio HiFi long reads, integrated with Hi-C data. The final 1.62-Gbp assembly achieves a scaffold N50 of 107.4 Mbp, a significant improvement in contiguity over the previous 6.3-Mbp N50. We anchored 94.45% of the assembly into 15 pseudomolecules, consistent with the known karyotype ( $n = 15$ ). The genome completeness (BUSCO v6.0.0 insecta\_odb12) reached 98.1%. We also updated the annotation, identifying 14,964 protein-coding genes. This gene set shows markedly improved completeness (BUSCO v6.0.0 insecta\_odb12: 95.7%) compared with the previous annotation (81.2%) and successfully recovers all 9 essential neuropeptide genes previously reported as missing from the draft assembly. This chromosome-scale genomic resource provides an essential foundation for comparative and functional genomics in *G. bimaculatus*.

**Keywords:** Orthoptera; Gryllidae; two-spotted cricket; white-eyed mutant strain; genome assembly

## Introduction

The two-spotted cricket, *Gryllus bimaculatus*, is a model organism for hemimetabolous insects (Horch et al. 2017) (Fig. 1). Unlike holometabolous insects such as fruit fly (*Drosophila melanogaster*) or red flour beetle (*Tribolium castaneum*), hemimetabolous insects undergo direct development, where nymphs hatch and grow through successive molts to become adults without larval or pupal stages. This ancestral developmental mode is crucially important for understanding insect evolution. Due to its ease of rearing, short generation time, and the applicability of efficient RNA interference (RNAi) (Miyawaki et al. 2004) and genome-editing techniques such as CRISPR/Cas9 (Matsuoka et al. 2025), *G. bimaculatus* has been established as a powerful experimental model in a wide range of fields, including evolutionary developmental biology (Donoughe and Extavour 2016), regeneration

biology (Nakamura et al. 2008), neuroscience (Matsumoto et al. 2018), and ethology (Abe et al. 2021; Kuriwada 2022). It is also gaining significant global attention as a novel, sustainable food source due to its high protein content and efficient rearing (Kataoka et al. 2020, 2022; Mito et al. 2022).

In the first report of a genomic resource for *G. bimaculatus*, the white-eyed mutant strain, which is standardly used in many functional studies, was sequenced (Ylla et al. 2021). The assembly (GenBank accession: GCA\_017312745.1) was a useful resource with a genome size of approximately 1.66 Gb and a scaffold N50 of 6.3 Mb, contributing to analyses of gene family evolution and DNA methylation. However, this previous assembly was fragmented into 47,877 scaffolds and was not assembled to chromosome scale. While chromosome-scale genomes have recently been reported for other related cricket species, such as *Gryllus assimilis* (Ito et al. 2025) and *Acheta domesticus* (Dossey et al. 2023), such a



**Fig. 1.** Adult male (left) and female (right) of the white-eyed mutant strain *G. bimaculatus*.

resource has remained unavailable for *G. bimaculatus*, which serves as the primary model for functional, developmental, and neurobiological studies. The lack of a chromosome-scale assembly for this specific species has been a significant limitation for large-scale comparative genomic analyses based on synteny (conservation of gene order), understanding the structural arrangement of transposons and repetitive sequences on chromosomes, and for quantitative trait locus mapping and the accurate identification of genome-editing off-target sites.

In this study, to fill the gap, we constructed a high-quality, chromosome-scale genome assembly using the same white-eyed mutant strain used in the previous study. By combining Nanopore and PacBio HiFi long-read sequencing methods, and Hi-C chromatin conformation capture technology, we anchored 94.45% of the entire genome into 15 pseudomolecules, corresponding to the *G. bimaculatus* karyotype ( $n = 14$  autosomes + X) (Yoshimura et al. 2006). This assembly achieves a contig N50 of 4.58 Mb and a scaffold N50 of 107.39 Mb, representing a significant improvement in contiguity. Furthermore, we report an updated gene annotation comprising 14,964 protein-coding genes. This chromosome-scale genome resource will strengthen the foundation for all genomic research using *G. bimaculatus* and accelerate new insights into the biology of hemimetabolous insects.

## Materials and methods

### Animals

A white-eyed mutant strain of *G. bimaculatus* (Mito and Noji 2008) was housed in plastic cages at  $30^{\circ}\text{C} \pm 1^{\circ}\text{C}$  and 30% to 40% relative humidity under a 10-h light and 14-h dark photoperiod. This strain has been widely used in genetic and developmental studies and was also used in the previous genome assembly, enabling direct comparison with earlier genomic resources (Miyawaki et al. 2004; Ylla et al. 2021). They were nourished with an artificial fish food (4971618–011312, Kyorin, Japan).

### Library preparation and sequencing

A single alive *G. bimaculatus* individual (Fig. 1) was used for genomic DNA extraction. Total genomic DNA was extracted from the head and hind legs of a male *G. bimaculatus* using NucleoBond HMW DNA (Macherey-Nagel, Germany) according to the manufacturer's instructions. This species exhibits an XX/XO sex determination system, in which males are hemizygous for the X chromosome; therefore, a male individual was selected to enable clear identification and chromosome-scale assembly of the X chromosome. The resulting genomic DNA was size selected using a Short Read Eliminator Kit (PacBio, CA, USA). DNA purity and concentrations were measured by spectrometry using

NanoPhotometer NP80-TOUCH (Implen, Germany) and fluorometry using Qubit 4 (Thermo Fisher Scientific, MA, USA).

For long-read sequencing, Oxford Nanopore Technologies (ONT) libraries were constructed using the Ligation Sequencing Kit V14 and sequenced on the PromethION 2 Solo platform (Oxford Nanopore Technologies, UK) with a Flow Cell R10.4.1. Base-calling was performed using Dorado v0.3.0 (model: dna\_r10.4.1\_e8.2\_400bps\_sup@v4.2.0). The resulting raw reads were adapter trimmed using Porechop\_ABI v0.5.0 (Bonenfant et al. 2023), and reads with a mean quality score below Q10 were removed using NanoFilt v2.8.0 (De Coster et al. 2018). Additionally, a SMRTbell library was prepared and sequenced on a PacBio Sequel IIe system.

For chromosome-scale scaffolding, 2 separate Hi-C libraries were prepared. The first library was prepared from the hind legs of a single male *G. bimaculatus* using the Dovetail Omni-C Kit (Dovetail Genomics, CA, USA) following the manufacturer's instructions. The second Hi-C library was generated from the thorax and legs of an adult male using the Proximo Hi-C (Animal) kit (KT2045) (Phase Genomics, Seattle, USA), following the manufacturer's instructions. Both libraries were sequenced on the Illumina NovaSeq 6000 platform, and the sequencing data were combined for downstream scaffolding analysis.

### Genome size estimation

Genome size estimation was performed using ONT reads. A total of 45.0 Gbp of ONT sequencing data was used for k-mer-based analysis. Reads were counted using Jellyfish v2.2.10 (Marçais and Kingsford 2011) with a k-mer size of 21. A k-mer frequency histogram was generated and used as input for GenomeScope v2.0 (Ranallo-Benavidez et al. 2020) to estimate genome size and sequence characteristics. GenomeScope was run with default parameters, and the resulting model was used to infer genome size based on the major k-mer peak.

### Genome de novo assembly

The initial draft genome was assembled by combining the filtered ONT long reads and the PacBio HiFi reads using Flye v2.9.5 (Kolmogorov et al. 2019). To correct errors in the resulting contigs, we retrieved publicly available Illumina short-read data [DDBJ Sequence Read Archive {DRA} accessions DRR272308–DRR272313], which were originally sequenced on an Illumina HiSeq 2000. These reads were downsampled to an approximate 93.9x coverage and used for polishing. The assembly underwent 2 rounds of error correction using POLCA v4.1.0 (Zimin and Salzberg 2020) with default settings.

Potential contamination in the assembly was removed using BlobToolKit v1.1.1 (Challis et al. 2020), which analyzes unexpected coverage, GC content, or similarity to bacterial and other contaminant sequences. Sequence coverage was determined by mapping Illumina reads with bwa v0.7.17-r1188. Similarity analysis was performed using BLASTn v2.13.0+ against NCBI nt database v5 (options: -task megablast culling\_limit 10 -evalue 1e-25 -outfmt "6 qseqid staxids bitscore std sscinames sskindoms title") to assign putative taxonomic origins of contigs as part of contamination screening and assembly quality control. Based on these analyses, no contigs could be confidently identified as apparent contaminants.

The resulting contigs were then corrected for misjoins, ordered, oriented, and anchored into a chromosome-scale assembly using Omni-C data with Juicer v1.9.9 (Durand et al. 2016) and 3D-DNA v180419 (Dudchenko et al. 2017). Candidate assembly was reviewed with Juicebox Assembly Tools v1.9.9 (Durand et al. 2016)

for quality control and interactive corrections. The contact map was visualized using Juicebox Assembly Tools v1.9.9. The completeness of the final genome assembly was assessed using BUSCO v6.0.0 (Tegenfeldt et al. 2025) against the insecta\_odb12 lineage datasets. In addition, k-mer-based assembly completeness and consensus accuracy were evaluated using Merqury v1.3 (Rhie et al. 2020).

### Mitochondrial genome assembly

The mitochondrial genome was assembled from PacBio HiFi reads using MitoHiFi v2 (Uliano-Silva et al. 2023), following the default workflow. Mitochondrial sequences were excluded from the nuclear genome assembly statistics but are provided as part of the released genome resources for transparency.

### Prediction of repeat regions

In de novo repeat prediction, RepeatModeler v2.0.6 (Flynn et al. 2020) was first used for de novo repeat identification. Because standard libraries are insufficient for effective masking in *Gryllus* genomes (Szrajer et al. 2024), this library was supplemented with the custom repeat library previously generated (Ylla et al. 2021). This custom library was constructed by integrating multiple complementary homology-based and de novo repeat identification approaches, enabling comprehensive annotation of diverse repeat classes ([https://github.com/guillemylla/Crickets\\_Genome\\_Annotation](https://github.com/guillemylla/Crickets_Genome_Annotation)). The combined repeat library was then used to identify and softmask repetitive elements in the *G. bimaculatus* genome using RepeatMasker v4.2.1 (Smit et al. 2015).

### Structural gene annotation

Structural annotation for protein-coding genes was performed on the softmasked genome using ab initio prediction and RNA-seq-based prediction. Both methods used publicly available RNA-seq data [Sequence Read Archive (SRA) accessions: SRR10619411, SRR10619415, SRR10619417, SRR10619418, SRR10619421, SRR10619423, SRR10619425, SRR10619429, SRR10619431, SRR10619432, SRR10619434, SRR10619437, SRR10619439, SRR10619440, SRR14026720–SRR14026726] as input.

To remove noisy RNA-seq reads potentially arising from erroneous transcription and splicing, de novo transcriptome assembly was first performed using Trinity v2.15.1 (Haas et al. 2013) to generate contigs. The original RNA-seq reads were then mapped back to these contigs using HISAT2 v2.2.1 (Kim et al. 2019) with default parameters, allowing filtration of reads that did not map correctly in the proper orientation as paired-end reads. After removing these noisy reads, the remaining reads were subsequently used for gene predictions.

The ab initio prediction was carried out using BRAKER v3.0.8 (Brůna et al. 2021; Gabriel et al. 2024; Hoff et al. 2016, 2019; Stanke et al. 2008, 2006), incorporating protein data from the OrthoDB 12 arthropods dataset (retrieved from [https://bioinf.uni-greifswald.de/bioinf/partitioned\\_odb12/](https://bioinf.uni-greifswald.de/bioinf/partitioned_odb12/)) and the mapping data of the filtered RNA-seq reads. This BRAKER prediction served as the foundation for our gene set. To complement this base set, StringTie2 v2.2.1 (Kovaka et al. 2019) was used for RNA-seq-based prediction. During this process, StringTie2-derived transcripts were compared against the BRAKER gene models, and only nonredundant transcript classes classified by GffCompare v0.12.6 (Pertea and Pertea 2020) as “u” (intergenic), “i” (intronic), or “y” (contains a reference transcript within its intron) were retained and added to the BRAKER-based gene set. Redundant or overlapping predictions were discarded. This procedure resulted in a final, nonredundant consensus gene set.

### Functional gene annotation

Gene functional annotation was conducted using eggNOG-mapper online (<http://eggno-mapper.embl.de/>) (Cantalapiedra et al. 2021) and BLASTp-based methods. For the BLASTp-based annotation, we used databases including *Homo sapiens*, *Mus musculus*, *Caenorhabditis elegans*, *D. melanogaster*, *T. castaneum*, and UniProt Swiss-Prot to identify the best hits for annotation (E-value <  $1.0 \times 10^{-10}$ ).

The completeness of the final predicted gene models was assessed using BUSCO v6.0.0 (Tegenfeldt et al. 2025) against the insecta\_odb12 lineage dataset. In addition, annotation quality was further evaluated using OMArk v0.4.1 (Nevers et al. 2025). For these analyses, the longest isoform for each gene was first extracted from the annotation file using AGAT v0.9.1 (Dainat et al. 2022).

### Phylogenomic analysis

We inferred a phylogenetic tree using single-copy orthologs identified by BUSCO (insecta\_odb12) from predicted protein-coding genes of *G. bimaculatus* and related insect species. BUSCO analyses were performed using BUSCO v6.0.0, and only single-copy BUSCO proteins shared across all taxa were retained. For each BUSCO ortholog, amino acid sequences were aligned using MAFFT v7.525 (Kato and Standley 2013) and subsequently trimmed to remove poorly aligned regions with trimAl v1.5.rev1 (Capella-Gutiérrez et al. 2009). The resulting alignments were concatenated into a supermatrix using AMAS v1.0 (Borowiec 2016). A maximum-likelihood phylogenetic tree was then inferred from the concatenated alignment using IQ-TREE v3.0.1 (Wong et al. 2025), with partitioned model selection and ultrafast bootstrap support.

### Validation of neuropeptide gene loci

To assess the completeness of our assembly regarding functionally important gene families, we specifically investigated the neuropeptide gene loci previously reported (Mochizuki et al. 2023). The neuropeptide cDNA sequences listed in the study were retrieved. These sequences were mapped against our final chromosome-scale genome assembly using Exonerate v2.4.0 (Slater and Birney 2005) with the est2genome model to accurately determine exon–intron boundaries. All resulting alignments were manually inspected to validate the gene structures.

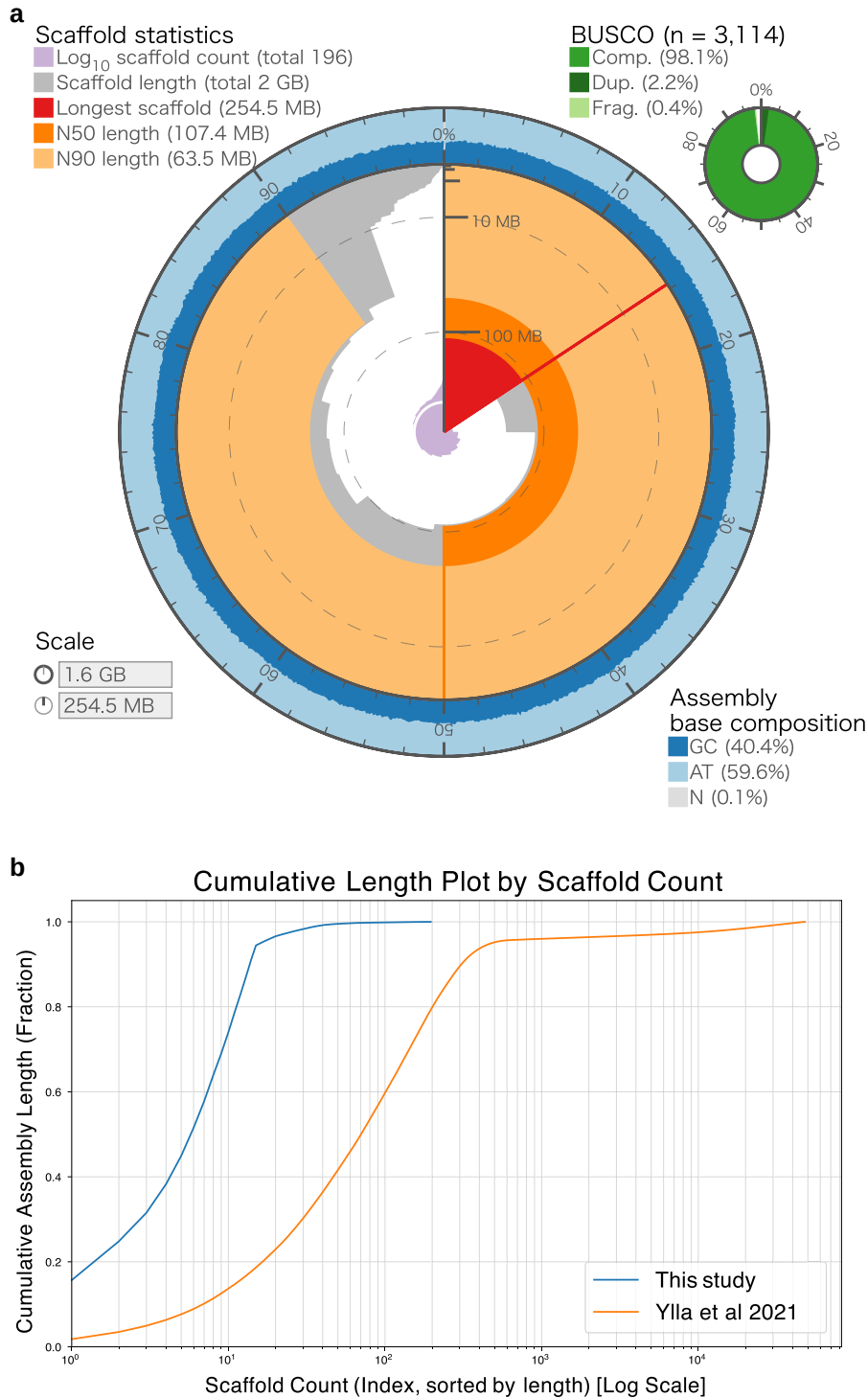
## Results and discussion

### Genome sequencing

To construct the genome assembly, we generated 3 types of sequencing data (Table 1). First, we obtained 45.00 Gbp of ONT sequencing data; the average and N50 read lengths were 13.13 Kbp and 24.23 Kbp, respectively. Second, we generated 13.44 Gbp of PacBio HiFi data, with an average read length of 13.61 Kbp and an N50 of 14.17 Kbp. Finally, for chromosome-scale scaffolding, 243.22 Gbp of Hi-C raw data was generated from the Illumina platform.

**Table 1.** Statistics for the DNA-seq data of the *G. bimaculatus* genome.

| Platform       | Raw data (bp)   | Average read length (bp) | N50 Read length (bp) | coverage (X) |
|----------------|-----------------|--------------------------|----------------------|--------------|
| ONT            | 45,004,338,081  | 13,129.8                 | 24,234               | 27.73        |
| PacBio HiFi    | 13,437,781,216  | 13,611.4                 | 14,170               | 8.28         |
| Hi-C libraries | 243,215,968,945 | 150                      | 150                  | 149.87       |



**Fig. 2.** Assembly statistics and contiguity of the *G. bimaculatus* (white-eyed strain) chromosome-scale genome. a) Snail plot visualizing the key statistics of the final assembly. The cumulative assembly length (1.62 Gbp) is plotted in light orange in the middle, with the longest scaffold (254.5 Mbp) shown in red in the center. The the N50 length (107.4 Mbp) is indicated by an orange arc. The inner purple spiral plots the cumulative number of scaffolds (196 total) on a log scale, with white scale lines drawn at successive orders of magnitude from 10 scaffolds onward. The circumferential axis indicates the base composition of the assembly (GC, 40.4%; AT, 59.6%; N, 0.1%). The donut chart (top right) displays the BUSCO completeness (insecta\_odb12), showing 98.1% total complete (“Comp.”) genes (light green), which includes 96.0% single-copy and 2.2% duplicated (“Dup.”) genes (dark green). An additional 0.4% were fragmented (“Frag.”) genes (pale green). An interactive version is available at <https://kataokaklab.github.io/snailplot-assembly-stats/>. b) Cumulative length plot comparing the contiguity of this assembly (blue line on the left) with the previous draft assembly (Ylla et al. 2021) (orange line on the right). The y axis shows the percentage of the total genome length covered, and the x axis (log scale) shows the number of scaffolds.

### Genome de novo assembly statistics

Genome size estimation based on Nanopore long-read data using a k-mer-based approach yielded an estimated genome size of

approximately 1.64 Gbp (Supplementary Fig. 1). This estimate is highly consistent with the final assembled genome size (1.62 Gbp) and with previous estimates reported for *G. bimaculatus* (Ylla et al. 2021).



**Table 2.** Summary statistics for the chromosome-scale assembly of *G. bimaculatus*.

| Features               | Statistic     |
|------------------------|---------------|
| Number of chromosomes  | 15            |
| Number of scaffolds    | 196           |
| Scaffold N50 (bp)      | 107,386,002   |
| GC content (%)         | 40.36         |
| Max scaffold size (bp) | 254,522,992   |
| Total N (bp)           | 1,440,159     |
| Gaps (%)               | 0.089         |
| Total size (bp)        | 1,622,898,960 |

**Table 3.** BUSCO completeness assessment of the *G. bimaculatus* genome assembly.

| BUSCO v6.0.0 (insecta)      | This study | Ylla et al. 2021 |
|-----------------------------|------------|------------------|
| Complete BUSCOs             | 98.1%      | 96.0%            |
| Single-copy complete BUSCOs | 96.0%      | 94.3%            |
| Duplicated complete BUSCOs  | 2.2%       | 1.7%             |
| Fragmented BUSCOs           | 0.4%       | 1.4%             |
| Missing BUSCOs              | 1.5%       | 2.6%             |

**Table 4.** Statistics of the assembled pseudochromosomes in *G. bimaculatus*.

| Chromosome | Length (bp)   | Proportion in genome |
|------------|---------------|----------------------|
| chrX       | 254,523,492   | 15.68%               |
| chr1       | 148,286,872   | 9.14%                |
| chr2       | 109,203,633   | 6.73%                |
| chr3       | 109,044,855   | 6.72%                |
| chr4       | 108,018,790   | 6.66%                |
| chr5       | 107,387,002   | 6.62%                |
| chr6       | 100,738,338   | 6.21%                |
| chr7       | 98,895,268    | 6.09%                |
| chr8       | 83,461,245    | 5.14%                |
| chr9       | 81,413,916    | 5.02%                |
| chr10      | 78,899,412    | 4.86%                |
| chr11      | 69,666,237    | 4.29%                |
| chr12      | 66,393,650    | 4.09%                |
| chr13      | 63,483,624    | 3.91%                |
| chr14      | 53,360,289    | 3.29%                |
| Total      | 1,532,776,623 | 94.45%               |

The hybrid assembly strategy combining ONT and HiFi reads, followed by 2 rounds of Illumina-based polishing, yielded a 1.63-Gbp draft genome. This polished assembly comprised 3,789 contigs and achieved a contig N50 of 4.58 Mbp (Supplementary Table 1).

Following Hi-C-based scaffolding, the final assembly resulted in a genome of 1.62 Gbp in size, visualized by a snail plot (Fig. 2a). This assembly consists of 196 scaffolds with a scaffold N50 length of 107 Mbp (Table 2). This represents a substantial improvement in contiguity compared to the previous *G. bimaculatus* assembly (Ylla et al. 2021), which had a scaffold N50 of 6.3 Mbp and was fragmented into 47,877 scaffolds (Fig. 2b). To assess genomic completeness, we performed a BUSCO v6.0.0 analysis using the insecta\_odb12 dataset. Our assembly achieved a completeness score of 98.1% (C:98.1%[S:96.0%, D:2.2%], F:0.4%, M:1.5%), demonstrating a higher level of completeness compared to the 96.0% (C:96.0%[S:94.3%, D:1.7%], F:1.4%, M:2.6%) of the previous genome (Table 3). A k-mer-based evaluation using Merqury showed that 85.5% of the reliable k-mers derived from the sequencing reads were represented in the genome assembly, with an estimated consensus quality value (QV) of 33.6, corresponding to a base-level accuracy of approximately 99.96%.

A total of 15 pseudochromosomes were identified based on Hi-C-guided scaffolding using 3D-DNA and accounted

**Table 5.** Classification of repetitive sequences of the *G. bimaculatus* genome assembly.

| Type                       | Length (bp) | Proportion in genome (%) |
|----------------------------|-------------|--------------------------|
| Retroelements              | 226,335,103 | 13.95                    |
| SINEs                      | 20,383,703  | 1.26                     |
| Penelope                   | 1,718,510   | 0.11                     |
| LINEs                      | 156,279,698 | 9.63                     |
| LTR elements               | 47,953,192  | 2.95                     |
| DNA transposons            | 170,814,501 | 10.53                    |
| Unclassified               | 294,555,500 | 18.15                    |
| Total interspersed repeats | 691,705,104 | 42.62                    |
| Rolling-circles            | 20,709,593  | 1.28                     |
| Other repeats              | 113,971,824 | 7.02                     |
| Total                      | 823,137,644 | 50.72                    |

**Table 6.** Summary statistics for gene prediction in *G. bimaculatus*.

| Features                       |        |
|--------------------------------|--------|
| Number of protein-coding genes | 14,964 |
| Mean exons per mRNA            | 7.0    |
| Mean gene length (bp)          | 29,659 |
| Mean exon length (bp)          | 2,32   |
| Mean intron length (bp)        | 4,696  |

for 94.45% of the total genome length (Table 4). This number ( $n = 15$ ) is in agreement with the established karyotype for *G. bimaculatus* ( $n = 14$  autosomes + X), which was previously determined by cytogenetic analysis (Yoshimura et al. 2006), and is also consistent with that of the recently reported species *G. assimilis* (Ito et al. 2025), which is closely related to *G. bimaculatus*. The X chromosome was identified as the longest, accounting for 15.68% of the genome, which is in accordance with the karyotype of this species (Yoshimura et al. 2006). This was further validated by the observation that the X chromosome displayed half of the read coverage compared to the autosomal chromosomes, calculated using a genomic short-read library (DRR272308) from a single hemizygous male (Supplementary Fig. 2).

Repetitive sequence annotation revealed that 823.1 Mbp of the genome, corresponding to 50.72% of the assembled sequence, is composed of repetitive elements (Table 5). This includes a substantial contribution from both Class I retroelements (13.95%) and Class II DNA transposons (10.53%), as well as a notable fraction of unclassified repeats (18.15%). The proportion of annotated repetitive sequences is markedly higher than that reported in the previous *G. bimaculatus* assembly, in which only 33.69% of the genome was annotated as repetitive (Ylla et al. 2021), likely reflecting improved assembly contiguity in the present study.

## Gene annotation

The final consensus gene set, derived from merging BRAKER ab initio predictions and StringTie2 RNA-seq-based predictions, comprised 14,964 protein-coding genes (Table 6). This total is less than the 17,871 genes reported previously (Ylla et al. 2021), likely because the improved scaffold contiguity of our assembly allows for the correct assembly of genes previously split across multiple scaffolds. Of the 14,964 predicted genes, functional annotation was assigned using eggNOG-mapper and BLASTp (E-value  $< 1.0 \times 10^{-10}$ ) against several model organisms' annotation datasets. eggNOG-mapper annotated 71.96% of the genes. The BLASTp searches (E-value  $< 1.0 \times 10^{-10}$ ) against databases including *H. sapiens*, *M. musculus*, *C. elegans*, *D. melanogaster*, *T. castaneum*, and UniProt Swiss-Prot yielded hit rates ranging from 44.84% to 75.76% (Table 7).

**Table 7.** Structural statistics of the *G. bimaculatus* gene set.

| Annotation sources                    |        |
|---------------------------------------|--------|
| <i>H. sapiens</i> (GRCh38)            | 44.84% |
| <i>M. musculus</i> (GRCm39)           | 71.89% |
| <i>C. elegans</i> (WBcel235)          | 55.18% |
| <i>D. melanogaster</i> (BDGP6.32)     | 67.16% |
| <i>T. castaneum</i> (icTriCast1.1)    | 75.76% |
| UniProt/Swiss-Prot (release: 2020_06) | 66.74% |
| eggNOG-mapper                         | 71.96% |

**Table 8.** Functional annotation summary for the *G. bimaculatus* gene set.

| BUSCO v6.0.0 (insecta)      | This study | Ylla et al. 2021 |
|-----------------------------|------------|------------------|
| Complete BUSCOs             | 95.7%      | 81.2%            |
| Single-copy complete BUSCOs | 93.6%      | 79.9%            |
| Duplicated complete BUSCOs  | 2.0%       | 1.3%             |
| Fragmented BUSCOs           | 1.5%       | 8.9%             |
| Missing BUSCOs              | 2.9%       | 10.0%            |

The completeness of this annotation set was validated using BUSCO v6.0.0. The analysis identified 95.7% of the expected complete insecta BUSCOs (C:95.7%[S:93.6%, D:2.0%], F:1.5%, M:2.9%) (Table 8). OMArk analysis indicated a completeness of 92.67% and a consistency score of 71.54% (Supplementary Table 2). Together, these independent metrics support the overall quality of the predicted gene models.

We next inferred a maximum-likelihood phylogeny using single-copy BUSCO orthologs shared across *G. bimaculatus* and closely related orthopteran species (Supplementary Fig. 3). The resulting topology places *G. bimaculatus* within Gryllidae and in the expected relationship to other *Gryllus* and cricket lineages, providing an additional consistency check and a comparative framework for future evolutionary analyses using this updated genome resource.

Recovery of missing neuropeptide genes

A recent comprehensive study (Mochizuki et al. 2023) highlighted significant gaps in the previous draft genome assembly (Ylla et al. 2021). They reported that several crucial neuropeptides genes [e.g. ACP [adipokinetic hormone/corazonin-related peptide], allatotropin, and kinin] were missing from the draft genome and could only be identified within de novo transcriptome assemblies, suggesting these loci were absent from the previous reference.

We asked if our new chromosome-scale assembly was more complete in this regard, by mapping the cDNA sequences of these previously missing neuropeptides. We successfully located all 9 of these genes encoding neuropeptides [i.e. ACP, allatostatin CC {Ast CC}, allatotropin, CCHamide-1, CCHamide-2, CRF/DH {corticotropin-releasing factor-like diuretic hormone}, kinin [leucokinin], neuropeptide F1a [NPF1a], neuropeptide F1b {NPF1b}], which are now correctly anchored onto our pseudomolecules. For example, the ACP gene, previously missing, was successfully mapped to chromosome X, where it spans 11,668 bp and is composed of 3 exons, revealing its complete exon-intron structure (Supplementary Fig. 4). This demonstrates that our assembly not only improves contiguity to the chromosome scale but also recovers functionally critical genes that were absent in the previous reference, providing a more complete and reliable resource for functional genomics in *G. bimaculatus*.

Conclusions

We have generated a high-quality, chromosome-scale genome assembly and updated gene annotation for the key hemimetabolous model organism *G. bimaculatus*. This assembly represents a substantial upgrade to the previous draft sequence (Ylla et al. 2021), increasing the scaffold N50 from 6.3 to 107.4 Mbp and anchoring 94.45% of the sequence into 15 pseudomolecules, consistent with the known karyotype (Yoshimura et al. 2006).

Crucially, our assembly resolves significant gaps present in the previous version, evidenced by the recovery of 9 essential neuropeptide genes previously reported as missing (Mochizuki et al. 2023). This improved completeness is further supported by superior BUSCO scores for both the genome (98.1% vs 96.0%) and the gene set (95.7% vs 81.2%). This highly contiguous and complete genome sequence provides an essential new foundation for the *G. bimaculatus* research community, facilitating advanced genetic and genomic analyses, such as synteny comparisons, QTL mapping, and the precise design of genome-editing experiments.

Data availability

The scripts used for the analyses in this study are available in GitHub ([https://github.com/Kataoka-K-Lab/Gryllus\\_bimaculatus\\_genome\\_gbim\\_v2.2](https://github.com/Kataoka-K-Lab/Gryllus_bimaculatus_genome_gbim_v2.2)). All bioinformatics tools used in this study followed their respective manuals and protocols. The software versions, codes, and parameters are provided in the Materials and Methods section. Unless otherwise specified, default parameters were used.

The assembled genome and annotation datasets are available in Figshare (<https://doi.org/10.6084/m9.figshare.30472754.v1>) (Kataoka 2025).

The genomic WGS sequencing data were deposited in the NCBI SRA database under the BioProject PRJNA1347939. All data and assembly statistics reported in this study correspond to the version before contamination screening and removal performed by NCBI.

Supplemental material available at G3 online.

Funding

This study was supported by the Cabinet Office, Government of Japan, Cross-ministerial Moonshot Agriculture, Forestry and Fisheries Research and Development Program, “Technologies for Smart Bio-industry and Agriculture” (BRAIN) (JPJ009237 to K.K., S.I., T.A., K.Y., and T.M.), the Strategic Programme Excellence Initiative at the Jagiellonian University—BioS PRA (to T.G. and G.Y.), Foundation for Polish Science (FNP) START 2025 (to T.G.), and the National Science Foundation Award [IOS-2220747] (to C.G.E.). C.G.E. is an investigator of the Howard Hughes Medical Institute.

Conflicts of interest

None declared.

Literature cited

Abe T, Tada C, Nagayama T. 2021. Winner and loser effects of juvenile cricket *Gryllus bimaculatus*. *J Ethol.* 39:47–54. <https://doi.org/10.1007/s10164-020-00671-1>.  
Bonenfant Q, Noé L, Touzet H. 2023. Porechop\_ABI: discovering unknown adapters in Oxford Nanopore Technology sequencing

- reads for downstream trimming. *Bioinform Adv.* 3:vbac085. <https://doi.org/10.1093/bioadv/vbac085>.
- Borowiec ML. 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ.* 4:e1660. <https://doi.org/10.7717/peerj.1660>.
- Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform.* 3:lqaa108. <https://doi.org/10.1093/nargab/lqaa108>.
- Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol.* 38:5825–5829. <https://doi.org/10.1093/molbev/msab293>.
- Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. Trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England).* 25:1972–1973. <https://doi.org/10.1093/bioinformatics/btp348>.
- Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. 2020. BlobToolKit—interactive quality assessment of genome assemblies. *G3 (Bethesda).* 10:1361–1374. <https://doi.org/10.1534/g3.119.400908>.
- Dainat J et al. 2022. NBISweden/AGAT: AGAT-v0.9.1. Zenodo. <https://doi.org/10.5281/ZENODO.6488306>.
- De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics (Oxford, England).* 34:2666–2669. <https://doi.org/10.1093/bioinformatics/bty149>.
- Donoughe S, Extavour CG. 2016. Embryonic development of the cricket *Gryllus bimaculatus*. *Dev Biol.* 411:140–156. <https://doi.org/10.1016/j.ydbio.2015.04.009>.
- Dossey AT et al. 2023. Genome and genetic engineering of the house cricket (*Acheta domesticus*): a resource for sustainable agriculture. *Biomolecules.* 13:589. <https://doi.org/10.3390/biom13040589>.
- Dudchenko O et al. 2017. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science.* 356:92–95. <https://doi.org/10.1126/science.aal3327>.
- Durand NC et al. 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* 3:95–98. <https://doi.org/10.1016/j.cels.2016.07.002>.
- Flynn JM et al. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A.* 117:9451–9457. <https://doi.org/10.1073/pnas.1921046117>.
- Gabriel L et al. 2024. BRAKER3: fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* 34:769–777. <https://doi.org/10.1101/gr.278090.123>.
- Haas BJ et al. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc.* 8:1494–1512. <https://doi.org/10.1038/nprot.2013.084>.
- Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics.* 32:767–769. <https://doi.org/10.1093/bioinformatics/btv661>.
- Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-genome annotation with BRAKER. *Methods in Molecular Biology (Clifton, N.J.).* 1962:65–95. [https://doi.org/10.1007/978-1-4939-9173-0\\_5](https://doi.org/10.1007/978-1-4939-9173-0_5).
- Horch HW, Mito T, Popadic A, Ohuchi H, Noji S. 2017. The cricket as a model organism: development, regeneration, and behavior. Springer.
- Ito Y et al. 2025. Chromosome-scale whole genome assembly and annotation of the Jamaican field cricket *Gryllus assimilis*. *Sci Data.* 12: 826. <https://doi.org/10.1038/s41597-025-05197-0>.
- Kataoka K. 2025. Chromosome-scale genome assembly and annotation of the two-spotted cricket *Gryllus bimaculatus* (Orthoptera: Gryllidae) [Data set]. figshare. <https://doi.org/10.6084/M9.FIGSHARE.30472754.V1>.
- Kataoka K, Togawa Y, Sanno R, Asahi T, Yura K. 2022. Dissecting cricket genomes for the advancement of entomology and entomophagy. *Biophys Rev.* 14:75–97. <https://doi.org/10.1007/s12551-021-00924-4>.
- Kataoka K et al. 2020. The draft genome dataset of the Asian cricket *Teleogryllus occipitalis* for molecular research toward entomophagy. *Front Genet.* 11:470. <https://doi.org/10.3389/fgene.2020.00470>.
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30:772–780. <https://doi.org/10.1093/molbev/mst010>.
- Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol.* 37:907–915. <https://doi.org/10.1038/s41587-019-0201-4>.
- Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol.* 37: 540–546. <https://doi.org/10.1038/s41587-019-0072-8>.
- Kovaka S et al. 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* 20:278. <https://doi.org/10.1186/s13059-019-1910-1>.
- Kuriwada T. 2022. Encounter with heavier females changes courtship and fighting efforts of male field crickets *Gryllus bimaculatus* (Orthoptera: Gryllidae). *J Ethol.* 40:145–151. <https://doi.org/10.1007/s10164-021-00742-x>.
- Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics (Oxford, England).* 27:764–770. <https://doi.org/10.1093/bioinformatics/btr011>.
- Matsumoto Y, Matsumoto CS, Mizunami M. 2018. Signaling pathways for long-term memory formation in the cricket. *Front Psychol.* 9:1014. <https://doi.org/10.3389/fpsyg.2018.01014>.
- Matsuoka Y et al. 2025. Establishment of CRISPR/Cas9-based knock-in in a hemimetabolous insect: targeted gene tagging in the cricket *Gryllus bimaculatus*. *Development.* 152:dev199746. <https://doi.org/10.1242/dev.199746>.
- Mito T, Noji S. 2008. The two-spotted cricket *Gryllus bimaculatus*: an emerging model for developmental and regeneration studies. *Cold Spring Harb Protoc.* 2008:db.emo110. <https://doi.org/10.1101/pdb.emo110>.
- Mito T et al. 2022. Cricket: the third domesticated insect. *Curr Top Dev Biol.* 147:291–306. <https://doi.org/10.1016/bs.ctdb.2022.02.003>.
- Miyawaki K et al. 2004. Involvement of Wingless/Armadillo signaling in the posterior sequential segmentation in the cricket, *Gryllus bimaculatus* (Orthoptera), as revealed by RNAi analysis. *Mech Dev.* 121:119–130. <https://doi.org/10.1016/j.mod.2004.01.002>.
- Mochizuki T et al. 2023. Best practices for comprehensive annotation of neuropeptides of *Gryllus bimaculatus*. *Insects.* 14:121. <https://doi.org/10.3390/insects14020121>.
- Nakamura T, Mito T, Bando T, Ohuchi H, Noji S. 2008. Dissecting insect leg regeneration through RNA interference: dissecting insect leg regeneration through RNA interference. *Cell Mol Life Sci.* 65: 64–72. <https://doi.org/10.1007/s00018-007-7432-0>.
- Nevers Y et al. 2025. Quality assessment of gene repertoire annotations with OMArk. *Nat Biotechnol.* 43:124–133. <https://doi.org/10.1038/s41587-024-02147-w>.
- Pertea G, Pertea M. 2020. GFF utilities: GffRead and GffCompare. *F1000Res.* 9:304. <https://doi.org/10.12688/f1000research.23297.1>.

- Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun.* 11:1432. <https://doi.org/10.1038/s41467-020-14998-3>.
- Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* 21:245. <https://doi.org/10.1186/s13059-020-02134-9>.
- Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics.* 6:31. <https://doi.org/10.1186/1471-2105-6-31>.
- Smit AFA, Hubley R, Green P. 2015. *RepeatMasker* (Open-4.0) [Computer software].
- Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics.* 24:637–644. <https://doi.org/10.1093/bioinformatics/btn013>.
- Stanke M, Schöffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics.* 7: 62. <https://doi.org/10.1186/1471-2105-7-62>.
- Szrajer S, Gray D, Ylla G. 2024. The genome assembly and annotation of the cricket *Gryllus longicercus*. *Sci Data.* 11:708. <https://doi.org/10.1038/s41597-024-03554-z>.
- Tegenfeldt F et al. 2025. OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. *Nucleic Acids Res.* 53:D516–D522. <https://doi.org/10.1093/nar/gkae987>.
- Uliano-Silva M et al. 2023. Mitohifi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. *BMC Bioinformatics.* 24:288. <https://doi.org/10.1186/s12859-023-05385-y>.
- Wong TKF et al. 2025. IQ-TREE 3: Phylogenomic Inference Software using Complex Evolutionary Models [preprint]. *EcoEvoRxiv.* <https://ecoevorxiv.org/repository/view/8916/>
- Ylla G et al. 2021. Insights into the genomic evolution of insects from cricket genomes. *Commun Biol.* 4:733. <https://doi.org/10.1038/s42003-021-02197-9>.
- Yoshimura A, Nakata A, Mito T, Noji S. 2006. The characteristics of karyotype and telomeric satellite DNA sequences in the cricket, *Gryllus bimaculatus* (Orthoptera, Gryllidae). *Cytogenet Genome Res.* 112:329–336. <https://doi.org/10.1159/000089889>.
- Zimin AV, Salzberg SL. 2020. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS Comput Biol.* 16:e1007981. <https://doi.org/10.1371/journal.pcbi.1007981>.

Editor: A. Sethuraman