scientific data

OPEN

DATA DESCRIPTOR

Chromosome-scale whole genome assembly and annotation of the Jamaican field cricket *Gryllus assimilis*

Yuki Ito[1,7], Ryuto Sanno[1,7], Seiya Ashikari[2], Kei Yura[1,3,4], Toru Asahi[1,4], Guillem Ylla[5] & Kosuke Kataoka[4,6] ✉

***Gryllus assimilis*, commonly known as Jamaican field cricket, is an edible insect with significant economic value in sustainable food production. Despite its importance, a high-quality reference genome of *G. assimilis* has not yet been published. Here, we report a chromosome-level reference genome of *G. assimilis* based on Oxford Nanopore Technologies (ONT) sequencing, Illumina sequencing, and Hi-C technologies. The assembled genome has a total length of 1.60 Gbp with a scaffold N50 of 102 Mbp, and 96.80% of the nucleotides was assigned to 15 chromosome-scale scaffolds. The assembly completeness was validated using BUSCO, achieving 99.5% completeness against the arthropoda database. We predicted 27,645 protein-coding genes, and 825 Mb repetitive elements were annotated in the reference genome. This reference genome of *G. assimilis* can provide a basis for the subsequent development of genomic resources, offering insights for future functional genomic studies, comparative genomics, and DNA-informed breeding of this species.**

Background & Summary

Gryllus assimilis, belonging to the Gryllidae family, is widely distributed across the West Indies, Southern United States, Mexico, and South America[1]. This species generally inhabits lawns, weedy fields, roadsides, and other open areas[2].

Globally, there is growing interest in integrating cricket-based ingredients into food products to combat food and nutrition insecurity[3]. Because of the high content of lipids, proteins, and carbohydrates, *G. assimilis* could be an excellent alternative future source of crude protein and fat[4–6]. In addition, protein concentrate from this species presents high antioxidant and anti-inflammatory activities, making it a functional ingredient in the food industry[7]. Given the economic importance of *G. assimilis*, it is important to obtain a high-quality chromosome-level genome assembly and annotation that can facilitate the generation of genomic tools and resources for this species, directly benefiting scientists and insect breeders.

Genome information provides a foundational resource for various research in insect biology[8,9]. Several cricket genomes have been sequenced and made publicly available, including *Gryllus bimaculatus*[10], *Gryllus. longicercus*[11], *Teleogryllus oceanicus*[12], *T. occipitalis*[13], *Laupala kohalensis*[14], *Acheta domesticus*[15], and *Apteronemobius asahinai*[16]. These genomic resources have enabled diverse studies, including evolutionary biology and entomophagy. For example, the *T. occipitalis* and *A. domesticus* genomes have provided insights into their potential as edible insect species[13,15]. These genomic data not only enhance our understanding of cricket biology but also provide a valuable platform for future research in areas such as pest management, biodiversity conservation, and the development of crickets as a sustainable protein source.

Here, we used short reads generated by an Illumina platform, long reads generated by Oxford Nanopore Technologies (ONT) sequencing, and high-throughput chromosomal conformation capture (Hi-C) analysis

[1]Graduate School of Advanced Science and Engineering, Waseda University, Tokyo, Japan. [2]Ecologgie Inc., Tokyo, Japan. [3]Graduate School of Humanities and Sciences, Ochanomizu University, Tokyo, Japan. [4]Comprehensive Research Organization, Waseda University, Tokyo, Japan. [5]Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, Kraków, Poland. [6]Institute of Engineering, Tokyo University of Agriculture and Technology, Tokyo, Japan. [7]These authors contributed equally: Yuki Ito, Ryuto Sanno. ✉e-mail: kataokak@go.tuat.ac.jp

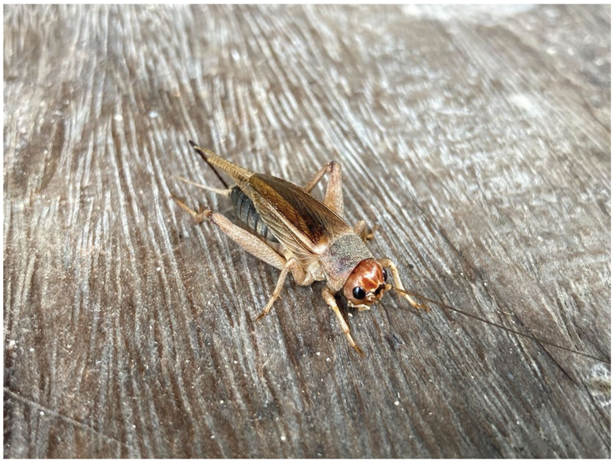

Fig. 1 Photograph of an adult female *Gryllus assimilis*, provided by Ecologgie Inc.

Platform	Raw data (Gbp)	Average read length (bp)	N50 read length (bp)	Coverage (X)
ONT	39.62	17,963	28,647	24.70
Illumina	148.56	150	150	92.63
Hi-C	128.86	150	150	80.35
RNA-seq	50.70	150	150	31.61

Table 1. Statistics for the DNA-seq and RNA-seq data of the *G. assimilis* genome.

to construct a chromosome-scale *G. assimilis* genome. The genome sequences were assembled into 1,100 scaffolds, with an N50 length of 102 Mbp and a total length of 1.60 Gbp. Chromosome scaffolding resulted in 1,101 sequences corresponding to 15 chromosomes. The 15 largest scaffolds, representing chromosome-scale sequences, account for 96.80% of the total scaffolds length. Using *de novo* and homology-based strategies, 27,645 protein-coding genes were revealed by gene annotation. BUSCO analysis against the Arthropoda database showed 99.0% completeness for the gene set and 99.5% for the genome assembly, indicating a high-quality assembly and annotation. The *G. assimilis* genome assembly has a large proportion of repeat sequences (51.42%). This genome assembly and its annotations provide a valuable resource for different fields of science, as well as for the food production sector focused on usage of crickets as food.

Methods

Sample collection and genome sequencing.

The *G. assimilis* individuals used in this study was obtained from a local cricket farm in Takeo Province, Cambodia (Fig. 1). The cricket population at this farm has been maintained without introduction of outside specimens. They were farmed outdoors, exposed to natural environmental conditions. The feed provided during rearing was a mixture of cassava leaves, mung bean residues, rice snack residues, soybean milk residues, commercial poultry feed, supplemented amino acids, and calcium carbonate. Alive *G. assimilis* was used to extract its genomic DNA. Total genomic DNA was extracted from the head and hind legs of a male *G. assimilis* using NucleoBond® HMW DNA (Macherey-Nagel, Germany) according to the manufacturer's instructions. The resulting genomic DNA was size-selected using a Short Read Eliminator Kit (PacBio, CA, USA). Oxford Nanopore Technologies (ONT) sequencing libraries were then constructed and sequenced on the PromethION 2 Solo platform (Oxford Nanopore Technologies, UK) with the Ligation Sequencing Kit V14 and Flow Cell R10.4.1. Base-calling was performed using Dorado v0.3.0 + 88df11b + dirty with the model dna_r10.4.1_e8.2_400bps_sup@v4.2.0[17]. Finally, we obtained 39.62 Gbp ONT sequencing data; average and N50 read lengths were 17.96 Kbp and 28.65 Kbp, respectively (Table 1). Additionally, using the same DNA sample, we prepared a whole genome sequencing library using the TruSeq DNA PCR-Free Library Prep Kit (Illumina, CA, USA) following the manufacturer's protocol. This library was sequenced on the Illumina NovaSeq 6000 platform, generating 148.56 Gbp of short-read data (Table 1). For chromosome-scale scaffolding by the Dovetail™ Omni-C™ Kit (Dovetail Genomics, CA, USA), the head and hind legs of another single male *G. assimilis* were used according to the manufacturer's instructions. The Hi-C sequencing library was built on the Illumina NovaSeq 6000 platform and generated 128.86 Gbp raw data (Table 1). DNA purity and concentrations were measured by spectrometry using NanoPhotometer NP80-TOUCH (Implen, Germany) and fluorometry using Qubit 4 (Thermo Fisher Scientific, MA, USA).

Total RNA was extracted from ten samples: eggs, small nymphs, large nymphs, and sex-specific samples of heads, thoraxes, abdomens, and hind legs from both males and females. The extracted total RNA was purified using RNA Clean & Concentrator Kits (Zymo Research, CA, USA). The RNA-seq library was constructed using NEBNext Ultra II Directional RNA Library Prep Kit following mRNA enrichment by NEBNext Poly(A) mRNA

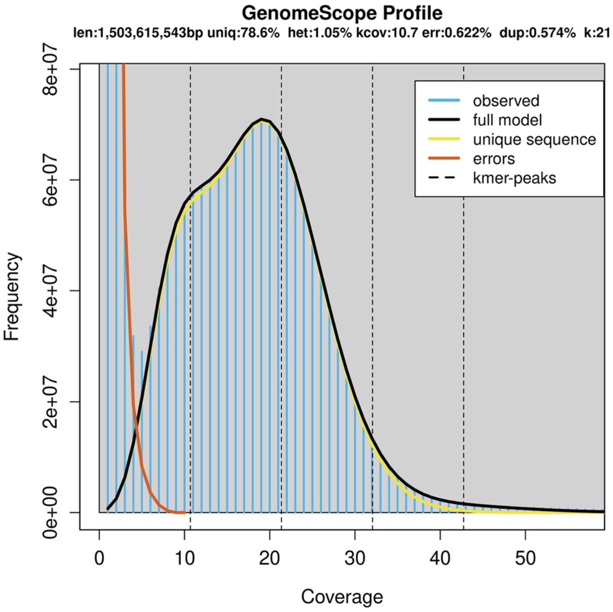

Fig. 2 GenomeScope K-mer distribution of Illumina paired-end reads.

Features	Contig-level	Scaffold-level
Number of sequences	2,432	1,100
Contigs N50 (bp)	4,674,341	102,719,816
GC content (%)	40.37	40.37
Largest sequence (bp)	37,468,416	279,078,345
Total size (bp)	1,627,797,038	1,603,838,333

Table 2. Statistics of the *G. assimilis* genome.

Magnetic Isolation Module, and the RNA sequences were read on Illumina NovaSeq 6000 platform. Finally, we obtained 50.70 Gbp paired-end raw reads (Table 1).

Genome size estimation.

Low-quality reads from the original ONT sequences were filtered by NanoFilt v2.8.0[18] with the parameter -q 10. Then, 39.60 Gbp of the clean reads were used to estimate the genome size, heterozygosity, and repeat content of the genome using Jellyfish v2.3.0[19], with a 21-mer frequency and the parameter set as reads_cutoff = 1k. GenomeScope v2.3.0[20] was then used to analyze the K-mer frequency distribution. The genome size was estimated at 1.50 Gbp with 1.05% heterozygosity and 78.6% repetitive sequences (Fig. 2).

De novo genome assembly.

The genome was assembled by integrating the clean ONT long reads, Illumina short reads, and Hi-C reads. Long reads generated from the PromethION sequencer were assembled using Flye v2.9.1[21]. Assembly continuity and gene completeness were evaluated using gVolante[22,23]. Gene completeness was specifically assessed using Benchmarking Universal Single-Copy Orthologs (BUSCO v5)[24,25], which is implemented in gVolante, with the arthropoda database. The resulting contigs underwent three rounds of error correction using POLCA v4.1.0[26] with default setting and Illumina pair-end read data. This process yielded a 1.63 Gbp draft genome, comprising 2,432 contigs with an N50 of 4.67 Mbp (Table 2).

Potential contamination in the assembly was removed using BlobToolKit v1.1.1[27], 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[28]. Similarity analysis was performed using BLASTn v2.13.0+[29] against NCBI NT database v5 (options: -task megablast culling_limit 10 -evalue 1e-25 -outfmt '6 qseqid staxids bitscore std sscinames sskingdoms stitle'). Mitochondrial genomes were also identified through gene prediction using the MITOS2[30] webserver (accessed on November 28, 2023) and subsequently removed. As a result, one contig was removed as a bacterial genome and another as a mitochondrial genome.

The final draft contig assembly was produced after removing duplicated contigs with Purge Haplotigs v1.1.2[31], using input generated from the long-read mapping data by bwa v0.7.17-r1188. The assembled contigs were then corrected for misjoins, ordered, oriented, and anchored into a chromosome-scale assembly using Omni-C™ data with Juicer v1.9.9[32] and 3D-DNA v180419[33]. Candidate assembly was reviewed with Juicebox Assembly Tools v1.9.9 for quality control and interactive corrections. The contact map (Fig. 3) was visualized using Juicebox, displaying interactive signals between each pair of bins. The resulting genome was 1.60 Gbp in size, consisting of 1,100 scaffolds with an N50 length of 102 Mbp (Table 2). This includes 15 pseudochromosomes

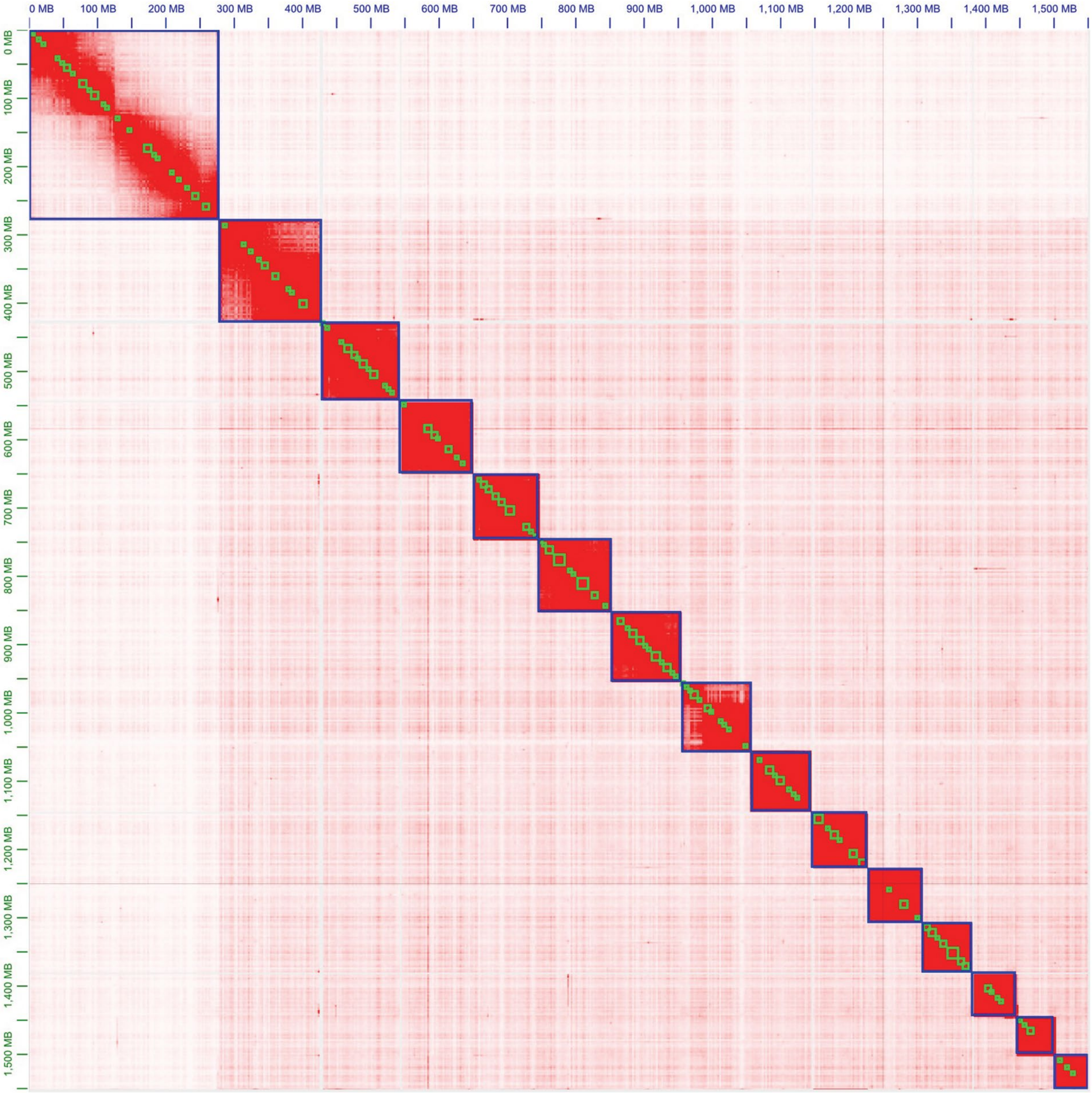

Fig. 3 Hi-C contact heatmap of the *G. assimilis* genome assembly.

accounting for 96.8% of the total genome length (Table 3). Additionally, a Circos plot, illustrating the distribution of genomic elements (Fig. 4), was generated using Circos v0.69-9[34].

Prediction of repeat regions and functional annotation of protein-coding genes.

In *de novo* repeat prediction, RepeatModeler v2.0.5[35] was first used for *de novo* repeat identification, and this library was supplemented with known repeat sequences from a closely related species, *G. bimaculatus*[10]. The combined repeat library was then used to identify and softmask repetitive elements in the *G. assimilis* genome using RepeatMasker v4.1.5[36] (Table 4). The structural annotation for protein-coding genes was performed on the softmasked genome using *ab inito* prediction, homology-based prediction, and RNA-seq-based prediction. Each prediction used RNA-seq data as input.

To remove noisy RNA-seq reads potentially arising from erroneous transcription and splicing, *do novo* transcriptome assembly was first performed using Trinity v2.15.1[37] to generate contigs. The original RNA-seq reads were then mapped back to these contigs using HISAT2 v2.2.1[38] 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 v3.0.2[39–44], incorporating protein data from OrthoDB 11's arthropods dataset[45] and the mapping data of the filtered RNA-seq reads. This BRAKER prediction served as the foundation for our gene set. To complement and improve this base set, we employed two additional

Chromosome	Length (bp)	Proportion in genome (%)
chrX	279,078,345	17.40
chr1	150,802,759	9.40
chr2	114,372,951	7.13
chr3	106,619,747	6.65
chr4	106,313,648	6.63
chr5	102,719,816	6.40
chr6	102,347,583	6.38
chr7	96,493,873	6.02
chr8	88,467,910	5.52
chr9	81,245,571	5.07
chr10	80,799,356	5.04
chr11	72,350,121	4.51
chr12	64,377,000	4.01
chr13	57,033,950	3.56
chr14	49,462,500	3.08
Total	1,552,485,130	96.80

Table 3. Statistics of chromosomes in the *G. assimilis* genome.

approaches. GeMoMa v1.9.0[46] was used for the homology-based prediction with gene sets from four species (*Apis mellifera*, *Drosophila melanogaster*, *Tribolium castaneum*, *Teleogryllus occipitalis*), while StringTie2 v2.2.1[47] was used for RNA-seq-based prediction. The predictions from GeMoMa and StringTie2 were used to identify and add genes that BRAKER3 had missed, resulting in an additional 14,080 genes to our gene set. These combined predictions were merged and the duplicate genes were discarded using GffCompare v0.12.6[48] to form a final, comprehensive consensus gene set.

Gene functional annotation was conducted using eggNOG-mapper online (http://eggnog-mapper.embl.de/)[49] and BLASTp-based methods. For the BLASTp-based annotation, we used databases including *Homo sapiens*, *Mus musculus*, *Caenorhabditis elegans*, *D. melanogaster*, and UniProt Swiss-Prot[50] to identify the best hits for annotation (E-value $< 1.0 \times 10^{-10}$) (Table 5).

Data Records

The raw sequencing data (Illumina, ONT, and Hi-C) used for genome assembly have been deposited in the Sequence Read Archive (SRA) under the accession number SRP530093[51].

The assembled genome has been deposited at Genbank under the accession number GCA_046254815.1[52].

The assembled genome and annotation datasets are also available in figshare[53].

Technical Validation

Genome assembly and annotation completeness evaluation.

To assess the genome quality, the completeness of the final genome assembly was evaluated using BUSCO v5.1.2[24,25] with the arthropoda gene set in the gVolante[22,23] webserver. Out of 1,013 single-copy orthologues, 99.5% were completely identified in the *G. assimilis* genome. The full BUSCO results were as follows: C: 99.5%[S: 97.0%, D: 2.5%], F:0.3%, M:0.2%. (Table 6).

Moreover, 27,645 protein-coding genes were obtained by combining *ab initio*, homology-based, and RNA-seq-based prediction. Of the predicted genes, 16,938 were functionally annotated with significant hits (E-value $< 1.0 \times 10^{-10}$) in at least one of these annotation resources: eggNOG-mapper or BLASTp searches against *H. sapiens*, *M. musculus*, *C. elegans*, *D. melanogaster*, and UniProt Swiss-Prot databases. A BUSCO analysis was performed to assess the completeness of our gene annotations. This analysis identified 99.0% of the expected complete arthropod BUSCOs. The full results were: C: 99.0%[S: 96.5%, D: 2.5%], F:0.6%, M:0.4%. These results collectively indicate a high-quality gene set for this species.

Genome assembly accuracy evaluation.

To identify the X chromosome, we sequenced a male (XO) *G. assimilis* individual. The sequenced reads were mapped to the G. *assimilis* genome, and read depth was calculated in 100 Kbp windows. We observed that the longest scaffold exhibited nearly half the read depth of the other chromosomes, suggesting it is the X chromosome[54] (Fig. 5). This finding is consistent with karyotype studies in the related species *G. bimaculatus*, where the X chromosome is also the largest[54]. The remaining chromosomes have similar coverage, indicating an absence of X chromosome–autosome chimeras.

We also compared the gene structure between *G. assimilis* and a published chromosome-scale cricket genome from *A. domesticus*[15] using MCScanX[55] and SynVisio[56]. For the MCScanX, BLASTp was carried out with the following options: -evalue 1e-10 -outfmt 6 -max_target_seqs. 5. This comparison revealed a strong collinearity relationship between the two species, particularly for the X chromosome, while highlighting notable autosome rearrangements (Fig. 6).

Additionally, we built a phylogenetic tree together with the genomes of *G.assimilis*, *A domesticus*, *A asahinai*, *G. bimaculatus*, *G. longicercus*, *L. kohalensis*, *T. occipitalis*, *Locusta migratoria*[57], and *Schistocerca gregaria*[58] (Fig. 7). First, we used OrthoFinder v2.5.5[59] (option: -S blast) to identify single-copy orthologs among these species. Each orthologous gene was then aligned with MAFFT v7.520[60] (option:–auto), and poorly aligned regions

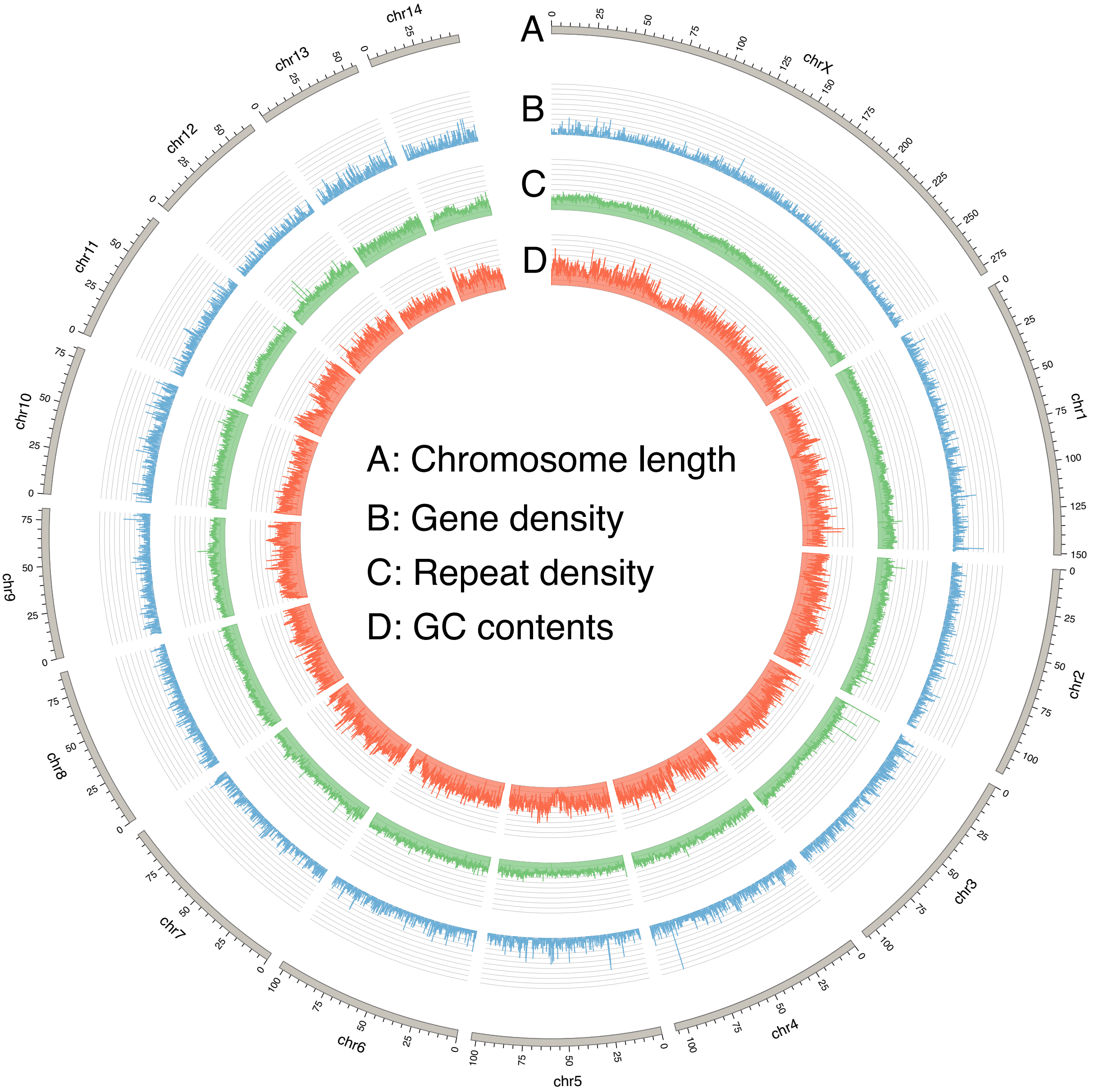

Fig. 4 Circos plot of 15 chromosomes of *G.assimilis*. From outer to inner layers were chromosome length in Mbp (**A**), gene density (**B**), repeat density (**C**), GC contents (**D**).

Repeat type	Length (bp)	Proportion in genome (%)
SINEs	22,853,434	1.42
Penelope	169,953	0.01
LINEs	162,567,307	10.14
LTR elements	46,532,505	2.90
DNA transposons	181,726,613	11.33
Unclassified	272,324,563	16.98
Total interspersed repeats	686,174,375	42.78
Small RNA	4,188,571	0.26
Satellites	8,166,710	0.51
Simple repeats	101,219,664	6.31
Low complexity	4,842,055	0.30
Total	824,753,490	51.42

Table 4. Classification of repetitive sequences of the *G. assimilis* genome.

Features	
Number of genes	27,645
Number of mRNA	27,777
Number of CDSs	133,786
Homo sapiens (GRCh38)	44.84%
Mus musculus (GRCm39)	43.08%
Caenorhabditis elegans (WBcel235)	37.99%
Drosophila melanogaster (BDGP6.32)	45.64%
Uniprot/Swissprot (release: 2020_06)	48.66%
eggNOG-mapper	56.78%

Table 5. Statical analysis of the gene annotation of the *G. assimilis* genome.

BUSCO	Assembly	Gene model
Complete BUSCOs	99.5	99.0
Single-copy complete BUSCOs	97.0	96.5
Duplicated complete BUSCOs	2.5	2.5
Fragmented BUSCOs	0.3	0.6
Missing BUSCOs	0.2	0.4

Table 6. Statistics for genome assessment using BUSCO.

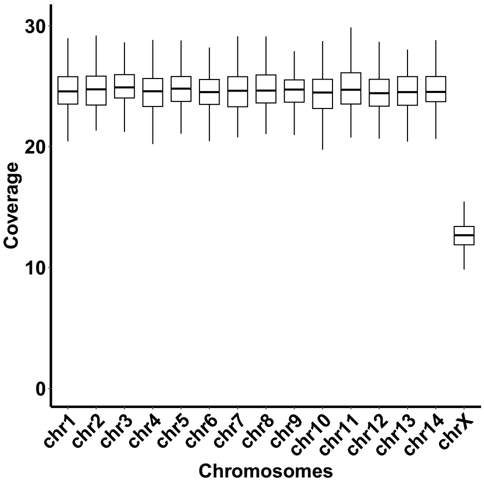

Fig. 5 Sequencing coverage distribution across 15 chromosome-scale scaffolds in male *G. assimilis*. Boxplot shows the sequencing coverage distribution for each chromosome-scale scaffold in a male (XO) *G. assimilis* individual. Coverage was calculated in 100 Kbp windows.

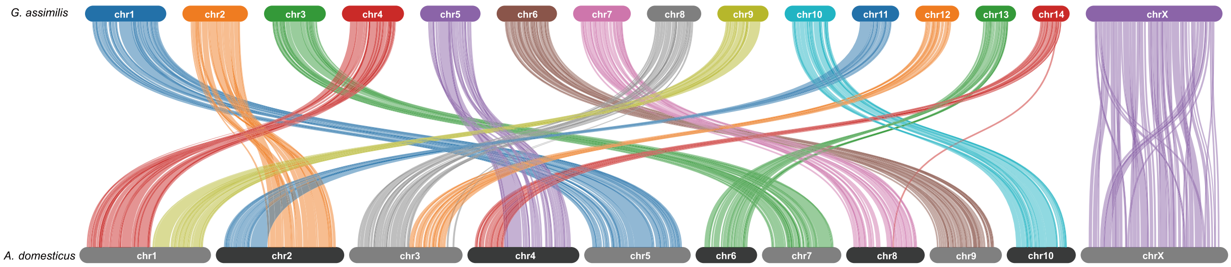

Fig. 6 Genomic synteny analysis between *Gryllus assimilis* (top) and *Acheta domesticus* (bottom). Each colored line represents conserved syntenic blocks between the two species, with different colors corresponding to different *G. assimilis* chromosomes. The width of each line is proportional to the number of syntenic genes in the block, with each line representing a minimum of 5 consecutive orthologous genes.

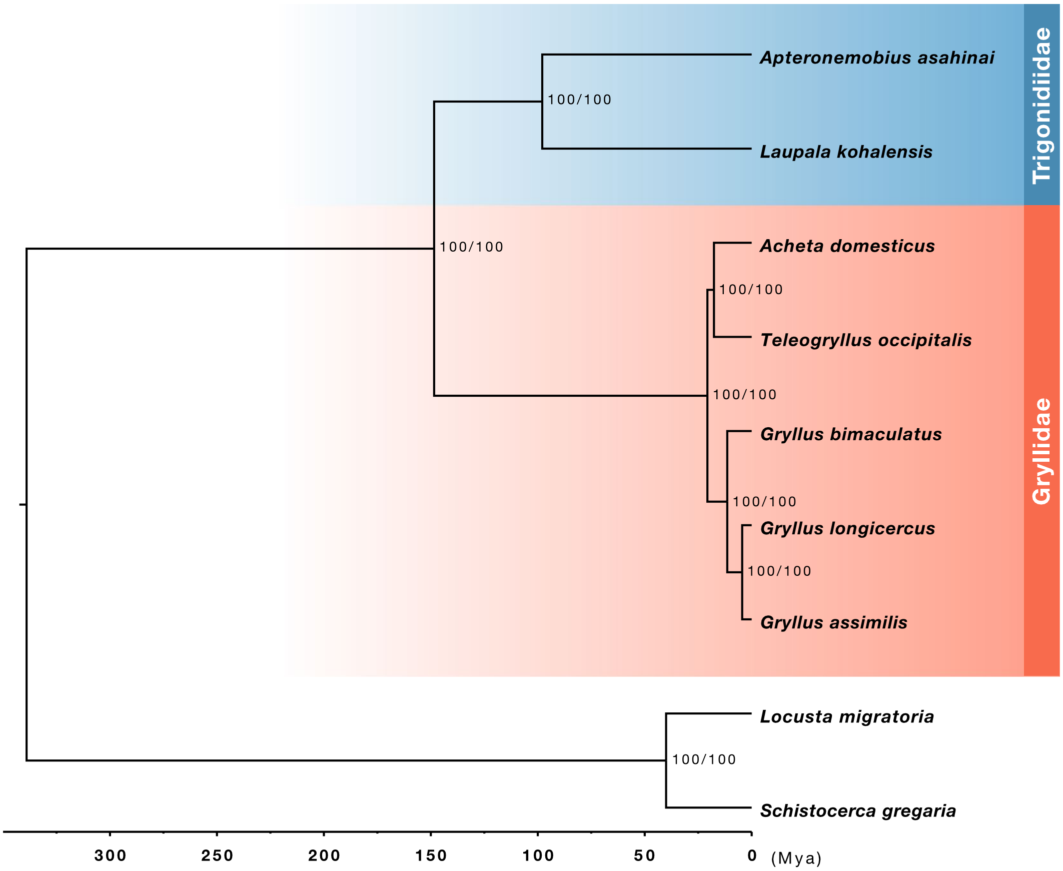

Fig. 7 Phylogenetic tree of crickets inferred by maximum likelihood analysis from a partitioned dataset of 371 concatenated single-copy orthologs. Numbers on the branches denote bootstrap support values. *Locusta migratoria* and *Schistocerca gregaria* were used as outgroups. The timeline at the bottom indicates estimated divergence times (in millions of years ago, Mya).

were removed using trimAl v1.4[61] (option: -automated1). After that, using a custom Python script, we generated per-ortholog partitions from all single-copy FASTA files and subsequently inferred phylogenetic relationships with IQ-TREE v2.2.3[62] (options: -nt AUTO -bb 1000 -m MFP -alrt 1000). Divergence time was obtained from TimeTree[63] (https://timetree.org/) and is indicated on the resulting phylogeny. This analysis confirmed the expected phylogenetic relationships among cricket species.

Code availability

The scripts used for the analyses in this study are available in figshare[51] and GitHub (https://github.com/kataokaklab/Gryllus_assimilis_genome). All bioinformatics tools used in this study followed their respective manuals and protocols. The software versions, codes, and parameters are provided in the Methods section. Unless otherwise specified, default parameters were used.

Received: 30 December 2024; Accepted: 14 May 2025;
Published online: 20 May 2025

References

1. Alexander, R. D. & Walker, T. J. Two introduced field crickets new to eastern United States (Orthoptera: Gryllidae). *Annals of the Entomological Society of America.* **55**, 90–94 (1962).
2. Walker, T.J. Jamaican field cricket, *Gryllus assimilis* (Fabricius 1775). *Sing. Insects N. Am.* (2011).
3. Murugu, D. K. *et al.* From farm to fork: crickets as alternative source of protein, minerals, and vitamins. *Front. Nutr.* **8**, 704002 (2021).
4. Mlček, J. *et al.* Selected nutritional values of field cricket (*Gryllus assimilis*) and its possible use as a human food. *Indian J. Tradit. Knowl.* **17** (2018).
5. Ribeiro, G. H. M. *et al.* Dietary supplementation with black cricket (*Gryllus assimilis*) reverses protein-energy malnutrition and modulates renin-angiotensin system expression in adipose tissue. *Food Res. Int.* **114570** (2024).
6. Quinteros, M. F., Martínez, J., Barrionuevo, A., Rojas, M. & Carrillo, W. Functional, antioxidant, and anti-inflammatory properties of cricket protein concentrate (*Gryllus assimilis*). *Biology.* **11**, 776 (2022).
7. Hassen, H. *et al.* Effect of diets with the addition of edible insects on the development of atherosclerotic lesions in ApoE/LDLR—/— mice. *Int. J. Mol. Sci.* **25**, 7256 (2024).
8. Kataoka, K., Togawa, Y., Sanno, R., Asahi, T. & Yura, K. Dissecting cricket genomes for the advancement of entomology and entomophagy. *Biophys Rev.* **14**, 75–97 (2022).

9. Sanno, R. *et al*. Comparative analysis of mitochondrial genomes in Gryllidea (Insecta: Orthoptera): Implications for adaptive evolution in ant-loving crickets. *Genome Biol. Evol.* **13** (2021).
10. Ylla, G. *et al*. Insights into the genomic evolution of insects from cricket genomes. *Commun. Biol.* **4**, 733 (2021).
11. Szrajer, S., Gray, D. & Ylla, G. The genome assembly and annotation of the cricket *Gryllus longicercus*. *Sci. Data.* **11**, 708 (2024).
12. Pascoal, S. *et al*. Field cricket genome reveals the footprint of recent, abrupt adaptation in the wild. *Evol. Lett.* **4**, 19–33 (2019).
13. Kataoka, K. *et al*. The draft genome dataset of the Asian cricket *Teleogryllus occipitalis* for molecular research toward entomophagy. *Front. Genet.* **11**, 470 (2020).
14. 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).
15. Dossey, A. T. *et al*. Genome and genetic engineering of the house cricket (*Acheta domesticus*): A resource for sustainable agriculture. *Biomolecules.* **13**, 589 (2023).
16. Satoh, A., Takasu, M., Yano, K. & Terai, Y. *De novo* assembly and annotation of the mangrove cricket genome. *BMC Res. Notes.* **14**, 387 (2021).
17. Mike, V. Dorado – a modern, C++, Totch-based basecaller. *Oxford Nanopore Technologies*. https://nanoporetech.com/ja/resource-centre/lc22-dorado-a-modern-torch-based-basecaller (2022).
18. De Coster, W., D'hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics.* **34**, 2666–2669 (2018).
19. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics.* **27**, 764–770 (2011).
20. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics.* **33**, 2202–2204 (2017).
21. Kolmogorov, M. *et al*. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods.* **17**, 1103–1110 (2020).
22. Nishimura, O., Hara, Y. & Kuraku, S. gVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics.* **33**(22), 3635–3637 (2017).
23. Nishimura, O., Hara, Y. & Kuraku, S. Evaluating genome assemblies and gene models using gVolante. *Methods Mol. Biol.* **1962**, 247–256 (2019).
24. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. 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).
25. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: assessing genomic data quality and beyond. *Curr. Protoc.* **1**, e323 (2021).
26. Zimin, A. V. & Salzberg, S. L. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS Comput. Biol.* **16**, e1007981 (2020).
27. Laetsch, D. R. & Blaxter, M. L. BlobTools: interrogation of genome assemblies. *F1000Res.* **6**, 1287 (2017).
28. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2014).
29. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinformatics.* **10**, 1–9 (2009).
30. Bernt, M. *et al*. MITOS: improved *de novo* metazoan mitochondrial genome annotation. *Mol. Phylogenet. Evol.* **69**, 313–319 (2013).
31. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics.* **19**, 1–10 (2018).
32. Robinson, J. T. *et al*. Juicebox.js provides a cloud-based visualization system for Hi-C data. *Cell Syst.* **6**, 256–258 (2018).
33. Dudchenko, O. *et al*. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science.* **356**, 92–95 (2017).
34. Krzywinski, M. *et al*. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
35. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA.* **117**, 9451–9457 (2020).
36. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0 http://www.repeatmasker.org (2015).
37. Grabherr, M. G. *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
38. Kim, D. *et al*. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
39. Gabriel, L. *et al*. BRAKER3: fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* (2024).
40. Brůna, 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 Genom. Bioinform.* **3**, lqaa108 (2021).
41. Hoff, K. J., Lomsadze, A., Borodovsky, M. & Stanke, M. Whole-genome annotation with BRAKER. *Methods Mol. Biol.* **1962**, 65–95 (2019).
42. 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).
43. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics.* **24**, 637–644 (2008).
44. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics.* **7**, 1–11 (2006).
45. Kuznetsov, D. *et al*. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res.* **51**, D445–D451 (2023).
46. Keilwagen, J., Hartung, F. & Grau, J. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Gene Prediction: Methods and Protocols.* **1962**, 161–177 (2019).
47. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 1–13 (2019).
48. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Res.* **9**, 304 (2020).
49. 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**, D309–314 (2019).
50. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
51. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP530093 (2024).
52. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046254815.1 (2024).
53. Yuki, I. & Kosuke, K. Chromosome-scale whole genome sequences and annotation of the Jamaican field cricket *Gryllus assimilis*. *Figshare* https://doi.org/10.6084/m9.figshare.26761927 (2024).
54. Yoshimura, A., Nakata, A., Mito, T. & Noji, S. The characteristics of karyotype and telomeric satellite DNA sequences in the cricket, Gryllus bimaculatus (Orthoptera, Gryllidae). *Cytogenet Genome Res.* **112**, 329–336 (2006).
55. Wang, Y. *et al*. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
56. Bandi, V. & Gutwin, C. Interactive exploration of genomic conservation. In *Proc. 46th Graphics Interface Conf. on Graphics Interface 2020 (GI'20)*. Canadian Human-Computer Communications Society, Waterloo, CAN (2020).

57. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_026315105.1 (2022).
58. Verlinden, H. *et al*. First draft genome assembly of the desert locust, Schistocerca gregaria. *F1000Res.* **9**, 775 (2020).
59. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
60. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
61. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
62. Minh, B. Q. *et al*. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
63. Kumar, S. *et al*. TimeTree 5: An expanded resource for species divergence times. *Mol. Biol. Evol.* **39** (2022).

Acknowledgements

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].

Author contributions

Y.I. and K.K. designed and led the project. Y.I., R.S. and K.K. performed the analyses. S.A. provided the *G. assimilis* samples. G.Y. contributed to the custom repeat library. Y.I. prepared the figures and wrote the first draft of the manuscript. K.K. and G.Y. edited the final version. K.K., T.A., and K.Y. supervised this study. All authors reviewed and accepted the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

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

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.