scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A chromosomal-level genome assembly of *Trichogramma chilonis* Ishii, 1941 (Hymenoptera: Trichogrammatidae)

Chengxing Wang[1,2,3,4], Zhenjuan Yin[1,5], Yan Liu[1,2,3,4], Xiaoyan Dai[1,2,3,4], Shan Zhao[1,2,3,4], Ruijuan Wang[1,2,3,4], Yu Wang[1,2,3,4], Long Su[1,2,3,4], Hao Chen[1,2,3,4], Li Zheng[1,2,3,4] & Yifan Zhai[1,2,3,4] ✉

Trichogramma **spp. is a genus of minute egg parasitoids frequently used in agricultural pest management that can feed on the eggs of various lepidopteran pests. Currently, there is a scarcity of high-quality genomic resources for this category of tiny parasitoids, which impedes our comprehension of the population evolution and parasitic ecology of this collective. In this case, a chromosome-level genome of** ***Trichogramma chilonis*** **was produced by integrating PacBio HiFi, Illumina, and Hi-C data. The genome size totals 202.48 Mb, with a scaffold N50 length of 40.00 Mb. A total of 98.59% (199.63 Mb) of contigs were effectively mapped onto five chromosomes. The BUSCO assessment revealed that the genome assembly achieved 98.1% (n = 1,367) completeness, with 95% representing single-copy BUSCOs and 3.1% duplicated BUSCOs. Also, the genome comprises 24.16% (48.91 Mb) repeat elements and 12,163 predicted protein-coding genes. The high-quality genome of** ***T. chilonis*** **presented in this study offers an invaluable asset for elucidating its evolutionary path and ecological interactions.**

Background & Summary

Globally, there exist over 600 recognized members of the Trichogrammatidae clan, with approximately 210 species of *Trichogramma* spp. capable of parasitizing the eggs of diverse crop pests, rendering them the most extensively employed taxon of natural enemies in a range of biological control initiatives[1,2]. The *Trichogramma* wasp deposits their eggs inside the eggs of the host pest, and their larvae ingest the nutrients within the eggs, destroying the growth and development of the embryos, thus adequately controlling the pest population[3,4]. The application of *Trichogramma* wasps substantially decreases dependence on chemical pesticides and alleviates the adverse effects of agriculture on the environment and human health.

Trichogramma chilonis serves a crucial role in biological control as one of the main *Trichogramma* species systematically bred and released[5]. Despite its minute size, measuring only 0.5–1 mm[6], this species exhibits a broad host range and can manage diverse agricultural pests, including the *Cnaphalocrocis medinalis*, *Spodoptera frugiperda*, *Tuta absoluta*, *Chilo infuscatellus*, *Chilo suppressalis*, *Helicoverpa armigera*, and others[7–12]. Therefore, *T. chilonis* is extensively employed in pest management for crops like rice, corn, cotton, sugarcane, and vegetables[13].

The swift advancement of genome sequencing technology in recent years has underscored the importance of utilizing genomes to unveil the biological traits and parasitic mechanisms of *T. chilonis*, essential for producing the efficacy of biological control strategies[14]. However, genome data has been disclosed for only three *Trichogramma* wasps[14], with none reported for *T. chilonis*. The inadequate availability of genetic resources has impeded our comprehension of its biological traits and parasitic mechanisms.

[1]Institute of Plant Protection, Shandong Academy of Agricultural Sciences, Jinan, 250100, China. [2]Shandong Key Laboratory for Green Prevention and Control of Agricultural Pests, Jinan, 250100, China. [3]Key Laboratory of Natural Enemies Insects, Ministry of Agriculture and Rural Affairs, Jinan, 250100, China. [4]Shandong Engineering Research Center of Resource Insects, Jinan, 250100, China. [5]College of Agriculture, Guizhou University, Guiyang, 550025, China. ✉e-mail: saaszyifan@163.com

Libraries	Insert sizes (bp)	High-quality data (Gb)	Sequencing coverage (x)
Illumina	350	11.82	58.38
PacBio HiFi	20 Kb	19.29	95.27
Hi-C	350	18.55	91.63
RNA	350	11.43	—

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

Here, we constructed the chromosome-level genome of *T. chilonis* by integrating PacBio HiFi, Illumina, and Hi-C data, which annotated the repeats, non-coding RNAs, and protein-coding genes. This high-quality genome establishes a valuable dataset for delving deeper into the evolution and essential traits of Trichogrammatidae species.

Methods

Insect preparation. The *T. chilonis* utilized in this research was obtained from the Natural Enemies and Pollinator Breeding Center at the Institute of Plant Protection, Shandong Academy of Agricultural Sciences (Jinan, China). The internal transcribed spacer 2 (ITS2) sequence was amplified and performed for species identification at the molecular level[15]. The findings indicated that the *ITS2* sequence amplified from the randomly chosen *Trichogramma* individuals (n = 8) exhibited the highest similarity with the *ITS2* sequence (Accession number: FN665797.1) of *T. chilonis* in the NCBI database, ranging from 99.57% to 99.78% (Fig. S1). To minimize heterozygosity, an isofemale line was established through the backcrossing of male offspring with a virgin female. *Corcyra cephalonica* eggs served as the host for *T. chilonis*, and following a minimum of six generations of cultivation, substantial populations of female adults were gathered.

DNA and RNA sequencing. The TaKaRa MiniBEST Universal Genomic DNA Extraction Kit Ver. 5.0 (TaKaRa, Tokyo, Japan) was utilized to extract genomic DNA from 1000 female adults. For constructing the PacBio HiFi 20Kb library (insert size), the SMRTbell® Express Template Prep Kit 2.0 (Pacific Biosciences) was employed, and sequencing was conducted in HiFi mode on the PacBio Sequel II platform. Illumina whole genome sequencing (WGS) short-read libraries were created using the TruSeq DNA PCR-free kit, comprising 150 bp paired-end reads and 350 bp insert size. High-throughput chromosome conformation capture (Hi-C) was carried out following established protocol[16], involving cross-linking, restriction enzyme digestion, end repair, DNA cyclization, and purification. All short-read libraries were sequenced on the illumina NovaSeq 6000 platform.

RNA was isolated from 200 female adults employing Trizol (Invitrogen, California, CA, USA) as per the manufacturer's instructions. Libraries were prepared with the TruSeq RNA Library Prep Kit v2 (Illumina, California, CA, USA) and then underwent Illumina sequencing. Notably, our genome sequencing yielded 19.29 Gb (95.27×) of PacBio long-reads, 11.82 Gb (58.38×) of Illumina short-reads, 18.55 Gb (91.63×) of Hi-C data, and 11.43 Gb of transcriptome data (Table 1).

Genome survey. To ensure the reliability of the genomic data, the Illumina raw data underwent quality control using BBTools v38.82[17]. The clumpify.sh script was employed to eliminate duplicate sequences, while the bbduk.sh script was utilized for specific quality control measures such as removing sites with base quality values below 20, sequences shorter than 15 bp, and poly-A/G/C endings longer than 10 bp. GenomeScope v2.0[18] was used for k-mer analysis, setting the maximum k-mer coverage cutoff at 1,000. The k-mer frequencies were assessed using klist.sh (BBTools) with a length of 21 k-mers. By analyzing the coverage and frequency distribution of k-mers, the estimated genome size of *T. chilonis* was determined to be approximately 203.22 Mb with a heterozygosity rate of 0.261% (Fig. 1).

Genome assembly. The high-quality PacBio HiFi long-reads underwent primary assembly using Hifiasm v0.16.1[19]. Contigs with a sequencing depth greater than 10 × were selectively retained to mitigate the inclusion of potentially contaminated or erroneous low-depth sequences. Subsequently, the alignment tool Minimap2 (v2.23)[20] was executed using Purge_dups (v1.2.5)[21] to remove haplotigs, with a haploid cutoff of 70 ('-s 70') applied[20] to identify contigs as haplotigs. Juicer v1.6.2[22] was utilized to align Hi-C reads with refined genomes, and the 3D-DNA v180922[23] procedure was employed to anchor contigs for chromosome assembly with default settings. Assembly errors were manually reviewed and rectified using Juicebox v1.11.08[22]. To identify potential bacterial and human contaminants in the assembly outcomes, a BLASTN-like search was conducted using MMseq. 2 v13[24] against the NCBI Nucleotide database. Contaminants from vectors were examined against the UniVec database employing BLAST + v2.11.0[25] with the VecScreen parameters. The chromosome-level genome of *T. chilonis* was assembled to a size of 202.48 Mb, consisting of 31 scaffolds and 63 contigs. The longest scaffold and contig measured 49.97 Mb and 24.21 Mb, respectively (Table 2). It featured a scaffold N50 length of 40.00 Mb and a contig N50 size of 10.27 Mb, and the GC content was 39.84% (Table 2). A sum of 199.63 Mb contigs was effectively arranged onto five chromosomes, achieving a mounting rate of 98.59% (Fig. 2), and chromosome sizes range from 33.83 to 49.97 Mb (Fig. 3). The count of haploid chromosomes assembled by us aligns with those documented for *Trichogramma brassicae*[26], *Trichogramma embryophagum*[27], and *Trichogramma pretiosum*[28] previously. Moreover, the BUSCO assessment of genome assembly completeness is demonstrated to be 98.1% (Table 3). All the indicators show that our genome assembly has achieved an exceptionally high level of continuity and integrity.

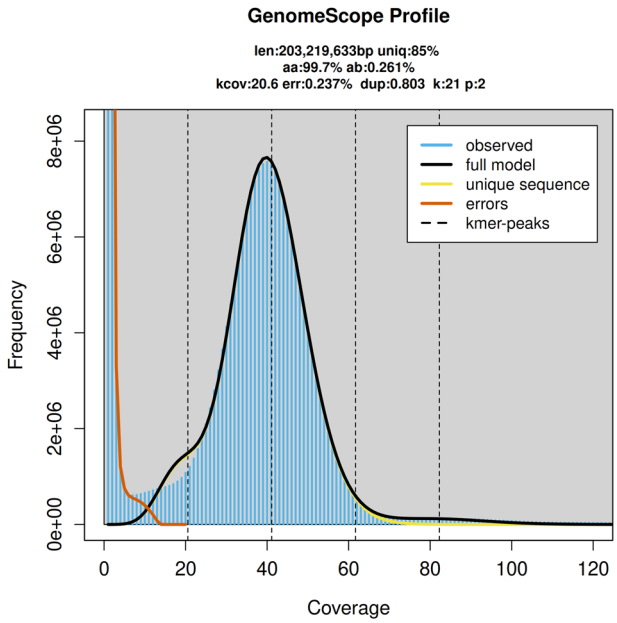

Fig. 1 Genome survey at 21-mer of *T. chilonis* was estimated using GenomeScope. The vertical dotted lines indicate the coverage peaks for heterozygous, homozygous, and duplicated sequences separately.

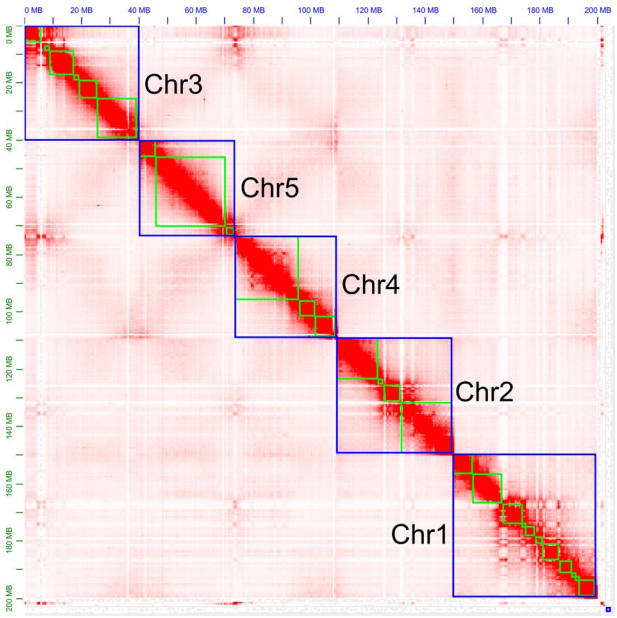

Fig. 2 Contact map. Chromosome-level heat map of *T. chilonis*, with each chromosome outlined in blue.

Assembly	Total length (Mb)	Number scaffolds (chromosomes)	Number contigs (chromosomes)	Scaffold N50 length (Mb)	Contig N50 length (Mb)	GC (%)	BUSCO (n = 1,367) (%)			
							C	D	F	M
Hifiasm	210.77	167	167	10.27	10.27	39.94	98.2	3.7	0.4	1.4
Purge_Dups	205.21	57	66	10.27	10.27	39.86	98.1	3.1	0.4	1.5
3D-DNA	205.22	57(5)	92(5)	40.00	10.27	39.86	98.1	3.1	0.4	1.5
Final	202.48	31(5)	63(5)	40.00	10.27	39.84	98.1	3.1	0.4	1.5

Table 2. Genome assembly statistics for *T. chilonis*. C: complete BUSCOs; D: complete and duplicated BUSCOs; F: fragmented BUSCOs; M: missing BUSCOs.

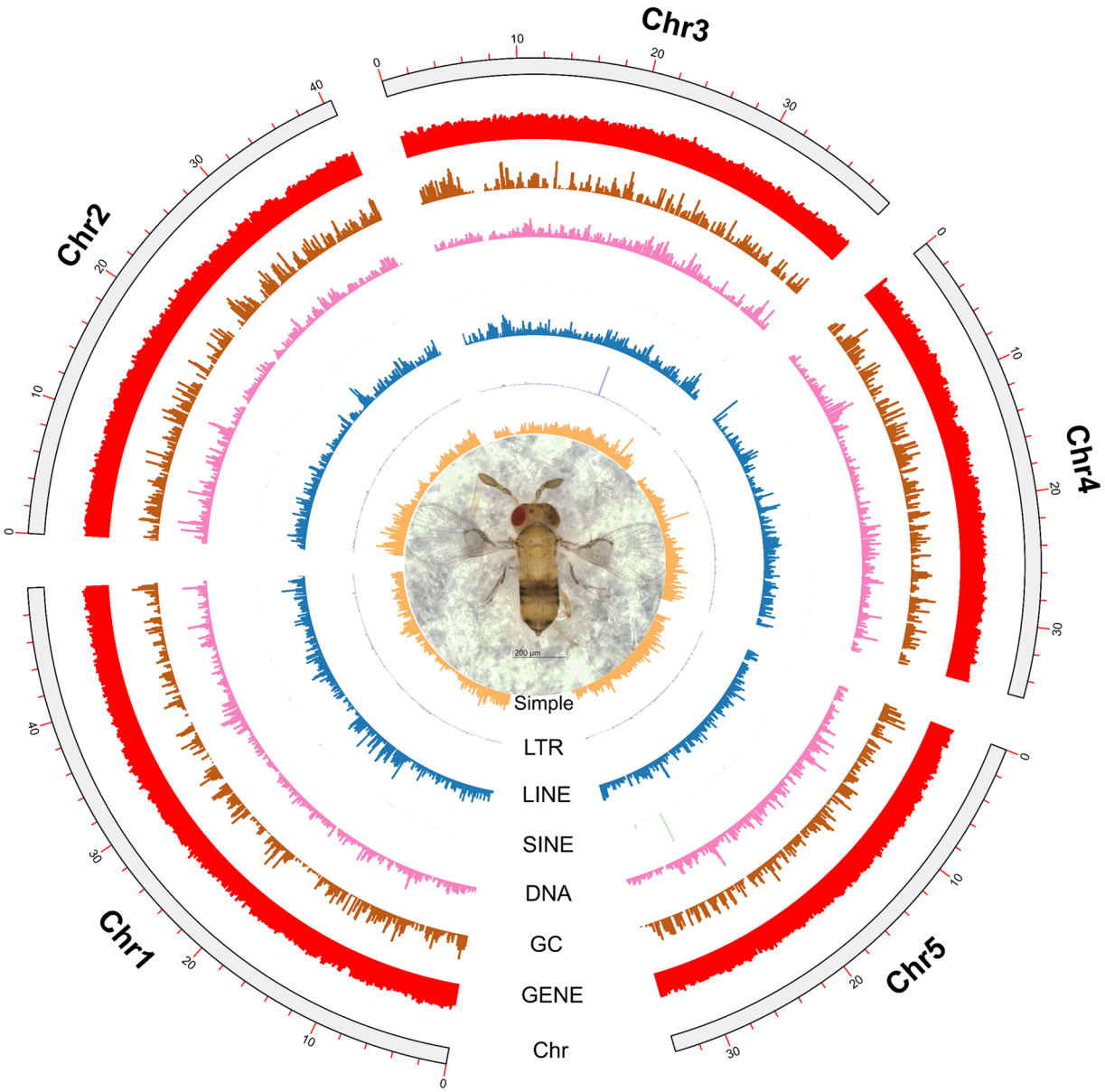

Fig. 3 Genomic features of *T. chilonis* are illustrated in a circular arrangement. Starting from the innermost circle moving outwards, they include simple repeats, long terminal repeats (LTR), long interspersed nuclear elements (LINE) and short interspersed nuclear elements (SINE), DNA transposons density, GC content, gene density (GENE), and chromosome length (Chr).

Genome assembly metrics	*Trichogramma chilonis*	*Trichogramma dendrolimi*
Total length (Mb)	202.48	215.2
Number scaffolds	31	316
Number contigs	63	364
Scaffold N50 length (Mb)	40.00	1.4
Contig N50 length (Mb)	10.27	1.3
GC (%)	39.84	39.8
BUSCO completeness (%)	98.1	93.4

Table 3. Comparative statistics of *Trichogramma chilonis* and *Trichogramma dendrolimi* genome assembly.

Genome annotation. The *de novo* repeat library was created through RepeatModeler v2.0.4[29] along with an extra LTR discovery pipeline (-LTRStruct). Subsequently, this library was combined with the Dfam 3.5[30] and RepBase-20181026[31] databases to produce the ultimate custom database. Following this, RepeatMasker v4.1.4[32]

Characteristics	*T. chilonis*
Genome annotation	
Repetitive elements	
Size (Mb)	48.91 (24.16%)
DNA transposons (Mb)	2.97 (1.43%)
SINEs (Mb)	0.32 (0.15%)
LINEs (Mb)	4.93 (2.42%)
LTRs (Mb)	3.26 (1.60%)
Unclassified (Mb)	32.33 (15.97%)
Protein-coding genes	
Number	12,163
Mean gene length (bp)	6,814.3
BUSCO completeness (%)	97.2
ncRNA	
Number of ncRNA	782
rRNA	250
miRNA	75
snRNA	1

Table 4. Genome annotation statistics of *T. chilonis*.

software was utilized to analyze the conclusive database, leading to the discovery of 319,998 repeat sequences covering 48,914,773 bp, representing 24.16% of the genome. The repetitive elements based on proportion included unclassified (15.97%), simple repeats (1.88%), long-interspersed elements (LINEs, 2.42%), long-terminal repeats (LTRs, 1.60%), DNA transposons (1.43%), and other elements (Fig. 3; Table 4).

We employed the MAKER3 v3.01.03[33] annotation pipeline to forecast protein-coding gene structures, including *ab initio* gene structure prediction, transcript sequence alignment with genomes, and comparison with protein sequences from known homologous species. BRAKER v2.1.6[34] and GeMoMa v1.8[35] were utilized to integrate transcriptome and protein evidence, with the generated files being used for MAKER's *ab initio* prediction. BRAKER automatically trained Augustus v3.3.4[36] and GeneMark-ES/ET/EP 4.68_3.60_lic[37] and cross-referenced the arthropoda protein sequence library (OrthoDB10 v1) to enhance prediction accuracy[38]. Transcript assembly guided by the genome was conducted using StringTie v2.1.6[39], and the outcomes were utilized as mRNA evidence for MAKER. Additionally, transcriptome alignment was carried out using HISAT2 v2.2.0[40], and homologous proteins from five species (*Apis mellifera* (GCF_003254395.2)[41], *Bombyx mori* (GCF_014905235.1)[42], *Nasonia vitripennis* (GCF_009193385.2)[43], *Drosophila melanogaster* (GCF_000001215.4)[44], *Tribolium castaneum* (GCF_000002335.3)[45]) were supplied to aid in the gene prediction process of GeMoMa with the parameters "GeMoMa.c = 0.4 GeMoMa.p = 10". The genome of *T. chilonis* predicted a total of 12,163 protein-coding genes, with an average length of 6,814.3 bp (Table 4). The completeness of the protein-coding gene sequences, assessed by BUSCO, reached 97.2% (Table 4).

Gene function annotation was carried out by comparing the UniProtKB database using the more sensitive mode (conditions "-e 1e-5") in Diamond v2.0.11.1[46]. Also, InterProScan 5.53–87.0[47] and eggNOG-mapper v2.1.5[48] were utilized for functional annotation, encompassing Gene Ontology (GO) terms, signaling pathways (KEGG and Reactome), and the recognition of conserved protein domains. This was accomplished through the comparison of the five databases: Pfam[49], SMART[50], Superfamily[51], CDD[52], and eggNOG v5.0[53]. The annotation outcomes produced by the above tools are merged to derive the ultimate list (Fig. 4) of predicted gene functions. As follows: The UniProtKB database contains entries for 11,515 genes, with 1,142 of them being individually annotated. Through InterProScan, 10,004 protein-coding genes were identified, with only 32 of them not being annotated by other databases. A total of 11,037 genes were annotated by the eggNOG database, with 1,599 of them being individually annotated. Additionally, 6,369 genes were annotated consistently by all databases. Finally, the structural features of the genome (Fig. 3) were visualized by using TBTools (v1.098769)[54].

Two strategies were employed for non-coding RNA (ncRNA) annotation. Initially, Infernal v1.1.4[55] was utilized to compare against the established ncRNA database (Rfam v14.10) for the annotation of ribosomal RNA (rRNA), small nuclear RNA (snRNA), and microRNA[56]. Subsequently, tRNAscan-SE v2.0.9[57] with the script "EukHighConfidenceFilter" was used to predict transfer RNA (tRNA) sequences within the genome. Our genome yielded 782 ncRNAs, comprising 2 long ncRNAs (lncRNAs), 250 rRNAs, 75 microRNAs, 96 snRNAs, and 265 tRNAs.

Data Records

The original sequencing data and genome assembly of *Trichogramma chilonis* have been submitted to the National Center for Biotechnology Information (NCBI). The data includes Hi-C, transcriptome, whole genome sequencing, and PacBio HiFi data, which can be accessed using the identification numbers SRR31510496-SRR31510499[58]. The assembled genome can be found on NCBI under the accession number GCA_045269175.1[59]. Results of the repeat analysis, gene structures, and functional predictions are available on figshare[60].

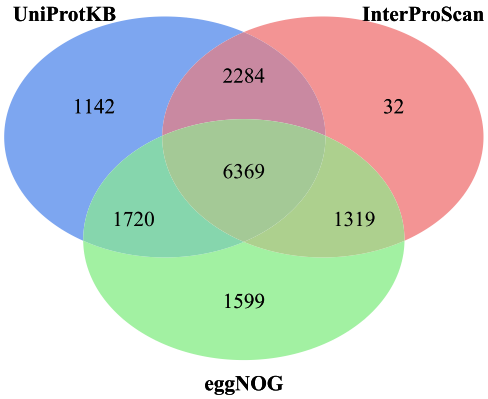

Fig. 4 Overview of the number of genes in the *T. chilonis* genome annotated by different databases.

Technical Validation

Compared to the genome of *T. dendrolimi*, our genome assembly demonstrates enhanced continuity and completeness attributed to the utilization of long reads and Hi-C sequencing. The number of scaffolds/contigs, as well as the N50 scaffold/contig length of *T. chilonis* are much better than those of *T. dendrolimi* (Table 3). Furthermore, the integrity of the genome assembly was evaluated using BUSCO v5.0.4[61]. The completeness assessment revealed 98.1% overall completeness, with 95% for single-copy BUSCOs, 3.1% for duplicated BUSCOs, 0.4% for fragmented BUSCOs, and 1.5% for missing BUSCOs, representing a notable improvement over *T. dendrolimi* in assembly integrity (Table 3). To further assess the assembly's integrity, the mapping rate was determined by aligning PacBio, Illumina, and RNA sequencing reads with the final assembly using Minimap2 and SAMtools. The mapping ratio for PacBio, Illumina, and RNA reads were 99.84%, 98.46%, and 98.03%, respectively. These analyses collectively confirm the high quality of the genome assemblies generated in this study.

Code availability

The scripts employed for genome assembly and annotation in this publication have been shared on figshare[60]. All commands and pipelines utilized in data processing were executed in accordance with the manuals and protocols provided by the respective bioinformatic software.

Received: 10 December 2024; Accepted: 7 March 2025;
Published online: 19 March 2025

References

1. Ashok Kumar, G. *et al.* Internal transcribed spacer-2 restriction fragment length polymorphism (ITS-2-RFLP) tool to differentiate some exotic and indigenous trichogrammatid egg parasitoids in India. *Biol. Control.* **49**, 207–213 (2009).
2. Zang, L., Wang, S., Zhang, F. & Desneux, N. Biological control with *Trichogramma* in China: History, Present Status, and Perspectives. *Annu. Rev. Entomol.* **66**, 463–484 (2021).
3. Woelke, J. B., Bukovinszky, T. & Huigens, M. E. Nocturnal parasitism of moth eggs by *Trichogramma* wasps. *Biocontrol Sci. Technol.* **27**, 769–780 (2017).
4. Wang, Z., He, K., Zhang, F., Lu, X. & Babendreier, D. Mass rearing and release of *Trichogramma* for biological control of insect pests of corn in China. *Biol. Control.* **68**, 136–144 (2014).
5. Zhang, Y. *et al.* Parasitism and suitability of *Trichogramma Chilonis* on large eggs of two factitious hosts: *Samia Cynthia Ricini* and *Antheraea Pernyi*. *Insects* **15**, 2 (2024).
6. Wang, C. *et al.* Knockdown of vitellogenin receptor based on minute insect RNA interference methods affects the initial mature egg load in the pest natural enemy *Trichogramma dendrolimi*. *Insect Sci.* **0**, 1–14 (2024).
7. Tian, J. *et al.* The effects of temperature and host age on the fecundity of four *Trichogramma* species, egg parasitoids of the *Cnaphalocrocis Medinalis* (Lepidoptera: Pyralidae). *J. Econ. Entomol.* **110**, 949–953 (2017).
8. Yuan, X., Guo, Y. & Li, D. Field control effect of *Telenomus Remus* Nixon and *Trichogramma chilonis* Ishii compound parasitoid balls against *Spodoptera Frugiperda* (J. E. Smith). *Insects* **15**, 28 (2024).
9. Jiang, Z. *et al.* An evaluation of the growth, development, reproductive characteristics and pest control potential of three *Trichogramma* species on *Tuta Absoluta* (Meyrick) (Lepidoptera: Gelechiidae). *Pest Manag. Sci.* **80**, 6107–6116 (2024).
10. Shahid, R., Suhail, A., Gogi, D., Shahzad, M. & Hussain, S. Effectiveness of *Trichogramma Chilonis* (Ishii) (Hymenoptera: Trichogrammatidae) against sugarcane stem borer *Chilo Infuscatellus* (Snell) (Lepidotera: Pyrallidae). *Pak. Entomol.* **29**, 141–146 (2007).
11. Yang, X. *et al.* Parasitism and suitability of fertilized and nonfertilized eggs of the rice striped stem borer, *Chilo Suppressalis* (Lepidoptera: Crambidae), for *Trichogramma* parasitoids. *J. Econ. Entomol.* **109**, 1524–1528 (2016).
12. Pawar, P., Murali Baskaran, R. K., Sharma, K. C. & Marathe, A. Enhancing biocontrol potential of *Trichogramma Chilonis* against borer pests of wheat and chickpea. *iScience.* **26**, 106512 (2023).
13. Haoxiang, Z. *et al.* Insights from the biogeographic approach for biocontrol of invasive alien pests: Estimating the ecological niche overlap of three egg parasitoids against *Spodoptera Frugiperda* in China. *Sci. Total Environ.* **862**, 160785 (2023).
14. Ye, X., Yang, Y., Zhao, X., Fang, Q. & Ye, G. The state of parasitoid wasp genomics. *Trends Parasitol.* **40**, 914–929 (2024).
15. Wang, A. *et al.* Which molecular marker is better? Comparative analyses of COI and ITS2 in molecular identification of *Trichogramma* (Hymenoptera: Trichogrammatidae). *Biocontrol.* **68**, 483–494 (2023).
16. Belton, J. *et al.* Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods.* **58** (2012).

17. Bushnell, B., Rood, J. & Singer, E. BBMerge – Accurate paired shotgun read merging via overlap. *PloS One* **12**, e185056 (2017).
18. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).
19. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods.* **18**, 170–175 (2021).
20. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*. **34** (2018).
21. Guan, D. *et al*. Identifying and removing haplotypic Duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
22. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst*. **3**, 95–98 (2016).
23. Dudchenko, O. *et al*. De novo assembly of the *Aedes Aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
24. Steinegger, M. & Söding, J. MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
25. Camacho, C. *et al*. Blast+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
26. Fatemeh, F. & Jamasb, N. Status of cell division, cytogenetical index, and chromosome details on the immature stage of *Trichogramma brassicae* (Hymenoptera: Trichogrammatidae). *Arthropods.* **10**, 130–139 (2021).
27. Farsi, F., Ero Lu, H. E., Nozari, J. & Hosseininaveh, V. Karyotype Analysis of *Trichogramma embryophagum* Htg. (Hymenoptera: Trichogrammatidae) using a new method and estimate its karyotype symmetry. *Caryologia*. **73** (2021).
28. Gokhman, V., Pereira, F. & Costa, M. A Cytogenetic study of three parasitic wasp species (Hymenoptera, Chalcidoidea, Eulophidae, Trichogrammatidae) from Brazil using chromosome morphometrics and base-specific fluorochrome staining. *Comp. Cytogenet.* **11**, 179–188 (2017).
29. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA*. **117**, 9451–9457 (2020).
30. Hubley, R. *et al*. The Dfam database of repetitive DNA Families. *Nucleic Acids Res.* **44**, v1272 (2015).
31. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA*. **6**, 11 (2015).
32. Tarailo-Graovac, M. & Chen, N. Using repeatmasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* **25**, 4–10 (2009).
33. Holt, C. & Yandell, M. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
34. 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).
35. Keilwagen, J., Hartung, F. & Grau, J. GeMoMa: Homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods Mol Biol.* **1962**, 161–177 (2019).
36. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: A web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312 (2004).
37. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-EP+: Eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom. Bioinform.* **2**, a26 (2020).
38. Kriventseva, E. V. *et al*. OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **47**, D807–D811 (2019).
39. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
40. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods.* **12**, 357–360 (2015).
41. Wallberg, A. *et al*. A hybrid de novo genome assembly of the honeybee, *Apis Mellifera*, with chromosome-length scaffolds. *BMC Genomics*. **20**, 275 (2019).
42. Kawamoto, M. *et al*. High-quality genome assembly of the silkworm, *Bombyx mori*. *Insect Biochem. Mol. Biol.* **107**, 53–62 (2019).
43. Pannebakker, B. A., Cook, N., van den Heuvel, J., van de Zande, L. & Shuker, D. M. Genomics of sex allocation in the parasitoid wasp *Nasonia vitripennis*. *BMC Genomics* **21**, 499 (2020).
44. Adams, M. D. *et al*. The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
45. Richards, S. *et al*. The genome of the model beetle and pest *Tribolium castaneum*. *Nature* **452**, 949–955 (2008).
46. Buchfink, B., Reuter, K. & Drost, H. Sensitive protein alignments at tree-of-life scale using diamond. *Nat. Methods.* **18**, 366–368 (2021).
47. Finn, R. D. *et al*. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.* **45**, D190–D199 (2017).
48. Huerta-Cepas, J. *et al*. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
49. El-Gebali, S. *et al*. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
50. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* **46**, D493–D496 (2018).
51. Wilson, D. *et al*. SUPERFAMILY-sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.* **37**, D380–D386 (2009).
52. Marchler-Bauer, A. *et al*. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* **45** (2016).
53. Huerta-Cepas, J. *et al*. eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
54. Chen, C. *et al*. TBtools: An integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant*. **13**, 1194–1202 (2020).
55. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
56. Griffiths-Jones, S. *et al*. Rfam: Annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–D124 (2005).
57. Chan, P. & Lowe, T. tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods Mol Biol.* **1962**, 1–14 (2019).
58. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP547575 (2024).
59. Wang, C. *Genebank* https://identifiers.org/ncbi/insdc.gca:GCA_045269175.1 (2024).
60. Wang, C. The annotation results for repeated sequences, gene structure, and functional prediction of *Trichogramma Chilonis*. *Figshare. dataset.* https://doi.org/10.6084/m9.figshare.27872361 (2024).
61. Waterhouse, R. M. *et al*. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).

Acknowledgements

This work was supported by the National Key R&D Program of China (2023YFD1401200) and the Taishan Scholars Program of Shandong Province (tsqn202312293).

Author contributions

Y.Z., C.W., Z.Y. and L.Z. contributed to the research design. C.W., Z.Y., S.Z., R.W. and Y.W. collected the samples. C.W., Y.L. and X.D. analyzed the data. L.S., H.C. and L.Z. contributed to data quality control. C.W., Y.Z. and Z.Y. wrote the draft manuscript and revised the manuscript. All co-authors contributed to this manuscript and approved it.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04792-5.

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

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.