



OPEN

DATA DESCRIPTOR

High-quality genome assembly of a cosmopolitan insect predator, *Chrysoperla zastrowi sillemi* (Esben-Petersen)

Muthugounder Mohan¹✉, Thiruvengadam Venkatesan¹✉, Selvapandian Upasna¹, Selvamani Selva Babu¹, P J Aneesha¹✉, Karuppannasamy Ashok^{1,2}✉, Budhwar Roli³, Gandhi R Gracy¹ & Satya N Sushil¹

Chrysoperla zastrowi sillemi is an extensively employed bio-control agent against crop pests. However, the genomic information on adaptive evolution and predator-prey interaction was lacking to enhance its predatory potential. Here, we presented a high contiguity, chromosomal-level genome assembly of *C. zastrowi sillemi* using short and long-read sequencing (Illumina and PacBio) coupled with Hi-C. The highly homozygous genome assembly has 597 Mb in span. The assembly contains six chromosomal pseudomolecules, occupying 84% of genome, which range from 29.36 to 170.05 Mb. Genome annotation using Refseq, EggNOG, SwissProt and Eukaryotic Orthologous Groups database has identified 14,495 protein coding genes. The mitochondrial genome is 16 Kb in length with 13 protein-coding genes, 22 tRNAs and three rRNA genes without any gene rearrangements. The genomic information furnished herein could be useful in choosing heritable traits for selective breeding to improve its bio-control potential.

Background & Summary

An essential component of integrated pest management is biological control which uses natural enemies to control insect pests¹. *Chrysoperla zastrowi sillemi* (Esben-Petersen) (Neuroptera: Chrysopidae), commonly known as the green lacewing or aphid lion is a potential predator that is a key component of biological control in many crop ecosystems^{2,3}. It has high predatory potential, and feeds mainly upon mites, aphids, coccids, leafhoppers and small lepidopteran larvae⁴. Many studies have proved the predatory efficacy of *C. zastrowi sillemi* under laboratory and field conditions^{5,6}. For unravelling the functional genomics, it is crucial to obtain the whole-genome information of the predatory species.

In the present study, the genome of *C. zastrowi sillemi* was sequenced using PacBio long read and Illumina paired-end short read sequencing. In addition, transcriptome sequencing was done to predict genes with more accuracy. The k-mer analysis estimated its genome size as 566 Mb. The k-mer distribution curve indicates the genome of the highly homozygous in nature (Fig. 1).

The raw sequence reads yielded 15 Gb of filtered PacBio HiFi reads with an average read length of 12,416 bp. Hifiasm assembler was used to create haplotype resolved assembly using HiFi and Hi-C data. The assembled genome of *C. zastrowi sillemi* has 597 Mb with scaffold N50 of 30.4 Mb (Fig. 2). The genome was scaffolded into six chromosomal pseudomolecules covering 84% of the genome. The size of the pseudochromosomes ranged from 29.37 Mb to 170.05 Mb with total size of 502 Mb (Figure S1). In Hi-C interaction, a broader view of interactions across the entire chromosome or genome is identified as linear (Figure S2a), which means the a strong interaction signal observed between two genomic regions that are located relatively close to each other on the linear chromosome⁷. The higher cis/trans ratio indicates enrichment for within the chromosomal reads (Figure S2b). The cis long/short ratio of >1.5 also suggested that the Hi-C experiment captured long - distance chromatin interactions. The summary statistics of Hi-C data is presented in Figure S3.

¹Division of Genomic Resources, ICAR- National Bureau of Agricultural Insect Resources, Hebbal, Bengaluru, 560024, India. ²Tata Institute for Genetics and Society, Bengaluru, 560065, India. ³Bionivid Technology Private Limited, Bengaluru, 560043, India. ✉e-mail: Mohan.M@icar.gov.in; Venkatesan.T@icar.gov.in; ashokg3s@gmail.com

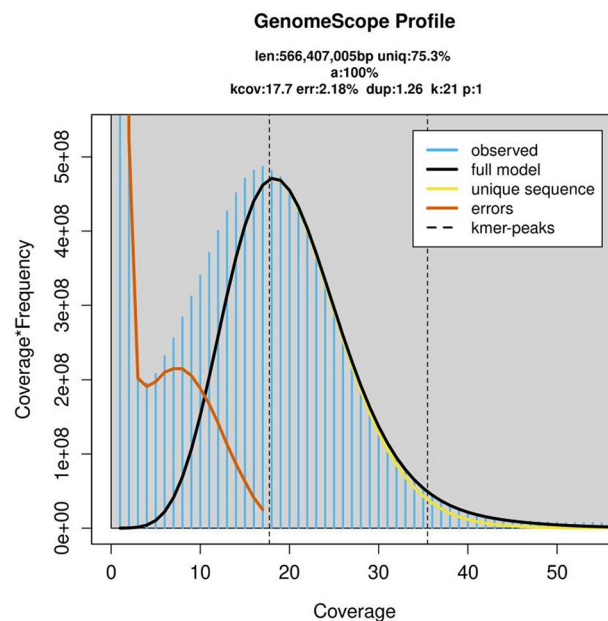


Fig. 1 K-mer frequency distribution plotted by GenomeScope from Illumina short reads for the estimation of genome size of *C. zastrowi sillemi*. The distribution was determined with jellyfish using a k-mer size of 21. The highly homozygous nature of the genome is visualized by the absence of k-mers in the heterozygous region.

The synteny analysis showed a strong collinearity between the *C. zastrowi sillemi* and *Chrysoperla carnea* genomes (Fig. 3). A total of 14,495 protein-coding genes, after removing the repeat sequences, were functionally annotated using Refseq, EggNOG, SwissProt and Eukaryotic Orthologous Groups (KOG) databases (Table S1).

Further, 4,572 non-coding RNAs were annotated and classified into 1,175 rRNAs, 54 miRNAs, 47 snRNAs, 125 spliceosomal RNAs and 2,070 tRNAs. The 342.5 Mb repeat elements constitute 57.31% of the genome, in which a high proportion of unclassified repeats (53.8%) were recognized. The unclassified repeats are a common phenomenon in early diverged insect orders such as Neuroptera (Table S2). The complete mitochondrial genome of *C. zastrowi sillemi* was predicted to be 16,048 bp with GC content of 21.1% containing 22 tRNA genes, 13 protein-coding genes and 3 rRNA genes (Figure S4 and Table S3).

The genome information provides a crucial resource in choosing heritable traits for a selective breeding programme. It also helps in further in-depth understanding of various aspects of its life history traits including the evolution of predator-prey interaction and its ecological adaptation.

Materials and Methods

Insect culture. A laboratory reared population of *C. zastrowi sillemi* (National Accession No. NBAIL-MP-TRI-15) is being maintained at ICAR-National Bureau of Agricultural Insect Resources (NBAILR), Bengaluru for more than 100 generations⁸. Freshly emerged and starved adults were used for genome sequencing. DNA and RNA Extraction DNA from both male and female adults was isolated using the DNAeasy[®] blood and tissue kit (Qiagen) following the manufacturer's protocol. The insects were starved for 24 hours before DNA isolation. Total RNA was isolated from adult, larva and pupae of *C. zastrowi sillemi* using an RNAeasy[®] kit (Qiagen) following the manufacturers' protocol.

Genome sequencing. For the primary assembly, the DNA library was prepared with the SMRTbell[®] express template prep kit 2.0 to generate Hifi reads on the PacBio Sequel II platform. In addition, short-read deep sequencing was done by generating high-quality RNA and DNA libraries following the manufacturer's protocols of NEBNext's Ultra II FS DNA Library Prep Kit and Ultra[™] II Directional RNA Library Prep Kit for Illumina[®], respectively for assembly polishing and validation. Hi-C sequencing was done to detect the chromatin structures and interactions in the nucleus. Arima Genome-Wide HiC+ kit was used to prepare a high-quality library and sequenced with paired-end chemistry on Illumina Novaseq6000. Quality-based filtering (phred score threshold of 30) and adapter trimming were performed using fastp. Transcriptome sequencing was done for enhanced gene prediction on a flow cell using the Illumina NovaSeq6000 platform. The genome size was estimated using Jellyfish⁹, a k-mer based counter. GenomeScope2 was used to analyze the Jellyfish results to obtain information such as heterozygosity, repeat length and genome size.

Genome assembly. PacBio HiFi reads were directly used to generate the primary assembly. The Hifiasm¹⁰ assembler was used in the Hifi-only mode to obtain a haplotype-resolved assembly. A transcriptome assembly from the RNA-seq data was generated using the Trinity¹¹ de novo assembler. CDHIT-EST¹² was used to cluster together similar nucleotide sequences within the assembly based on the sequence and length similarity. Transcripts that were 90% identical in length and sequence similarity were grouped together and only a representative transcript (Unigene) was retained.

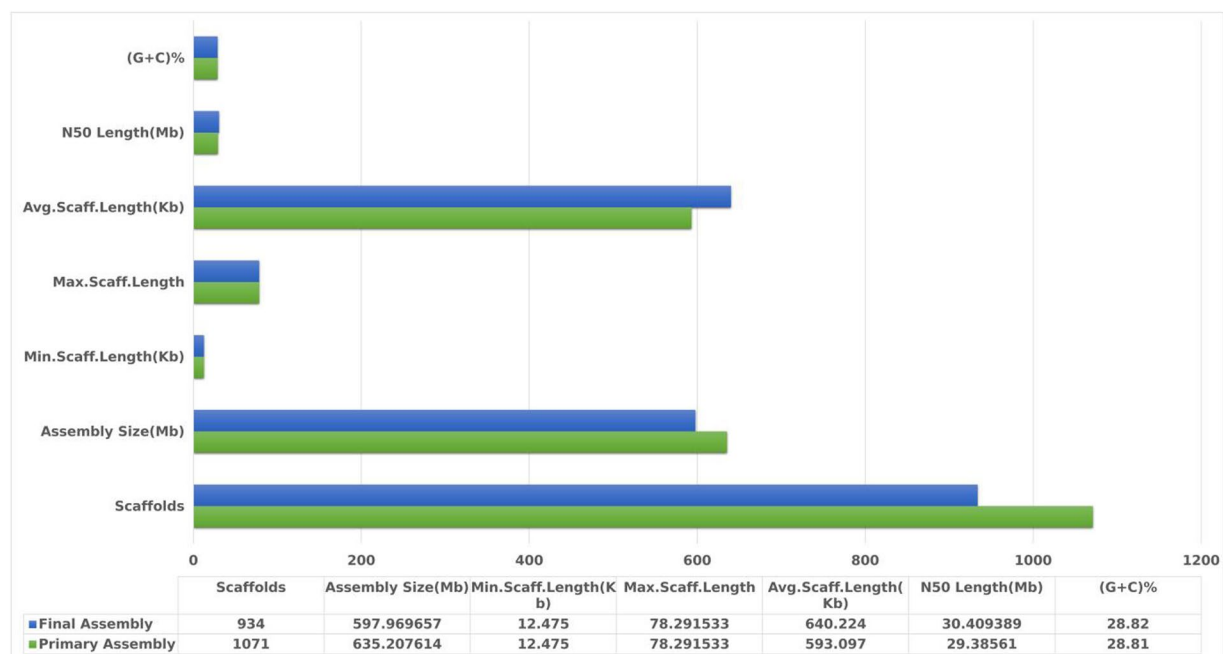


Fig. 2 Scaffold level genome assembly statistics of *C. zastrowi sillemi*.

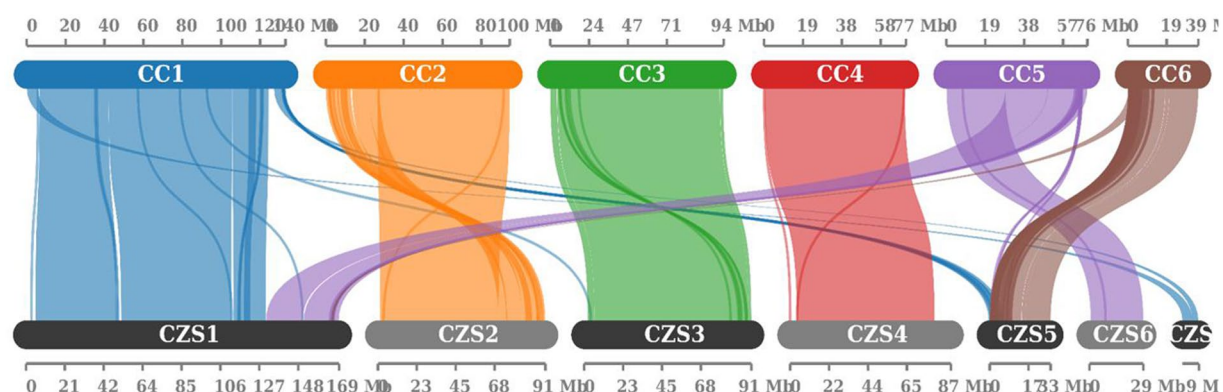


Fig. 3 Synteny blocks display high degree of conservation of regions of chromosomes between the genomes of *C. carnea* (CC) and *C. zastrowi sillemi* (CZS).

Hi-C scaffolding and assembly polishing. The HiCUP¹³ pipeline was used to align the reads against the assembly and filter out the artificial experimental artefacts. Further, this alignment information was used by the YAHS¹⁴ pipeline to join the gaps, re-orient sequences and produce more contiguous scaffolds/pseudo- molecules. The assembly was screened for possible bacterial contamination using Kraken2¹⁵. The standard database containing archaea, bacteria, viral, plasmid, human and UniVec Core sequences was used to screen the assembly.

Assembly validation. The whole genome and RNA-seq data were mapped back to the assembly with BWA¹⁶ and Hisat2¹⁷ aligners as a means of validation. Benchmarking Universal Single-Copy Orthologs (BUSCO)¹⁸ was used as a qualitative measure to ensure assembly completeness and annotation. The insecta_db10 database was used to assess the completeness of the *C. zastrowi sillemi* assembly. Similarly, Core Eukaryotic Genes Mapping Approach (CEGMA)¹⁹ that works with a database of 248 conserved eukaryotic genes was used to assess assembly completeness and annotation using Hidden Markov Models (HMMs).

Repeat prediction. Repeat sequences were identified in the assembly using RepeatMasker²⁰ that was used to soft mask the repeat regions identified with the Dfam (v3.6)²¹ database and a custom library was created using predictions from RepeatModeller²², Transposon-psi²³ and LTRharvest²⁴. The custom FASTA sequences were clustered together using USEARCH to remove redundant sequences and provided to RepeatMasker as a reference for the second round of repeat masking.

Gene prediction. The Maker2 pipeline was utilized for gene prediction. It is a wrapper package that encompasses the ab initio gene predictors Augustus²⁵ and GeneMark-ES and correlates transcript and protein evidence from related species to predict genes in the assembly. The Prothint²⁶ pipeline that uses the arthropoda_odb10 dataset from OrthoDB was used to score hints in the form of splice sites, introns, start and stop codons from the assembly. GeneMark EP + and the protein hints obtained from the orthologous protein sequences were used for accurate gene predictions. Chrysopidae family proteins were downloaded from the NCBI database and the de novo assembled transcriptome assembly was used as evidence for the gene prediction in Maker2²⁷. An Annotation Edit Distance (AED) value of less than one and sufficient exons and splice site coverage by alignment evidence were kept as parameters for filtering the predicted gene.

Genome annotation. Protein sequences from NCBI Refseq database were used to annotate the predicted transcripts/proteins. NCBI Blast + (v2.11)²⁸ was used for the annotation. The proteins were further classified into functional annotation categories such as Gene Ontology, eukaryotic orthologous groups (KOGs), KEGG pathways and PFAM domains using the EggNOG mapper²⁹ based on the EggNOG database²⁹.

Miscellaneous chromosomes. The X chromosome was identified from the synteny analysis with the published *C. carnea* X chromosome (NCBI accession: NC058342.1). The synteny was generated with the D- GENIES³⁰ software which computes large genome alignments with the minimap2 aligner. In addition, the mitochondrial assembly was identified using the assembly coverage information and homology alignment with other published assemblies of the Chrysopidae family. Mitochondrial genes were predicted using the Mitos2³¹ web server.

Synteny analysis. MCScanX³² which uses the MCScan algorithm was used to compute the syntenic blocks and co-linear segments among chromosomes. The *C. zastrowi sillemi* genome assembly was compared with *C. carnea* genome assembly was compared with and the synteny was viewed on the SynVisio online portal (<https://synvisio.github.io/>)³³.

Data Records

This Whole Genome Shotgun project has been deposited at NCBI GenBank under the accession JBFCE000000000.1³⁴ and GenBank assembly accession GCA_040669965.1. All the raw data files used in genome assembly such as PacBio, Illumina, and Hi-C sequencing data were deposited at the Sequence Read Archive (SRA) at NCBI with the accession number SRP410139³⁵ under BioProject accession number PRJNA905226. The Annotation file published in Figshare (<https://doi.org/10.6084/m9.figshare.28099868>)³⁶.

Technical Validation

Validations were employed to assess the contiguity, accuracy and completeness of the genome assembly. The final assembly validation resulted in the mapping of 85.76% and 95.50% of whole genome sequencing and Illumina raw data, respectively (Figure S5). The BUSCO analysis showed a high degree of completeness (99.0% and 94.3%) when compared with the Arthropoda data set of the Benchmark of Universal Single-Copy Orthologs (BUSCO) and Core Eukaryotic Genes Mapping Approach (CEGMA) using the eukaryotic data sets. This also included 95.3% complete and single-copy BUSCO genes and 3.7% duplicated BUSCOs. NCBI RefSeq database records matched with 80.42% of the annotated genes, and 76.28%, 67.05% and 49.79% of the genes matched with EggNOG, SwissProt and KOG databases, respectively (Table S2).

Code availability

The present study did not use any custom scripts. All the data processing was done using the guidelines of standard pipelines and bioinformatics tools given in the methods section. Parameters for each software were detailed.

Received: 5 June 2024; Accepted: 31 January 2025;

Published online: 16 February 2025

References

1. Baker, B. P., Green, T. A. & Loker, A. J. Biological control and integrated pest management in organic and conventional systems. *Biological Control* **140**, 104095, <https://doi.org/10.1016/j.biocontrol.2019.104095> (2020).
2. Pappas, M. L. & Koveos, D. S. Life-history traits of the predatory lacewing *Dichochrysa prasina* (Neuroptera: Chrysopidae): temperature-dependent effects when larvae feed on nymphs of *Myzus persicae* (Hemiptera: Aphididae). *Annals of Entomological Society of America* **104**, 43–49, <https://doi.org/10.1603/AN10036> (2011).
3. Venkatesan, T. *et al.* Occurrence of *Chrysoperla zastrowi arabica* Henry *et al.* (Neuroptera: Chrysopidae), a cryptic song species of *Chrysoperla* (carnea-group), in India. *Journal of Biological Control* **143**–147 (2008).
4. Halder, J. & Rai, A. B. Suitability of different prey aphids on the growth, development and reproduction of *Chrysoperla zastrowi sillemi* (Esben-Petersen) (Chrysopidae: Neuroptera). in: *Proceedings of the Zoological Society* **69**, 89–95, <https://doi.org/10.1007/s12595-014-0131-6> (2016).
5. Nair, I. J., Sharma, S. & Kaur, R. Efficacy of the green lace wing, *Chrysoperla zastrowi sillemi* (Esben-Peterson) (Neuroptera: Chrysopidae), against sucking pests of tomato: an appraisal under protected conditions. *Egyptian Journal of Biological Pest Control* **30**, 74, <https://doi.org/10.1186/s41938-020-00277-2> (2020).
6. Kubavat, A. K., Jethva, D. M. & Wadaskar, P. S. Biology and feeding potential of *Chrysoperla zastrowi sillemi* (Esben-Peterson) on eggs of *Corcyra cephalonica* (Stainton). *Environment and Ecology* **35**, 1948–1952 (2017).
7. Lajoie, B. R., Dekker, J. & Kaplan, N. The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods* **72**, 65–75 (2015).
8. Venkatesan, T., Mahiba, S. H., Jalali, S. K., Ramya, S. L. & Prathibha, M. Detection of insecticide resistance and mechanisms of resistance in field populations of *Chrysoperla zastrowi sillemi* (Neuroptera: Chrysopidae) collected from different geographical locations in India. *Journal of Biological Control* **31**, 159–167, <https://doi.org/10.18311/jbc/2017/16333> (2017).

9. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770, <https://doi.org/10.1093/bioinformatics/btr011> (2011).
10. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170–175, <https://doi.org/10.1038/s41592-020-01056-5> (2021).
11. Grabherr, M. G. *et al.* Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology* **29**, 644, <https://doi.org/10.1038/nbt.1883> (2011).
12. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659, <https://doi.org/10.1093/bioinformatics/btl158> (2006).
13. Wingett, S. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res* **4**, <https://doi.org/10.12688/f1000research.7334.1> (2015).
14. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, btac808, <https://doi.org/10.1093/bioinformatics/btac808> (2023).
15. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biology* **20**, 1–13, <https://doi.org/10.1186/s13059-019-1891-0> (2019).
16. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595, <https://doi.org/10.1093/bioinformatics/btp698> (2010).
17. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**, 357–360, <https://doi.org/10.1038/nmeth.3317> (2015).
18. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212, <https://doi.org/10.1093/bioinformatics/btv351> (2015).
19. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067, <https://doi.org/10.1093/bioinformatics/btm071> (2007).
20. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* **25**, 4–10, <https://doi.org/10.1002/0471250953.bi0410s25> (2009).
21. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA* **12**, 1–14, <https://doi.org/10.1186/s13100-020-00230-y> (2021).
22. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451–9457, <https://doi.org/10.1073/pnas.1921046117> (2020).
23. Haas, B., 2007. TransposonPSI: an application of PSI-Blast to mine (retro-) transposon ORF homologies. Broad Institute, Cambridge, MA, USA.
24. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 1–14, <https://doi.org/10.1186/1471-2105-9-18> (2008).
25. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research* **33**, 465–467, <https://doi.org/10.1093/nar/gki458> (2005).
26. 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**, lqaa026, <https://doi.org/10.1093/nargab/lqaa026> (2020).
27. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 1–14, <https://doi.org/10.1186/1471-2105-12-491> (2011).
28. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 1–9, <https://doi.org/10.1186/1471-2105-10-421> (2009).
29. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution* **38**, 5825–5829, <https://doi.org/10.1093/molbev/msab293> (2021).
30. Cabanettes, F. & Klopp, C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**, e4958, <https://doi.org/10.7717/peerj.4958> (2018).
31. Donath, A. *et al.* Improved annotation of protein-coding genes boundaries in metazoan mitochondrial genomes. *Nucleic Acids Research* **47**, 10543–10552, <https://doi.org/10.1093/nar/gkz833> (2019).
32. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* **40**, 49–49, <https://doi.org/10.1093/nar/gkr1293> (2012).
33. Bandi, V. K. SynVisio: a multiscale tool to explore genomic conservation. (University of Saskatchewan, 2020).
34. Muthugoundar, M. Genome Assembly of *Chrysoperla zastrowi sillemi*. GenBank <https://identifiers.org/ncbi/insdc:JBFCZE000000000.1> (2024).
35. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRP410139> (2024).
36. Mohan, M. *et al.* Annotation files of *Chrysoperla zastrowi sillemi*. figshare. Dataset <https://doi.org/10.6084/m9.figshare.28099868.v1> (2024).

Acknowledgements

The authors extend gratitude to the Indian Council of Agricultural Research (ICAR) for generous funding under the Consortium Research Platform (CRP) on the Genomics project. The authors are thankful to the sequencing service extended by Bionivid Technology Pvt. Ltd., Bengaluru and the ICAR- Network project on Agricultural Bioinformatics and Computational Biology for bioinformatic support. The infrastructure facilities provided by the Director, ICAR- National Bureau of Agricultural Insect Resources, Bengaluru are gratefully acknowledged.

Author contributions

M.M. contributed to the Conception and design of work and interpretation of data. T.V. contributed to the rearing and maintenance of iso-female colonies of *C. zastrowi sillemi* and DNA barcoding. S.U. contributed to Assembly validations. S.B.S. contributed to Gene and orthology predictions. A.P.J. contributed to Hi-C scaffolding. G.R.G. contributed to CAFE analysis & manuscript editing. R.B. contributed to the synteny analysis. S.N.S. contributed to the preparation of samples for transcriptome studies. A.K. contributed to the drafting of the manuscript.

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/s41597-025-04571-2>.

Correspondence and requests for materials should be addressed to M.M., T.V. or K.A.

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.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025