scientific data

OPEN

DATA DESCRIPTOR

Chromosome–level genome assembly of the seasonally polyphenic scorpionfly (*Panorpa liui*)

Jiuzhou Liu[1,2,5], Yuetian Gao[1,5], Shuangmei Ding[3], Shuai Zhan[4], Ding Yang[1] ✉ & Xiaoyan Liu[2] ✉

Mecoptera is a small relict order of insects within the Holometabola. Panorpidae is the most speciose family in Mecoptera. They are also known as scorpion flies due to the enlarged and upward recurved male genital bulb. *Panorpa liui* Hua, 1997, a member of Panorpidae, is a bivoltine species of seasonal color polyphenism in the lowland plain of northeastern China. In this study, we applied PacBio HiFi and Hi–C sequencing technologies to generate a chromosome-level genome reference of *P. liui*. The assembled genome is 678.26 Mbp in size, with 91.3% being anchored onto 23 pseudo–chromosomes. Benchmarking Universal Single–Copy Orthologs (BUSCO) estimation reveals the completeness of this assembly as 95.1%. By integrating full-length transcriptome and homologs of related species, we generated full annotation of this assembly, yielding a total of 15,960 protein–coding genes, of these, 15,892 genes were anchored on the 23 chromosomes. The high-quality genome provides critical genomic resources for population genetics and phylogenetic research on Mecoptera. It also offers valuable information for exploring the mechanisms underlying seasonal color polymorphism.

Background & Summary

The origin and diversification of species have always been central themes in evolutionary biology. Reconstructing the evolutionary relationships among species is fundamental to comprehending biodiversity[1]. Accurate and comprehensive reference genome assemblies are crucial for genetic and whole–genome studies of individuals and multiple species. The Mecoptera, commonly known as scorpionflies, belongs to the Panorpida under Holometabola, along with the Siphonaptera, Diptera, Lepidoptera, and Trichoptera[2]. It is composed of 9 families with 40 genera and more than 800 extant species[3]. Scorpionflies are renowned for their varied feeding habits. Although traditionally perceived as active predators, mounting evidence indicates that they are omnivorous. Their diets primarily consist of dead or dying arthropods, supplemented by other invertebrates, nectar, pollen, and animal feces[4]. Additionally, certain scorpionfly taxa (e.g. *Panorpa*) exhibit kleptoparasitism on insect carcasses in spider webs[5]. Some studies propose that scorpionflies possess research significance in forensic entomology[6]. Furthermore, they can also serve as indicator organisms for environmental monitoring[7,8].

The phylogenetic relationship among Mecoptera, Siphonaptera, and Diptera has long been a focal topic of investigation, representing one of the most enduring issues in insect evolution and systematics[9,10]. The analysis of larval mouthpart characteristics and the structure of the proventriculus suggests a close relationship between scorpionflies and Siphonaptera[11,12], which is also supported by subsequent molecular phylogenetic studies[2,13]. However, the relationship among the early branches of holometaboles remains controversial. Some scholars have proposed that Mecoptera is a paraphyletic group, with Siphonaptera placement within Mecoptera, forming a sister group to Boreidae or Nannochoristidae[14,15]. Furthermore, transcriptomic and mitochondrial genome data in phylogenetic analysis corroborate the hypothesis that fleas are specialized scorpion flies, implying that Siphonaptera should be regarded as an infraorder within Mecoptera[16]. Nonetheless, some studies suggest a closer phylogenetic relationship between Diptera and Mecoptera or Siphonaptera, thereby challenging the

[1]Department of Entomology, China Agricultural University, Beijing, 100193, China. [2]Hubei Insect Resources Utilization and Sustainable Pest Management Key Laboratory, College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, 430070, China. [3]The Institute of Scientific and Technical Research on Archives, NAAC, Beijing, 100053, China. [4]Key Laboratory of Plant Design, CAS Center for Excellence in Molecular Plant Sciences, Chinese Academy of Sciences, Shanghai, 200032, China. [5]These authors contributed equally: Jiuzhou Liu, Yuetian Gao. ✉e-mail: dyangcau@126.com; liuxiaoyan@mail.hzau.edu.cn

Sequencing strategy	Platform	Usage	Clean data (Gb)
Long–reads	PacBio (PacBio Sequel II)	Assembly	28.01
Hi–C	Illumina (HiSeq)	Hi–C assembly	58.14
Short–reads	BGI (DNBSEQ-T7)	Genome survey	27.22
RNA–seq	Nanopore (PromethION)	Annotation	13.42

Table 1. Library sequencing methods and data used in this study.

proposed sister group relationship between scorpionflies and fleas[17,18]. In summary, disputes persist regarding the phylogenetic relationships among Mecoptera, Siphonaptera, and Diptera.

High-quality genomic data offers a wealth for understanding the biology of a species at the molecular level and clarifying the evolutionary relationship among these three groups. Additionally, *P. liui* is a very representative ubiquitous taxon of scorpion flies in China, this species exhibits seasonal polymorphism, with the spring generation being mostly black and the summer generation being yellow[19]. Hence this study is also helpful for the research of insect's seasonal color polymorphism.

In this study, we used PacBio HiFi and high–throughput Chromosome Conformation Capture (Hi–C) sequencing technologies to assemble the chromosome–level genome of *P. liui*. The final genome size was 678.26 Mb, with assembled genome sequences successfully anchored onto 23 chromosomes, which is consistent with previous studies on other scorpionfly species[20]. A total of 15,960 protein-coding genes were identified, and 83.87% of them were functionally annotated. BUSCO analysis showed that the completeness of the genome assembly is 95.1%. The presented assembly was demonstrated with a significant degree of continuity and completeness. Overall, the newly presented *P. liui* genome will provide a fundamental resource for studying the adaptive evolution of the Mecoptera, including the famous seasonal color polymorphism in *P. liui*, and for supporting the future comparative genomic analyses in holometabolous insects.

Methods

Sampling and sequencing. The specimens of *P. liui* used in this study were collected in Donglin, Shenyang, Liaoning, China on July 20th, 2022. Since most reported karyotypes for Mecoptera are XX/XO (sex chromosome deletion), we selected the female specimens for short reads, Illumina Hi–C and Pacbio Hifi with Circular Consensus Sequencing (CCS). The male specimen was used for RNA extraction to supplement additional sequencing data.

Total genome DNA was extracted using the Cetyltrimethylammonium Bromide (CTAB) method to obtain high–quality genomic DNA and assessed by 0.75% agarose gel electrophoresis. The concentration of DNA was then precisely quantified using a Qubit 3.0 fluorometer (Life Technologies, Carlsbad, CA, USA).

For Pacbio Hifi with Circular Consensus Sequencing (CCS), the high–quality purified genomic DNA samples were utilized for constructing PCR–free SMRT bell libraries which were sequenced using the PacBio Sequel II platform. Following sequencing, reads with a length less than 50 bp, quality value lower than 0.8 and adapter sequences were eliminated from the polymerase reads to obtain only the insert fragments, referred to as Subreads. The Subreads were then processed using SMRTLink v8.0 (https://www.pacb.com/support/software-downloads) to generate HiFi reads, which yielded 28.01 Gb (coverage: 41.30 X) of valid data with 1,878,183 HIFI reads and the HIFI reads length N50 is 14,916 bp. The Hi–C technology was used to assist the genome assembly at the chromosome–level. The purified nuclei were digested with DpnII enzyme. Hi-C samples were then prepared through end repair, biotin labeling, flat-end ligation, DNA purification, and random shearing into 300 to 700 bp fragments. Finally, the Hi-C fragment library was quantified and sequenced using the Illumina HiSeq platform with paired-end 150 bp reads. In total, 58.14 Gb (coverage: 85.72 X) of Hi–C raw data was obtained. For short reads sequencing, the qualified DNA samples were randomly fragmented using a Covaris ultrasonicator. Sequencing libraries were constructed using Plus DNA Library Prep Kit for MGI V2, with an insert size of 200–400 bp. After the libraries were qualified, we performed 150 bp paired-end sequencing on the DNBseq-T7 platform. The raw reads obtained from sequencing were filtered using Fastp v0.21.0[21], resulting in 27.22 Gb (coverage: 40.13 X) of clean data totally. For Nanopore full–length transcriptome sequencing, a library was constructed using the SQK–PCS109 kit (Oxford Nanopore Technologies, Oxford, UK). Subsequently, a specific concentration and volume of cDNA library was added to the flow cell, which was then transferred to the Oxford Nanopore PromethION sequencer for real–time single molecule sequencing. Sequences with an average quality value of 7 or lower were filtered out, resulting in a final dataset of 13.42 Gb of valid data, comprising 15,587,197 reads. The read lengths for N50 and N90 are 1,079 bp and 448 bp, respectively. Data from all sequencing results are available in Table 1.

Genome size estimation and assembly. Based on the reads obtained from Illumina sequencing, the genome size and heterozygosity were estimated using the k–mer analysis method by Jellyfish v.2.3.1[22] and GenomeScope v.2.0[23]. The k–mer frequency–depth distribution analysis suggested that the genome size is approximately 658,37 Mbp, with a heterozygosity of 0.99% based on the 19–mer frequency distribution. The final assembled genome size was 678.26 Mbp. This discrepancy arises from k-mer analysis being affected by sequencing errors, repetitive sequences, and heterozygous regions. Thus, the final genome size, provides a more accurate representation of the actual genome.

The PacBio long–reads were used for *de novo* genome assembly. The genome assembly was performed using the hifiasm v.0.16.1[24] based on the OLC (Overlap Layout Consensus) algorithm.

Genome features	HIFI assemble	Hi–C assemble
Total length (bp)	750,774,560 bp	678,258,355 bp
Contigs N50 (bp)	941,472 bp	25,266,336 bp
Contigs N90 (bp)	116,414 bp	15,620,259 bp
Longest Contig length (bp)	5,541,207 bp	72,422,398 bp
GC content (%)	29.79	29.79

Table 2. Features of *P. liui* genome.

	Summary	Number	Percent (%)
Genome assembly	Complete BUSCOs	1,299	95.1
	Complete and single–copy BUSCOs	1,253	91.7
	Complete and duplicated BUSCOs	46	3.4
	Fragmented BUSCOs	20	1.5
	Missing BUSCOs	48	3.4
Gene annotation	Complete BUSCOs	1,263	92.3
	Complete and single–copy BUSCOs	1,168	85.4
	Complete and duplicated BUSCOs	95	6.9
	Fragmented BUSCOs	15	1.1
	Missing BUSCOs	89	6.6

Table 3. Statistical result of BUSCO for *P. liui*.

The HiFi reads were assembled into 2419 contigs, with a contig N50 length of 0.94 Mb (Table 2). Benchmarking Universal Single–Copy Orthologs (BUSCO) v5.7.1 was used to evaluate the completeness of the assembly based on the insecta_odb10 database (1,367 genes)[25] and the results was shown in Table 3.

Hi–C libraries and genome scaffolding. The Hi–C technique was used to capture genome-wide chromatin interactions for assisting the chromosome–level assembly[26].

HICUP v. 0.8.0[27] was used to map the Hi–C data to the assembled genome sequence and assess the proportions of self–circle, dangling end, dumped pairs and valid interaction pairs in the effective Hi–C data (Supplementary Table 1). Then, according to agglomerative hierarchical clustering, contigs in the sketch were clustered and grouped into n chromosome clusters using ALLHIC v.0.9.8[28]. Subsequently, the contigs within each chromosome cluster were ordered, oriented, and located. The interaction relationships between contigs were then converted into specified binary files using 3D–DNA v.180419[29] and Juicer v.1.6[30]. Manual sequencing and orientation of contigs were performed using Juicebox v.1.11.08[31]. Finally, manually remove heterozygous sequences which refer to segments that have no interaction with a sequence of the same size but interact normally with other sequences. After Hi–C assembly and manual adjustment, a total of 619 Mbp of genome sequence was positioned on 23 chromosomes, accounting for 91.30% of the genome. The length of 23 pseudo–chromosomes ranged from 72 Mbp to 16 Mbp (Supplementary Table 2), respectively. HiCExplorer v. 3.6[32] was used to plot the contig interaction intensity against position (Fig. 1a). We conducted self-comparisons of the protein sequences using blastp v. 2.6.0+[33], followed by the identification of syntenic blocks with MCScanX[34]. Finally, we generated a circular plot (Fig. 1b) using Tbtools (v2.097)[35].

Genome annotation. The genome annotation process comprises three parts: repeat sequence annotation, gene structure and function annotation, and non–coding RNA annotation. Repeat sequences were predicted using RepeatModeler (http://www.repeatmasker.org/RepeatModeler/) to generate model sequences based on the genome. The Long Terminal Repeat (LTR) sequences were predicted using LTR_FINDER[36], with redundancies subsequently removed using LTR_retriever[37] to obtain non-redundant LTR sequences. *De novo* sequences and the RepBase library (http://www.girinst.org/repbase, version: 20181026) were merged to create a repeat sequence library, and RepeatMasker[38] was used to predict repeat sequences. Transposable element (TE) protein–type repeat sequences were predicted using RepeatProteinMask subroutine of RepeatMasker. Finally, all repeat prediction results were merged and redundancies were removed to obtain the final combined TEs results which accounted for a total of 63.50% of the genome (Supplementary Table 3).

Gene structure prediction mainly used transcriptome prediction, homologous protein prediction and *ab initio* prediction. Transcripts were predicted from Nanopore full–length transcriptome sequencing data by initially filtering the data using NanoFilt v.2.8.0[39], identifying full–length sequences using Pychopper v2.7.2 (https://github.com/epi2me–labs/pychopper), and error–correcting these sequences with racon v1.4.21 (https://github.com/isovic/racon) based on original reads. The error–corrected full–length sequences were aligned to the genome using minimap2[40] and the alignment results were used to reconstruct transcripts with Stringtie[41]. Finally, TransDecoder v5.1.0 (https://github.com/TransDecoder/TransDecoder) was used to predict coding regions within the reconstructed transcript regions. For the prediction of homologous protein, five species were selected: *Anopheles gambiae* (Diptera), *Drosophila melanogaster* (Diptera), *Nephrotoma flavescens* (Diptera),

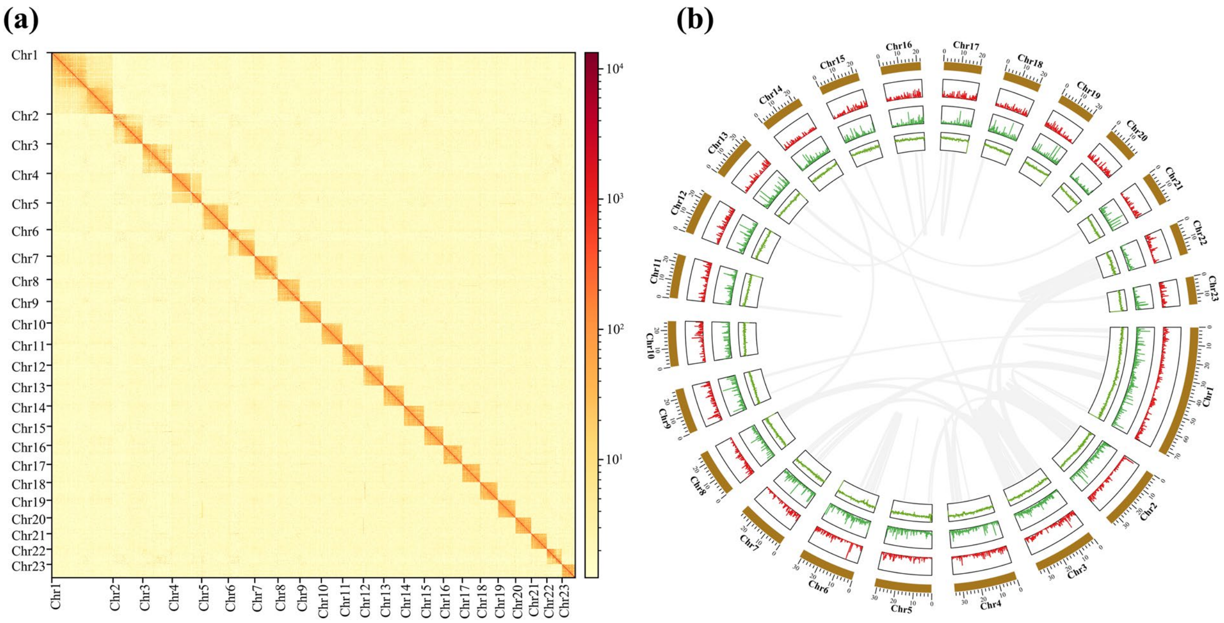

Fig. 1 The visualization of *Panorpa liui* genomic details resulting from high–quality assembly. (**a**) Hi–C interactive heatmap of the genome-wide organization of *Panorpa liui* chromosomal interval. The frequency of Hi–C interactive links represented by color, ranging from yellow (low) to red (high). Chr: chromosome. (**b**) Circular diagram depicting the characteristics of the *Panorpa liui* genome. The outer layer consists of colored blocks, representing the 23 linkage groups, with a circular demonstration of gene density (line), repeat density (line), and GC ratio (line) from the outer to the inner circle, respectively. Central gray lines represent syntenic links within and between chromosomes.

Bombyx mori (Lepidoptera) and *Ctenocephalides felis* (Siphonaptera). Tblastn v2.7.1[33] was used to align homologous protein sequences to the genome, then transcripts and coding regions were predicted based on the alignment results using Exonerate[42]. Furthermore, predicted genes based on BUSCO database were used as indirect homologous evidence. Augustus v.3.3.2[43] and Genscan v1.0[44] were used for *de novo* gene prediction analyses. Finally, MAKER v.2.31.10[45] was used to integrate the gene sets which predicted by the above methods. A total of 15,960 genes were predicted based on the results (Supplementary Table 4).

Gene functions were annotated using two methods: sequence similarity search and motif similarity search, referencing public databases such as Uniprot[46], NR[47], KEGG[48], KOG[49], Pfam[50] and Gene Ontology (GO)[51]. A total of 83.87% protein–coding genes were functionally annotated (Supplementary Table 5). Genomic non–coding RNA (ncRNA) was predicted using Infernal v.1.1.2[52] based on the Rfam database, while tRNA prediction used tRNAscan–SE v.1.23[53]. rRNA was predicted using RNAmmer v.1.250[54]. In total, 22,149 ncRNA sequences were annotated (Supplementary Table 6).

Data Records

The genome assembly data had been deposited at the National Center for Biotechnology Information (NCBI), under the accession number of JBDODE000000000[55]. The NCBI BioProject accession number is PRJNA1113301. Raw reads obtained for genome assembly have been deposited in the Sequence Read Archive (SRA) repository with the accession number of SRP508639[56]. The genome annotation GFF, CDS sequences, and protein sequences are available in Figshare[57].

Technical Validation

The quality assessment of *P. liui* genome assembly primarily focused on its completeness and accuracy. The genome completeness at the chromosome–level was evaluated using BUSCO with reference to the insects_odb10 database. A total of 1,299 (95.1%) complete genes were identified, including 1253 (91.7%) single–copy genes, 46 (3.4%) duplicated genes, 20 (1.5%) fragmented genes, and 48 (3.4%) missing genes. Furthermore, BUSCO was used to evaluate the predicted gene set, revealing 1,263 (92.3%) complete gene elements within the annotated gene set, comprising 1,168 (85.4%) single–copy genes, 95 (6.9%) duplicated genes, 15 (1.1%) fragmented genes and 89 (6.6%) missing genes. The accuracy of assembly was verified by calculating the mapping rates though aligning Illumina reads to the final assembly, resulting in successfully mapped rate of 99.40% and coverage rate of 99.5%, with an average coverage depth of 36.11. These results indicated that the majority of conservative genes were assembled and predicted with relatively high integrity, suggesting a high reliability of the predictions.

Code availability

No specific programs or codes were used in this study.

Received: 20 June 2024; Accepted: 27 December 2024;
Published online: 07 January 2025

References

1. Duchêne, D. A. Phylogenomics. *Curr. Biol.* **31**, R1177–R1181 (2021).
2. Wiegmann, B. M. *et al.* Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. *BMC Biol.* **7**, 1–16 (2009).
3. Wang, J. S. & Hua, B. Z. Morphological phylogeny of Panorpidae (Mecoptera: Panorpoidea). *Syst. Entomol.* **46**, 526–557 (2021).
4. Flügel, H. J. & Malec, F. Aktuelle schnabelfliegen-nachweise (Mecoptera) aus nord-hessen und ihre blütenbesuche. *Lebbimuk* **10**, 3–18 (2013).
5. Thornhill, R. Scorpionflies as kleptoparasites of web-building spiders. *Nature* **258**, 709–711 (1975).
6. Keiper, J. B. & Ivanov, K. Field observations of scorpionflies (Mecoptera: Panorpidae) and signal flies (Diptera: Platystomatidae) at animal carcasses. *Entomol. News* **129**, 301–306 (2020).
7. Dokjan, T. *et al.* Biodiversity and spatiotemporal variations of Mecoptera in thailand: Influences of elevation and climatic factors. *Insects* **15**, 151 (2024).
8. Vincent, A., Tillier, P., Vincent-Barbaroux, C., Bouget, C. & Sallé, A. Influence of forest decline on the abundance and diversity of raphidioptera and Mecoptera species dwelling in oak canopies. *Eur. J. Entomol.* **117** (2020).
9. Beutel, R. G., Yavorskaya, M. I., Mashimo, Y., Fukui, M. & Meusemann, K. The phylogeny of Hexapoda (Arthropoda) and the evolution of megadiversity. *In Proc. Arthropod. Embryol. Soc. Jpn* **51**, 1–15 (2017).
10. Kjer, K. M., Simon, C., Yavorskaya, M. & Beutel, R. G. Progress, pitfalls and parallel universes: a history of insect phylogenetics. *J. Royal Soc. Interface* **13**, 20160363 (2016).
11. Hinton, H. The phylogeny of the Panorpoid orders. *Annu. Rev. Entomol.* **3**, 181–206 (1958).
12. Richards, A. The proventriculus of adult Mecoptera and Siphonaptera. *Entomol. News* **76**, 253–256 (1965).
13. Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763–767 (2014).
14. Whiting, M. F. Mecoptera is paraphyletic: multiple genes and phylogeny of Mecoptera and Siphonaptera. *Zool. Scripta* **31**, 93–104 (2002).
15. Whiting, M. *Phylogeny of the Holometabolous Insects: The Most Successful Group of Terrestrial Organisms*, 345–361 (2004).
16. Tihelka, E. *et al.* Fleas are parasitic scorpionflies. *Palaeoentomology* **3** (2020).
17. Zhang, Y. *et al.* Mitochondrial phylogenomics provides insights into the taxonomy and phylogeny of fleas. *Parasites & Vectors* **15**, 223 (2022).
18. Zhao, X. *et al.* Mouthpart homologies and life habits of mesozoic long-proboscid scorpionflies. *Sci. Adv.* **6**, eaay1259 (2020).
19. Li, N., Jiang, L., Wang, J. S. & Hua, B. Z. Integrative taxonomy of the seasonally polyphenic scorpionfly *Panorpa liui* hua, 1997 (Mecoptera: Panorpidae). *Org. Divers. & Evol.* **21**, 533–545 (2021).
20. Miao, Y., Wang, J. S. & Hua, B. Z. Molecular phylogeny of the scorpionflies Panorpidae (Insecta: Mecoptera) and chromosomal evolution. *Cladistics* **35**, 385–400 (2019).
21. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
22. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
23. Vurture, G. W. *et al.* GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
24. 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).
25. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. Busco update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
26. Belaghzal, H., Dekker, J. & Gibcus, J. H. Hi-c 2.0: An optimized hi-c procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods* **123**, 56–65 (2017).
27. Wingett, S. *et al.* Hicup: pipeline for mapping and processing hi-c data. *F1000Research* **4** (2015).
28. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on hi-c data. *Nat. Plants* **5**, 833–845 (2019).
29. Dudchenko, O. *et al.* De novo assembly of the aedes aegypti genome using hi-c yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
30. Durand, N. C. *et al.* Juicer provides a one-click system for analyzing loop-resolution hi-c experiments. *Cell Syst.* **3**, 95–98 (2016).
31. Durand, N. C. *et al.* Juicebox provides a visualization system for hi-c contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
32. Wolff, J. *et al.* Galaxy hicexplorer 3: a web server for reproducible hi-c, capture hi-c and single-cell hi-c data analysis, quality control and visualization. *Nucleic Acids Res.* **48**, W177–W184 (2020).
33. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC bioinformatics.* **10**, 1–9 (2009).
34. Wang, Y. *et al.* MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49–e49 (2012).
35. Chen, C. *et al.* TBtools-II: a one for all, all for one bioinformatics platform for biological big-data mining. *Mol. Plant.* **16**, 1733–1742 (2023).
36. Ou, S. & Jiang, N. Ltr_finder_parallel: parallelization of ltr_finder enabling rapid identification of long terminal repeat retrotransposons. *Mob. DNA* **10**, 48 (2019).
37. Ou, S. & Jiang, N. Ltr_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
38. Tarailo-Graovac, M. & Chen, N. Using repeatmasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinforma.* **25**, 4.10.1–4.10.14 (2009).
39. De Coster, W., D'hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. Nanopack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
40. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
41. Kovaka, S. *et al.* Transcriptome assembly from long-read rna-seq alignments with stringtie2. *Genome Biol.* **20**, 1–13 (2019).
42. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinforma.* **6**, 1–11 (2005).
43. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cdna alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
44. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic dna. *J. Mol. Biol.* **268**, 78–94 (1997).
45. Holt, C. & Yandell, M. Maker2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinforma.* **12**, 1–14 (2011).

46. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
47. Deng, Y. *et al*. Integrated nr database in protein annotation system and its localization. *Comput. Eng.* **32**, 71–72 (2006).
48. Kanehisa, M. & Goto, S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
49. Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Expanded microbial genome coverage and improved protein family annotation in the cog database. *Nucleic Acids Res.* **43**, D261–D269 (2015).
50. Finn, R. D. *et al*. Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).
51. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
52. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster rna homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
53. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. trnascan-se 2.0: improved detection and functional classification of transfer rna genes. *Nucleic Acids Res.* **49**, 9077–9096 (2021).
54. Lagesen, K. *et al*. Rnammer: consistent and rapid annotation of ribosomal rna genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
55. Liu, J. Z. *GenBank. Panorpa liui isolate A01, whole genome shotgun sequencing project.* https://identifiers.org/ncbi/insdc:JBDODE000000000 (2024).
56. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP508639 (2024).
57. Annotation and protein sequences of *Panorpa liui* Hua, 1997 genome. *figshare* https://doi.org/10.6084/m9.figshare.25854778.v1 (2024).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.32130012).

Author contributions

D.Y. and X.L. conceived the project. Y.G. collected samples. J.L. and Y.G. performed the experiments. J.L., Y.G., S.D., S.Z., D.Y. and X.L. performed the analysis and wrote the manuscript. All authors contributed revising the manuscript and approved the final version 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/s41597-025-04365-6.

Correspondence and requests for materials should be addressed to D.Y. or X.L.

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.