A chromosome-level genome assembly of Caligula japonica as a resource for evolutionary studies in Lepidoptera

Article *in* Entomologia Generalis · November 2023

DOI: 10.1127/entomologia/2023/2004

CITATION
1

READS
277

6 authors, including:

Xu Chen
Guizhou University
25 PUBLICATIONS **242** CITATIONS
SEE PROFILE

Yong-Ming Chen
Guizhou University
27 PUBLICATIONS **273** CITATIONS
SEE PROFILE

Su Wang
Institue of plant and environment protection, Beijing academy of agriculture and …
165 PUBLICATIONS **3,510** CITATIONS
SEE PROFILE

Xinhai Ye
Zhejiang A&F University
64 PUBLICATIONS **673** CITATIONS
SEE PROFILE

Entomologia Generalis
Published online November 2023

A chromosome-level genome assembly of *Caligula japonica* as a resource for evolutionary studies in Lepidoptera

Xu Chen[1], Yong-Ming Chen[1], Su Wang[2], Xin-Hai Ye[3], Meng-Yao Chen[3], Lian-Sheng Zang[1,*]

[1] National Key Laboratory of Green Pesticide, Key Laboratory of Green Pesticide and Agricultural Bioengineering, Ministry of Education, Guizhou University, Guiyang, China
[2] Institute of Plant and Environment Protection, Beijing Academy of Agricultural and Forestry Sciences, Beijing, China
[3] Ministry of Agriculture Key Lab of Molecular Biology of Crop Pathogens and Insects, Institute of Insect Science, Zhejiang University, Hangzhou, China
* Corresponding author: lsz0415@163.com

With 4 figures and 2 tables

Abstract: *Caligula japonica*, a forestry pest known for its destructive impact on various trees, has recently shown promise as a potential natural medical mesh biomaterial in the medical industry. However, the lack of genomic resources has hindered in-depth studies on the karyotype evolution and functional genomics of this species. In this study, we conducted high-throughput sequencing of the *C. japonica* genome, yielding the first chromosome-level assembly. We successfully assembled a high-quality genome of 584,537,256 bp with Contig N50 of 12 Mb and 31 chromosomes. About 342 Mb repeat sequences were identified, accounting for 59.16% of *C. japonica* genome. Genome annotation by de novo gene prediction and homologous gene search yields 20,887 protein-coding genes. Furthermore, a phylogenetic analysis of gene families linked to detoxification and chemosensory revealed a pronounced expansion in the P450, ABC, and GR gene families within *C. japonica* and *Bombyx mori*. The genome obtained in this research will provide data basis support for research on resource utilization of *C. japonica* and to progress comparative genomic analyses and evolutionary adaptability in Lepidoptera.

Keywords: Japanese giant silkworm; Comparative genomic analyses; Evolutionary adaptability; Chromosome; Phylogenetic analysis

1 Introduction

The Japanese giant silkworm, *Caligula japonica* (Moore 1862) (Lepidoptera: Bombycoidea) (Fig. 1), is generally considered a forestry pest due to its damage to many trees. (Kawaguchi et al. 2003; Qiao et al. 2014). In China, 38 plant species from 30 genera across 20 families have been reported as hosts of this pest. These include several important fruit trees such as walnut (*Juglans regia* L.), chestnut (*Castanea mollissima* Blume), ginkgo (*Ginkgo biloba* L.), plums (*Prunus* spp.), pears (*Pyrus* spp.), apples (*Malus pumila* Mill.), and sumac (*Toxicodendron vernicifluum* (Stokes)) (Yang et al. 2008; Qiao et al. 2014). It is widely distributed in Asia, and its damage causes enormous economic losses every year in countries such as Japan, North Korea, Russia, and particularly in China (Chen et al. 2019). However, recent research indicated that the cocoon of pupa displays nontoxicity, biocompatibility, suitable mechanical properties, and porosity while showing no adverse effect in animal trials and even appears to enhance cell proliferation so that it could be used as natural medical mesh in the medical industry (Chen et al. 2022). Thus, mesh production holds significant economic potential. However, a critical factor hindering its development is that this moth undergoes a single generation per year and a lack of research on molecular genetic characteristics.

Previous studies on *C. japonica* have primarily focused on nucleotide diversity analysis, and control methods, especially biological control excavation (Li et al. 2009a; Zang et al. 2023; Chen et al. 2021a). The absence of research on the biological characteristics and evolutionary pathways of this insect hampers its potential as a valuable resource insect. Consequently, we aim to use genomic analysis to enhance the cultivation of this insect and harness its beneficial traits.

DOI: 10.1127/entomologia/2023/2004
E. Schweizerbart'sche Verlagsbuchhandlung, 70176 Stuttgart, Germany, www.schweizerbart.de

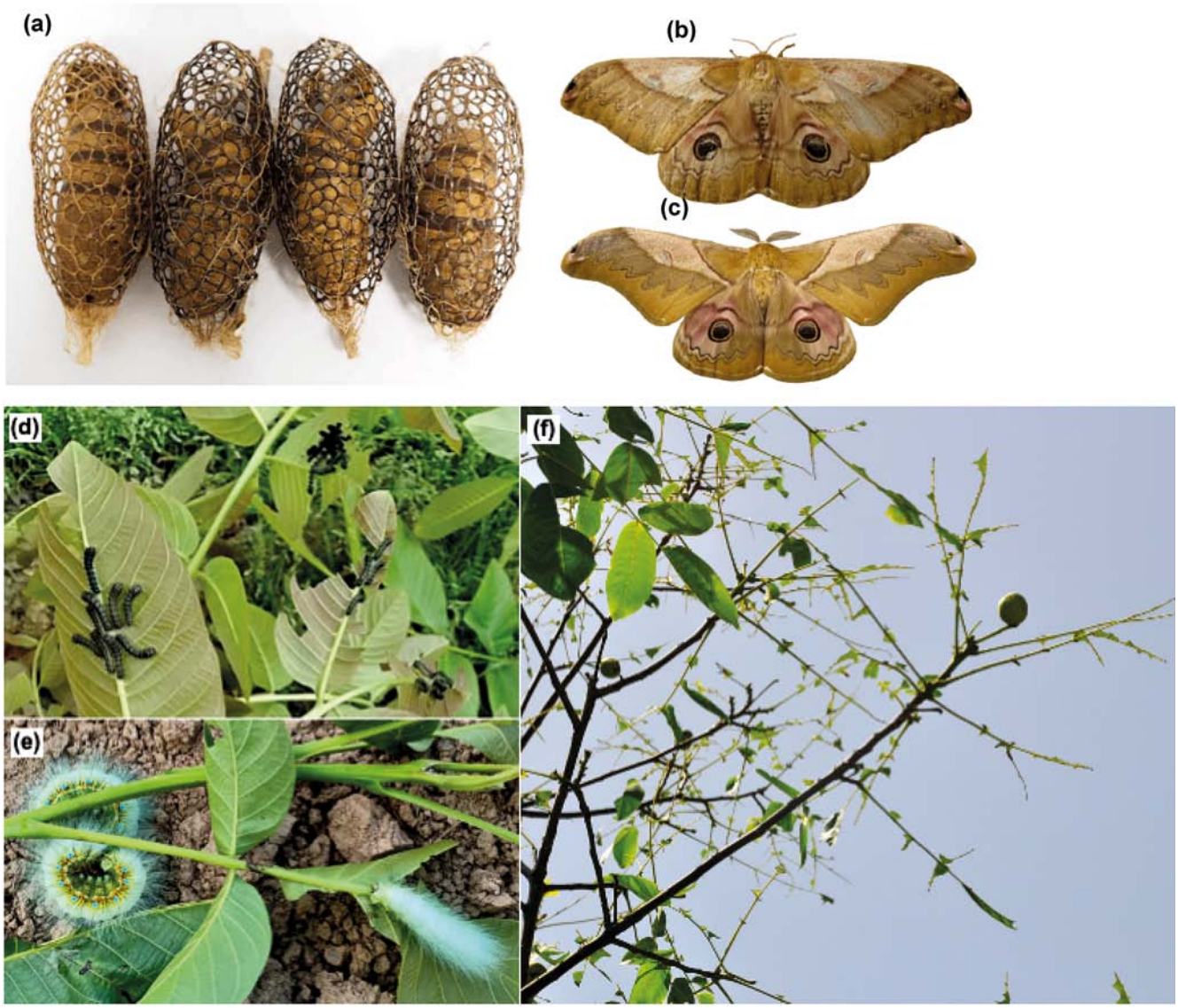

Fig. 1. Morphological and damage photographs of *Caligula japonica*. **(a)** Pupae with reticulated cocoon; **(b)** Female adult; **(c)** Male adult; **(d)** The 3rd instar larvae gather to feed on walnut plant leaves; **(e)** The 5th instar larvae feed on walnut plant leaves; **(f)** After a large-scale outbreak of larvae, walnut leaves are almost completely consumed.

A high-quality genome provides a valuable platform to explore the functions of crucial genes involved in insect physiological regulation (Chen et al. 2021b). By accurately identifying and annotating genes in the genome, researchers can focus on key regulatory genes. Here, in an initial effort to explore the evolutionary adaptability of *C. japonica*, we conducted high-throughput sequencing of its genome and obtained a high-quality chromosome-level genome. The detoxification metabolic genes and chemoreception gene families closely related to adaptive evolution were also identified in the *C. japonica* genome and conducted phylogenetic analysis. The availability of our high-quality genome sequence resources promotes the advancement of research on *C. japonica*.

2 Materials & methods

2.1 Insect collection

In its natural environment, *C. japonica* is known to primarily feed on walnut trees (*J. regia*). We collected the pupae from walnut trees on the Chen jialiang mountain from Longnan, Gansu province, China, in July 2019. Then pupae were placed in the Key Laboratory of Green Pesticide and Agricultural Bioengineering of Ministry of Education, Guizhou University, in a 30 cm × 30 cm × 30 cm nylon cage with a 200-mesh size under natural temperature, humidity, and light conditions for 40 days. After emerging, one male was collected and stored at a temperature of −80 °C for Illumine and PacBio sequencing. After fertilization, the eggs

hatched, and the emerging larvae were provided with walnut leaves. Upon reaching a specific developmental stage, they were preserved at −80 °C for later RNA extraction and transcriptome analysis.

2.2 Genome size estimation by flow cytometry

Flow cytometry was used to estimate the genome size of *C. japonica* according to the standard procedure (He et al. 2016). One head of male adult was homogenized completely with 500 μL ice-cold Galbraith's buffer (PH = 7). The homogenate was centrifuged at 5,000 rpm at 25 ℃ for 5 min and suspended with 400 μL phosphate buffer (PH = 7.4). To remove the RNA, RNaseA was added to a tube at 25 ℃ for 10 min (final concentration of 20 μg/mL). Finally, samples were stained with 50 μg/mL propidium iodide stock solution in darkness at 4 ℃ for 10 min. Samples were analyzed by the FACSCalibur platform (BD Biosciences) and FACScomp v4.0 under 488-nm wavelength. FlowJo v7.6.1 was used to obtain the nuclei peaks. *Drosophila melanogaster* was analyzed as a control following the same above parameter. The outputs were used to estimate the genome size.

2.3 Genome sequencing and assembly

High-quality genomic DNA for de novo sequencing was extracted from 1 male adult of *C. japonica* using the Genome DNA extraction Kit (TIANGEN, Cat. DP304-03) according to the protocol. Illumine sequencing was performed to evaluate genome size, heterozygosity, and rate of duplication and polish de novo assembly. A paired-end library (insert size: 350 bp) was constructed on Illumina NovaSeq platform. To ensure the quality of information analysis data, we filter the original sequence as followed: (a) Removing the contaminated Reads from the joint; (b) Removing low-quality Reads (bases with a mass value of Q ≦ 19 in Reads account for more than 50% of the total base count. For double-ended sequencing, if one end is low-quality Reads, both end Reads will be removed); (c) Removing Reads with a N content ratio greater than 5%. After filtering, we yielded a total of 112.81 Gb clean data with 176× sequence coverage.

High-quality genome DNA (extracted using the same method as Illumine sequencing) was fragmented to construct a PCR-free SMRT bell library. After the library size was tested to be qualified by Qubit 3.0 and Agilent 2100, it was sequenced on a SMRT cell by PacBio Sequel II sequencing platform (Pacific Biosciences) with ×186.17 Mean Depth. We obtained 169.37 Gb clean data after filtering and 7,960,820 subreads (mean subreads length: 21,275.65 bp, subreads length N50: 31,540 bp). CANU v2.2 (Koren et al. 2017) corrected row data generated from PacBio sequencing with the default parameters. In the assembly phase, reads were assembled into contig and output consensus sequences by WTDBG v2 (Ruan et al. 2020) with default parameters. PBMM2 (MINIMAP2) (https://github.com/PacificBiosciences/pbmm2) was used to map original data to the reference genome, and ARROW (RACON) (github.com/lbcb-sci/racon) was used for polishing. The previously polished FASTA sequence was indexed with BWA index, and the corrected genome was used as the reference genome. Then, the Illumina sequencing FASTQ data were compared with the BWA MEM to perform Pilon error correction for secondary polishing. To remove the redundancy of the genome after preliminary assembly and error correction, PURGE_HAPLOTIGS software was used to identify and remove the redundant heterozygous contigs according to the depth distribution of reads and sequence similarity.

2.4 Hi-C assisted assembly

To obtain the chromosome-level genome, we used (High-throughput/resolution chromosome conformation capture) Hi-C technology to assist assembly. After the collected pupae have emerged, we collected the fertilized eggs produced after mating. 50 eggs were treated with paraformaldehyde to fixed DNA conformation for 10 min and stopped crosslinking by 2.5 M glycine for 20 min. Crosslink DNAs were cut with a restriction enzyme and produced fill ends with biotin, which was used to build the library and subsequent sequencing via the Illumina platform. High-quality clean data 58.615 Gb (read length: 150 bp) was generated after sequencing and filtering, then used for preliminary assembly by applying a 3D-DNA pipeline using default parameters (Dudchenko et al. 2017). We employed Juicer to construct the chromosome interaction map and then utilized Juicebox for visual correction. This allowed us to identify potential errors in contig sequence, direction, or assembly within the contig, ensuring the accuracy and reliability of our genome assembly. The quality of the genome sequence was evaluated by BUSCO v4 with Lepidoptera_ODB10 and default parameters (Manni et al. 2021). We used BWA v0.7.17 (Li 2013), SAMtools v1.9 (Li et al. 2009) and bedtools v2.29.2 (Quinlan et al. 2010) to align reads, remove duplicates and calculate genome coverage. The number of Ns and total genome length were obtained using in-house Perl scripts. The N percentage was calculated by dividing Ns by genome length.

2.5 Transcriptome sequencing and analysis

To assist in genome annotation, transcriptomic libraries were prepared from the 1st, 2nd, 3rd, 4th, and 5th instar larvae of *C. japonica*. Each larva designated for sequencing had an individual library constructed for the procedure. Total RNA was isolated from individual *C. japonica* larvae utilizing the TRIzol (Invitrogen, Carlsbad, CA, USA) reagent method. Following homogenization, samples were allowed to stand at ambient conditions before chloroform was introduced. The mixture underwent centrifugation at 12,000 g at 4 °C, allowing for phase separation. The aqueous phase was subsequently subjected to isopropanol precipitation and centrifugation. The RNA pellet obtained was rinsed in 75% ethanol (prepared with RNase-free water) and centrifuged twice to ensure purity. The air-dried pellet was reconstituted

in DEPC-treated water, and its integrity and concentration were quantified using a Nanodrop-2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, United States) at 260 nm. The RNA samples that had good quality were then utilized for cDNA library construction. The Illumina HiSeq X platform was used for transcriptome sequencing with 150 bp paired-end reads. After removing low-quality reads, the sequence contaminated by the connector and containing more than 5% N, 34.9 Gb clean reads were obtained. The reads were used for transcriptome splicing by Trinity v2.14.0 (Grabherr et al. 2011; Haas et al. 2013) with the default parameters. The obtained spliced transcript was used for genome structure annotation to provide evidence of transcription level.

2.6 Annotation of repeats

The repeat sequences of *C. japonica* were marked by combining RepeatModeler v2.0.2 and RepeatMasker v4.1.2 (http://www.repeatmasker.org/). Firstly, a de novo repeat database was built by RepeatModeler. Subsequently, we utilized RepeatMasker to annotate the repeat sequences based on the Repbase database (https://www.girinst.org/server/RepBase/index.php). The random repeats were detected by TRF (Tandem Repeats Finder) v4.09 (Benson 1999). And the sequence search engine was RMBlast v2.11.0 (http://www.repeatmasker.org/RMBlast.html) for sequence alignment.

2.7 Gene prediction and annotation

For protein-coding gene annotation in the *C. japonica* genome, we employed de novo prediction and homologous gene search. The repeat-masked genome was then subjected to further analysis according to the MAKER v3.01.03 genome annotation pipeline (Cantarel et al. 2008). First, we utilized BRAKER v2 to construct the parametric species model for the *C. japonica* genome (Bruna et al. 2021; Hoff et al. 2016, 2019; Stanke et al. 2008, 2006). Next, we employed Trinity to perform transcript splicing for genome annotation. Finally, we executed MAKER incorporating the transcriptome, genome, parametric model of species, and the protein sequences of 10 Lepidoptera insects (*Actias luna, Antheraea pernyi, Antheraea yamamai, Bombyx mori, Cnaphalocrocis medinalis, Heliccoverpa armigera, Plutella xylostella, Spodoptera frugiperda, Spodoptera litura, Samia ricini*) with good annotations down from InsectBase 2.0 (http://v2.insect-genome.com) as input data to predict genes (Mei et al. 2022).

2.8 Comparative genomics analysis

OrthoFinder v2.5.1 (Emms & Kelly 2015, 2019) was used to analyze the orthologous and paralogous genes of 10 insect genomes, including *D. melanogaster* (assembly accession: GCF_000001215.4), *P. xylostella* (assembly accession: GCA_019096205.1), *Danaus plexippus* (assembly accession: GCF_009731565.1), *A. yamamai* (Kim et al. 2018), *B. mori* (assembly accession: GCF_014905235.1), *A. pernyi* (assembly accession: GCA_015888305.1), *C. japonica* (in this study), *S. ricini* (assembly accession: GCA_014132275.1), *S. frugiperda* (assembly accession: GCF_011064685.1), *H. armigera* (assembly accession: GCF_002156985.1). And *D. melanogaster* was selected as an outgroup.

2.9 Phylogenetic analysis

1189 single-copy Orthologues shared by 10 insects were used for phylogenetic analysis. Supergenes formed by multiple alignments of single-copy gene families were constructed for tree construction. The phylogenetic tree was constructed by maximum likelihood (ML) using IQ-TREE v2.1.2 with the best model (JTT + F + R5) and 1000 rapid bootstrap replicates to assess the robustness of the tree (Nguyen et al. 2015). Additionally, we used Astral-III (Zhang et al. 2017) to merge all gene trees obtained through OrthoFinder into a unified species tree. It is essential to emphasize that the two trees generated from these methods must be congruent, validating the consistency and accuracy of our phylogenetic analysis. Divergence time was estimated by MCMCtree (Puttick 2019) program in the PAML package v4.8 based on the multiple sequence alignment protein sequences. The calibration time points of *P. xylostella* (286 MYA), *D. plexippus* (179 MYA), *S. frugiperda* (140 MYA), *B. mori* (108MYA), and *H. armigera* (56 MYA) were obtained from TimeTree (http://timetree.org/) (Kumar et al. 2017). Gene family contraction and expansion were analyzed using CAFE v4.2, incorporating the results from OrthoFinder and the phylogenetic tree with divergence time information (Han et al. 2013). Finally, Evolview (https://www.evolgenius.info/evolview/#/) was used to visualize and enhance the appearance of the phylogenetic tree.

2.10 Comparative analysis of detoxification and chemosensory gene families

To identify cytochrome P450 (P450s), glutathione S-transferases (GSTs), ATP-binding cassette transporters (ABCs), odorant binding proteins (OBPs), olfactory receptors (ORs), gustatory receptors (GRs), ionotropic receptors (IRs), chemosensory protein (CSPs) and sensory neuron membrane proteins (SNMPs) protein sequences of well-annotated insects download from UniProt were used as queries to search against the protein sequences of *C. japonica* and other 9 insects by BLASTP (e-value = $1e^{-5}$). After obtaining predicted proteins, we performed additional validation using Hmmsearch to check for the presence of specific protein domains from Pfam, P450s: PF00067, GSTs: PF00043 or PF02798, ABCs: PF00005, OBPs: PF01395, ORs: PF02949 or PF13853, GRs: PF08395 or PF06151, IRs: PF00060, CSPs: PF03392, SNMPs: PF01130. Phylogenetic analysis of the detoxification and chemosensory genes was

performed respectively using maximum likelihood methods using VT + R7, LG + R3, PMB + F + R5, WAG + F + R5, VT + F + R4, VT + F + G4, VT + R4, LG + G4, WAG + F + R5 model determined by ModelFinder (Kalyaanamoorthy et al. 2017) in IQ-TREE. Statistical support for all phylogenetic trees was assessed by Ultrafast bootstrap (Hoang et al. 2018) analysis using 1000 repeats.

3 Results

3.1 Genome assembly

The genome size of *C. japonica* was estimated to be 504 Mb using flow cytometry, and this estimation was further corroborated by K-mer analysis (K = 35) using Illumina short reads. The heterozygosity of the genome was found to be low, at 0.47%. With 169.37 Gb PacBio long reads, we successfully assembled a high-quality genome spanning 584,506,556 bp, with Contig N50 of 12 Mb. This assembly resulted in a total of 771 contigs, with the longest contig reaching a length of 24,253,087 bp. BUSCO analysis with Lepidoptera_ODB10 revealed that the gene space is 95.3% complete genes, suggesting the assembled genome is highly quality and suitable for further analysis. The percentage of Ns in the assembled genome was 1.12% (Table 1). Using Hi-C scaffolding, we anchored a remarkable 97.15% of the assembled genome sequences to 31 chromosomes, resulting in a highly contiguous genome assembly with 584,537,256 bp. The Scaffold N50 reached 20,239,873 bp, indicating the substantial size and continuity of the scaffolds. (Fig. 2 & Fig. 3).

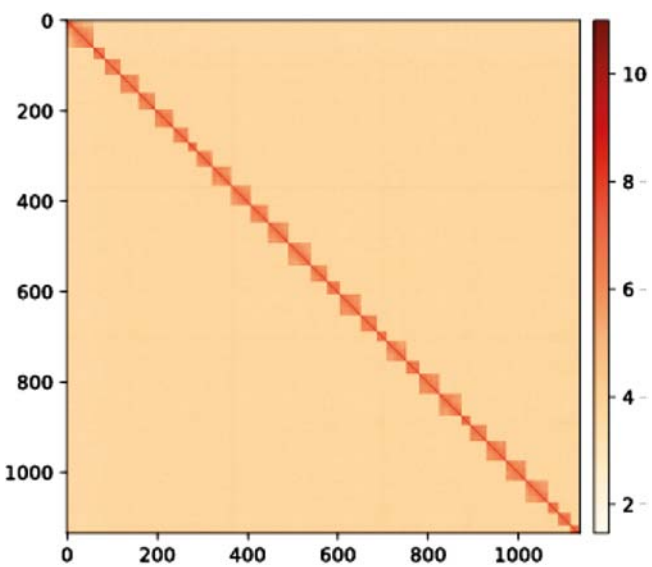

Fig. 2. The Chromosome Hi-C interaction map of *Caligula japonica* identified 31 linkage groups.

3.2 Genome annotation

A total of 345839401 bp repeat sequences were identified, accounting for 59.16% of the *C. japonica* genome (Table 1), which were found to be comparable to the close relatives *A. pernyi* (60.74%) but higher than *A. yamamai* (37.33%) and *S. ricini* (34.3%). The unclassified, long interspersed nuclear elements (LINEs) and DNA transposons were found to be the most abundant transposable elements (TEs) (Table 2). Through de novo gene prediction and homologous gene search, we identified a total of 20,887 protein-coding genes in the *C. japonica* genome, significantly surpassing the gene count in *A. yamamai* (14638), and similar to *A. pernyi* (20814) and *S. ricini* (20366) (Table 1).

Table 1. Summary of chromosome-level assembly for *Caligula japonica*.

Genome size (bp)	584,537,256
No. of chromosome	31
No. of contig	771
Chromosome-level Contig N50 (bp)	12,646,762
Chromosome-level Scaffold N50 (bp)	20,239,873
BUSCO genes (%)	
complete BUSCOs	95.3
complete and single-copy BUSCOs	94.8
complete and duplicated BUSCOs	0.5
fragmented BUSCOs	0.5
missing BUSCOs	4.2
Ns (%)	1.12
Heterozygosity (%)	0.47
Repeat (%)	59.16
G + C (%)	35.13
No. of genes	20,887

Table 2. Statistics of repeat elements of *Caligula japonica*.

Repeat types	Nb. elements	Length (bp)
SINEs	87098	14241746
LINEs	255941	71656528
LTR elements	97265	50439041
DNA transposons	171932	47731449
Unclassified	573747	90591293
Small RNA	29467	4582552
Simple repeats	92423	4067279
Low complexity	14218	661737
Bases masked		345839401

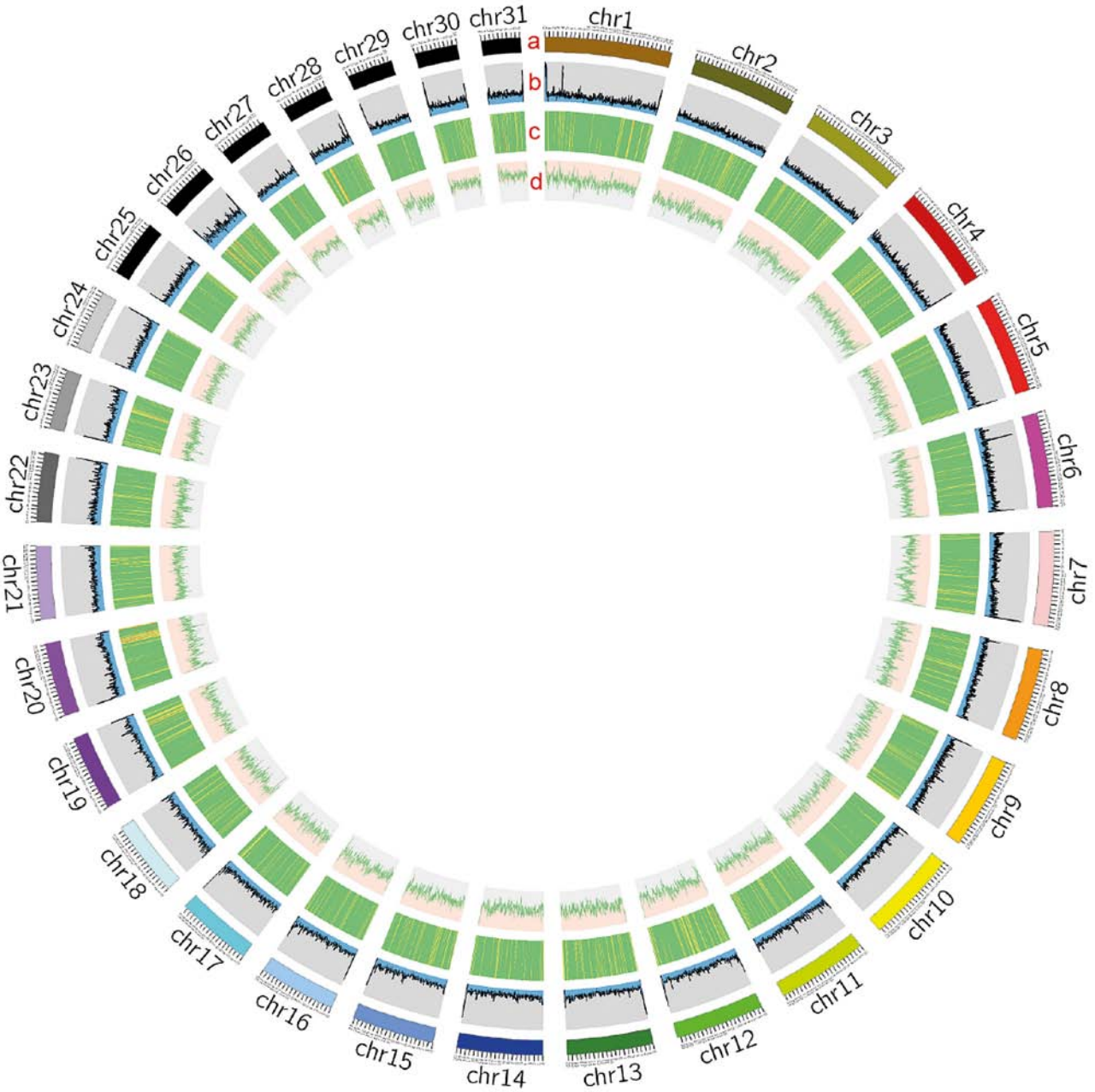

Fig. 3. Circos graph of characteristics of *Caligula japonica* genome. **a:** karyotype, **b:** count of the gene, **c:** repeats density and **d:** GC density in the genome of *C. japonica*.

3.3 Phylogenetic analysis

To infer the evolutionary status and trace the phylogenetic placement of *C. japonica*, we conducted phylogenetic analysis for 10 insects (*D. melanogaster, P. xylostella, D. plexippus, A. yamamai, B. mori, A. pernyi, C. japonica, S. ricini, S. frugiperda*, and *H. armigera*) (Fig. 4). A total of 173167 genes were assigned by OrthoFinder, and 157460 genes were found in orthogroups (90.0%) shared across the 10 insects, consistent with our findings. The analysis resulted in a total of 17,526 orthogroups, with 4,438 orthogroups (25.32%) being shared among all ten tested insects. For phylogenetic analysis, we used 1189 single-copy genes identified by OrthoFinder, and the resulting phylogenetic tree indicated that *C. japonica* diverged from Antheraea about 42.7 million years ago. Further examination revealed 906 genes showing expansion and 417 genes showing contraction in *C. japonica* genome.

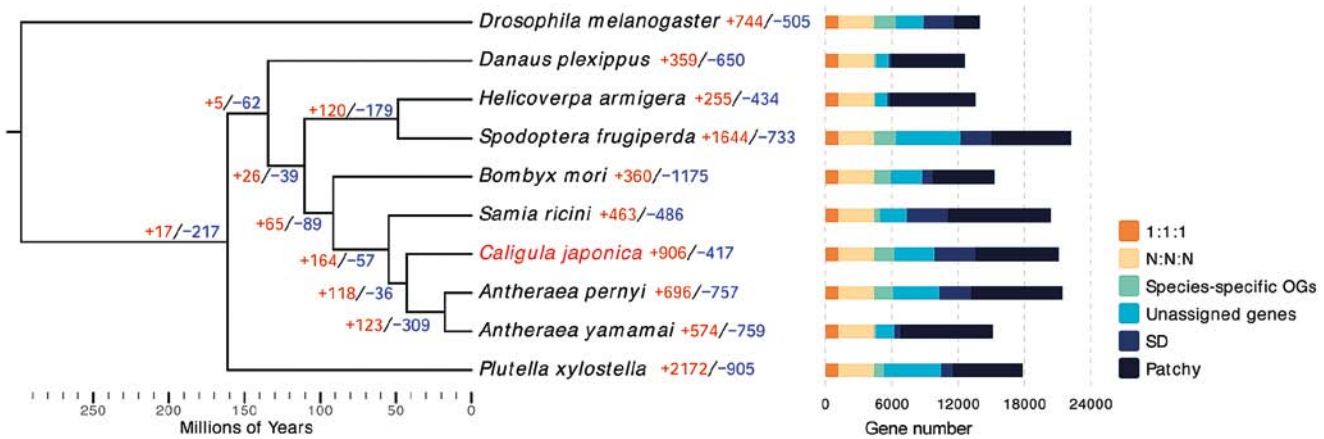

Fig. 4. Phylogenetic tree and orthologs between genomes of *Caligula japonica* and 9 other insects. The maximum likelihood phylogenomic tree was calculated based on 1167 single-copy genes. ("1:1:1": single-copy universal genes; "N:N:N": other multiply genes; "Species-species OGs": genes without in any other species; "Unassigned genes": genes not assigned to any homologous group.; "SD": specific duplicated genes; "Patchy": orthologous in all other species). The red and blue numbers on the branches and nodes respectively represent the contraction and expansion of the gene family during the evolutionary.

3.4 Identification and phylogenetic analysis of detoxification and chemosensory gene families in *C. japonica*

A total of 112 P450s, 39 GSTs, and 111 ABCs were manually identified in the *C. japonica* genome (Table S2). Notably, *C. japonica* shares a similar number of GST genes with three closely related species: *S. ricini, A. pernyi*, and *A. yamamai*. However, the number of *Abc* genes is twice that of *S. ricini* (51) and closely doubled compared to *A. yamamai* (62), making it the second highest among the 10 insects analyzed. The number of ABC genes in *C. japonica* is at an intermediate level. The chemosensory system plays a vital role in various behaviors of insects, such as locating the oviposition site, mates, shelter, and food. In this study, we identified 32 OBPs, 82 ORs, 121 GRs, 23 IRs, 23 CSPs, and 18 SNMPs protein sequences in *C. japonica* genome (Table S2). The number of chemosensory genes is comparable to other lepidoptera insects. However, the number of GR in *C. japonica* is more than double that of *S. ricini*, and *C. japonica* has the lowest number of IR, ranking below *A. yamamai* and *P. xylostella*.

Through rigorous bioinformatic analysis, a comprehensive phylogenetic tree was generated, elucidating the evolutionary relationships within the detoxification and chemosensory gene families annotated in the genomes of *C. japonica* and *B. mori*. Comparative genomic insights revealed a pronounced expansion of these gene families in *C. japonica* (Table S2) (Fig. S1). Particularly, the gene families P450 (Fig. S1a), ABC (Fig. S1c), and GR (Fig. S1f) manifested a significant evolutionary expansion in contrast to their counterparts in *B. mori*.

4 Discussion

C. japonica, while commonly identified as a forestry pest, has garnered attention due to its potential medicinal attributes (Chen et al. 2022). It becomes conceivable to reposition this forestry pest as a lucrative asset. However, establishing artificial rearing programs for this insect is complex, and there is currently no documented experience or report on rearing it. The physiological traits of this insect remain unclear and are susceptible to biotic or abiotic stress. Several parasitic wasp families, including the Eulophid and Trichogrammatid, are known to parasitize the eggs of *C. japonica*, potentially posing challenges for its artificial rearing (Chen et al. 2021). Of course, *C. japonica* also has the potential to serve as an alternative host for the industrial-scale breeding of these parasitic wasps, like *A. pernyi* (Chen et al. 2021) (Hassan et al. 2004). Importantly, genetic research and molecular breeding on *C. japonica* are currently hampered by the lack of a high-quality genome resource.

Detailed insights into its genetic composition might not only spur innovative management strategies but also foster widescale cultivation techniques, potentially turning *C. japonica* into a significant economic commodity. In this study, a robust, chromosome-level genome of *C. japonica* was generated using an integrated approach that harnessed the capabilities of Illumina NovaSeq, PacBio Sequel II, and Hi-C methodologies, revealing a genome comprising 31 chromosomes and spanning roughly 584.54 MB. This chromosome-level genome resource of *C. japonica* thus provides an important resource for key trait mapping through genome-wide association studies for the identification of

candidate genes underlying disease resistance and environmental adaption, which can be used in molecular breeding. *C. japonica* demonstrates remarkable sensitivity to environmental changes and employs two distinct types of diapause. This will also increase the time and cost associated with artificial breeding. The first type is known as egg diapause, which occurs during the overwintering period. The pupal period typically spans around 40 days, during which the insects enter the second type of diapause, known as summer diapause (Chen et al. 2021). At the molecular level, the regulation mechanism related to diapause *C. japonica* has remained unknown. In studies on other insects, research related to diapause has involved the functionality of genes associated with the circadian clock (Williams & Adkisson 1964; Xu et al. 2003; Ikeno et al. 2010, 2011a, 2011b; Cao et al. 2021). This will also be a future research direction for *C. japonica*, with the hope of regulating diapause and improving genetic resources. A high-quality genome serves as a fundamental and indispensable tool for advancing research on the regulation of diapause and harnessing the potential medical value of *C. japonica*.

Gene expression dynamics related to diapause are largely influenced by environmental stimuli. Direct interfaces with the environment, such as sensory modalities, play crucial roles in modulating these genetic responses. Comparative genomic analyses reveal that *C. Japonica* exhibits expansions in the P450s, ABCs, and GRs gene families when juxtaposed against the silkworm and other Lepidopteran insects (Fig. S1) (Table S2). Integrating this with phylogenetic information underscores that *C. japonica* has augmented its repertoire of metabolic and allelopathic genes, likely to exploit a broader range of host plants as nutritional sources. Such evolutionary tendencies could be consequential to broader environmental shifts, inclusive of contemporary climate change patterns and the intensifying deployment of chemical agents. Given these observations, it becomes evident that future research should adopt ecologically compatible prevention and control strategies, such as using some natural enemy insects for control (Chen et al. 2019).

In summary, we first reported a high-quality chromosome-scale genome of *C. japonica* and discussed the evolution of diapause-related genes. Through rigorous phylogenetic analyses, genes pertinent to metabolism and chemical perception were annotated and scrutinized. This foundational work establishes a valuable theoretical basis for the preventive strategies and resource reutilization of *C. Japonica*, as well as advances our understanding in Lepidopteran species.

Acknowledgments: This work was supported by the National Key R&D program of China (2023YFE0104800), the Natural Science Research Program of Guizhou University (202202) and the Foundation of Postgraduate of Guizhou Province (YJSKYJJ [2021]041).

Data accessibility: The initial reads produced (Hi-C reads, Illumina reads and the PacBio reads) in this study have been deposited in NCBI under bioproject ID: PRJNA814848. The genome sequence assembly and annotation file are available in the FigShare database: https://doi.org/10.6084/m9.figshare.24156021.v1.

References

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research, 27*(2), 573–580. https://doi.org/10.1093/nar/27.2.573

Bruna, T., Hoff, K. J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021). BRAKER2: Automatic Eukaryotic Genome Annotation with GeneMark-EP+ and AUGUSTUS Supported by a Protein Database. *NAR Genomics and Bioinformatics, 3*(1), 108. https://doi.org/10.1093/nargab/lqaa108

Cantarel, B. L., Korf, I., Robb, S., Parra, G., Ross, E., Moore, B., … Yandell, M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research, 18*(1), 188–196. https://doi.org/10.1101/gr.6743907

Cao, L. J., Song, W., Yue, L., Guo, S. K., Chen, J. C., Gong, Y. J., … Wei, S. J. (2021). Chromosome-level genome of the peach fruit moth *Carposina sasakii* (Lepidoptera: Carposinidae) provides a resource for evolutionary studies on moths. *Molecular Ecology Resources, 21*(3), 834–848. https://doi.org/10.1111/1755-0998.13288

Chen, M., Mei, Y., Chen, X., Chen, X., Xiao, D., He, K., … Li, F. (2021a). A chromosome-level assembly of the harlequin ladybird *Harmonia axyridis* as a genomic resource to study beetle and invasion biology. *Molecular Ecology Resources, 21*(4), 1318–1332. https://doi.org/10.1111/1755-0998.13342

Chen, Y. M., Gibson, G A P., Peng, L. F., Iqbal, A., & Zang, L. S. (2019). *Anastatus Motschulsky* (Hymenoptera, Eupelmidae): Egg parasitoids of *Caligula japonica* Moore (Lepidoptera, Saturniidae) in China. *ZooKeys, 881*, 109–134. https://doi.org/10.3897/zookeys.881.34646

Chen, Y. M., Pekdemir, S., Bilican, I., Koc-Bilican, B., Cakmak, B., Ali, A., … Kaya, M. (2021b). Production of natural chitin film from pupal shell of moth: Fabrication of plasmonic surfaces for SERS-based sensing applications. *Carbohydrate Polymers, 262*, 117909. https://doi.org/10.1016/j.carbpol.2021.117909

Chen, Y. M., Qu, X. R., Li, T. H., Iqbal, A., Wang, X., Ren, Z. Y., … Zang, L. S. (2021c). Performances of six eupelmid egg parasitoids from China on Japanese giant silkworm *Caligula japonica* with different host age regimes. *Journal of Pest Science, 94*(2), 309–319. https://doi.org/10.1007/s10340-020-01271-1

Chen, Y. M., Zang, L. S., Koc-Bilican, B., Bilican, I., Holland, C., Cansaran-Duman, D., … Kaya, M. (2022). Macroporous Surgical Mesh from a Natural Cocoon Composite. *ACS Sustainable Chemistry & Engineering, 10*(18), 5728–5738. https://doi.org/10.1021/acssuschemeng.1c06941

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., … Aiden, E. L. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science, 356*(6333), 92–95. https://doi.org/10.1126/science.aal3327

Emms, D. M., & Kelly, S. (2015). OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves

orthogroup inference accuracy. *Genome Biology, 16*(1), 157. https://doi.org/10.1186/s13059-015-0721-2

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology, 20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology, 29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols, 8*(8), 1494–1512. https://doi.org/10.1038/nprot.2013.084

Han, M. V., Thomas, G. W., Lugo-Martinez, J., & Hahn, M. W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular Biology and Evolution, 30*(8), 1987–1997. https://doi.org/10.1093/molbev/mst100

Hassan, S. A., Liscsinszky, H., & Zhang, G. (2004). The oak-silkworm egg *Antheraea pernyi* (Lepidoptera: Anthelidae) as a mass rearing host for parasitoids of the genus *Trichogramma* (Hymenoptera: Trichogrammatidae). *Biocontrol Science and Technology, 14*(3), 269–279. https://doi.org/10.1080/09583150410001665150

He, K., Lin, K., Wang, G., & Li, F. (2016). Genome sizes of nine insect species determined by flow cytometry and k-mer analysis. *Frontiers in Physiology, 7*, 569. https://doi.org/10.3389/fphys.2016.00569

Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution, 35*(2), 518–522. https://doi.org/10.1093/molbev/msx281

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., & Stanke, M. (2016). BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics (Oxford, England), 32*(5), 767–769. https://doi.org/10.1093/bioinformatics/btv661

Hoff, K. J., 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

Ikeno, T., Tanaka, S. I., Numata, H., & Goto, S. G. (2010). Photoperiodic diapause under the control of circadian clock genes in an insect. *BMC Biology, 8*(1), 116. https://doi.org/10.1186/1741-7007-8-116

Ikeno, T., Katagiri, C., Numata, H., & Goto, S. G. (2011a). Causal involvement of mammalian-type cryptochrome in the circadian cuticle deposition rhythm in the bean bug *Riptortus pedestris. Insect Molecular Biology, 20*(3), 409–415. https://doi.org/10.1111/j.1365-2583.2011.01075.x

Ikeno, T., Numata, H., & Goto, S. G. (2011b). Photoperiodic response requires mammalian-type cryptochrome in the bean bug Riptortus pedestris. *Biochemical and Biophysical Research Communications, 410*(3), 394–397. https://doi.org/10.1016/j.bbrc.2011.05.142

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., Von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods, 14*(6), 587–589. https://doi.org/10.1038/nmeth.4285

Kawaguchi, Y., Ichida, M., Kusakabe, T., & Koga, K. (2003). Chorion architecture in the Japanese giant silkmoth, *Caligula japonica* Moore. *Sericologia, 43*(1), 29–39.

Kim, S. R., Kwak, W., Kim, H., Caetano-Anolles, K., Kim, K. Y., Kim, S. B., … Park, S.-W. (2018). Genome sequence of the Japanese oak silk moth, *Antheraea yamamai*: The first draft genome in the family Saturniidae. *GigaScience, 7*(1), gix113. https://doi.org/10.1093/gigascience/gix113

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research, 27*(5), 722–736. https://doi.org/10.1101/gr.215087.116

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution, 34*(7), 39–78. https://doi.org/10.1093/molbev/msx116

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., & Li, H. (2009a). The sequence alignment/map format and SAMtools. *Bioinformatics (Oxford, England), 25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Y. P., Yang, B. S., Wang, H., Xia, R. X., Wang, L., Zhang, Z. H., … Liu, Y. Q. (2009b). Mitochondrial DNA analysis reveals a low nucleotide diversity of *Caligula japonica* in China. *African Journal of Biotechnology, 8*(12).

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN]. https://doi.org/10.48550/arXiv.1303.3997

Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO: Assessing genomic data quality and beyond. *Current Protocols, 1*(12), e323. https://doi.org/10.1002/cpz1.323

Mei, Y., Jing, D., Tang, S., Chen, X., Chen, H., Duanmu, H., … Li, F. (2022). InsectBase 2.0: A comprehensive gene resource for insects. *Nucleic Acids Research, 50*(D1), D1040–D1045. https://doi.org/10.1093/nar/gkab1090

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution, 32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Puttick, M. N. (2019). MCMCtreeR: Functions to prepare MCMCtree analyses and visualize posterior ages on trees. *Bioinformatics (Oxford, England), 35*(24), 5321–5322. https://doi.org/10.1093/bioinformatics/btz554

Qiao, X., Wang, Y. C., Wu, G., Wang, S. Y., Hu, J. Z., Liu, T. X., & Feng, S. Q. (2014). Occurrence reasons and control measures of *Caligula japonica* in Longnan city of China. *Plant Diseases and Pests, 5*, 38–41.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England), 26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ruan, J., & Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nature Methods, 17*(2), 155–158. https://doi.org/10.1038/s41592-019-0669-3

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*(1), 62. https://doi.org/10.1186/1471-2105-7-62

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve

de novo gene finding. *Bioinformatics (Oxford, England), 24*(5), 637–644. https://doi.org/10.1093/bioinformatics/btn013

Williams, C. M., & Adkisson, P. L. (1964). Physiology of insect diapause. XIV. An endocrine mechanism for the photoperiodic control of pupal diapause in the oak silkworm, *Antheraea pernyi*. *The Biological Bulletin, 127*(3), 511–525. https://doi.org/10.2307/1539252

Xu, W. H., & Denlinger, D. L. (2003). Molecular characterization of prothoracicotropic hormone and diapause hormone in *Heliothis virescens* during diapause, and a new role for diapause hormone. *Insect Molecular Biology, 12*(5), 509–516. https://doi.org/10.1046/j.1365-2583.2003.00437.x

Yang, B. S., Li, J., Li, Y. R., & Wang, Z. (2008). Genetic diversity assessment of *Dictyoploca japonica* from different areas. *Chinese Bulletin of Entomology, 45*, 418–421.

Zang, Z. Y., Chen, Y. M., Xu, W., & Zang, L. S. (2023). Evidence of two-sex life table analysis supporting *Anastatus japonicus*, a more effective biological control agent of *Caligula japonica* compared with other two *Anastatus* species. *Biological Control, 180*, 105188. https://doi.org/10.1016/j.biocontrol.2023.105188

Zhang C, Sayyari E, & Mirarab S. (2017): ASTRAL-III: increased scalability and impacts of contracting low support branches. In RECOMB international workshop on comparative genomics, 53–75. https://doi.org/10.1007/978-3-319-67979-2_4

Manuscript received: February 15, 2023
Revisions requested: June 5, 2023
Revised version received: October 9, 2023
Manuscript accepted: October 10, 2023

The pdf version (Adobe JavaScript must be enabled) of this paper includes an electronic supplement:
Table S1, S2, Figure S1, The code involved in bioinformatics analysis in this research